香港城市大學
City University of Hong Kong

南開大學
Nankai University

‘AI’ and AI:
Sharing some Progresses from EL&CL team

Qibin Ran
School of Liberal Arts, Nankai University

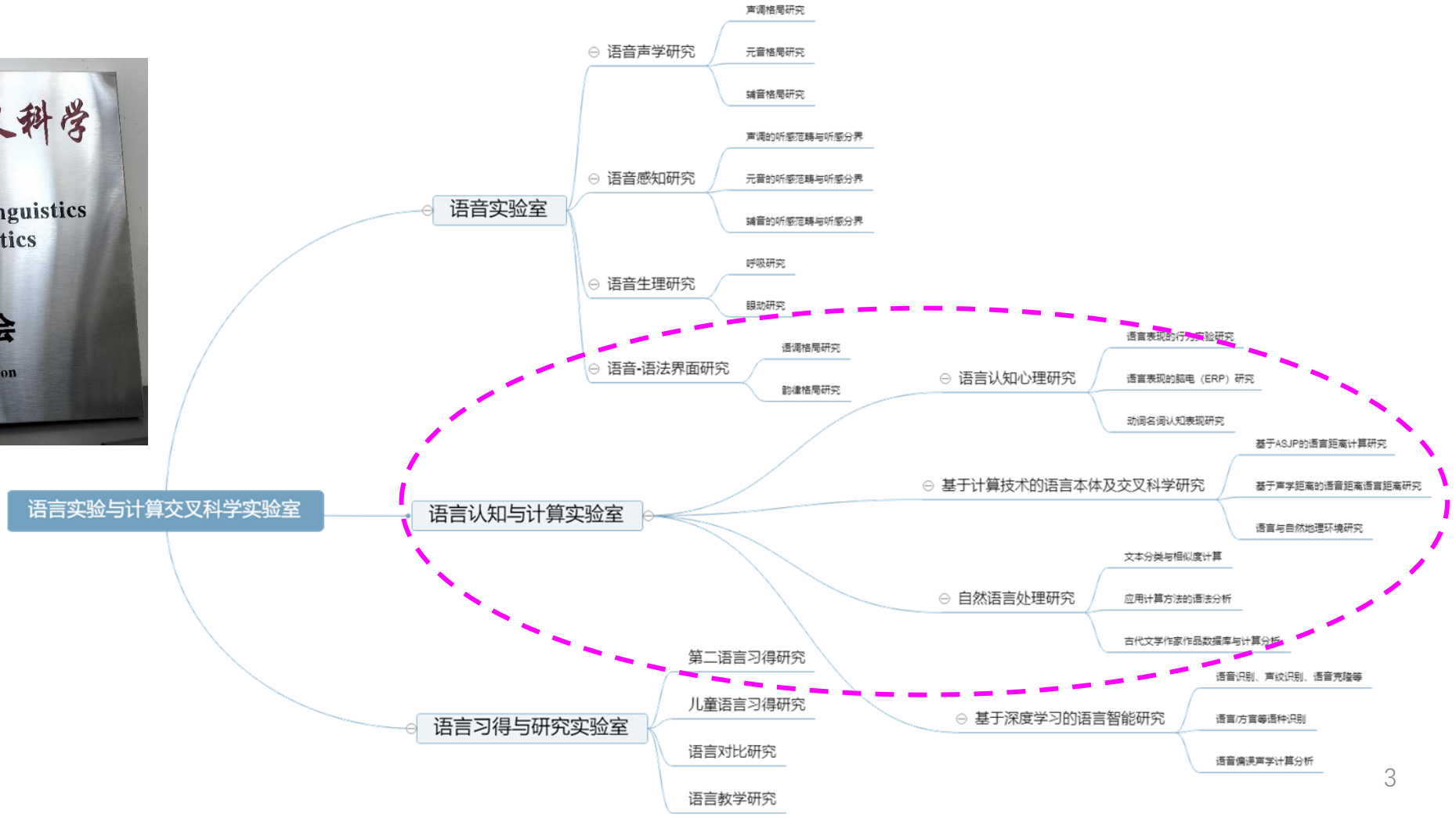
Jan.20, 2025

History of my academic career

- Laboratory name: EL&CL(Experimental Linguistics and Computational Linguistics)
- What actually we do: Linguistics study using quantitative approaches
- Several turns in recent years: Linguistic distance calculation（ASJP approach）
- （cross-disciplinary/interdisciplinary studies based on linguistics）
- Deep learning/AI(applications in linguistics studies)

Literature → Ancient Chinese → Phonetics → Language distance calculation → Language and environment → Deep learning/AI

- **Title: Social Science Laboratory of Tianjin - Interdisciplinary Laboratory for Experimental Linguistics and Computational Linguistics**



- **Social Science Laboratory of Tianjin - Interdisciplinary Laboratory for Experimental Linguistics and Computational Linguistics, Nankai University**

- **History and scope:**
- **Phonetics Laboratory(1986-), established very early in Mainland China**
- **Social Science Laboratory of Tianjin(2019-)**
- **Innovation Team of Tianjin: Language Intelligence calculation(2024-)**

- **The lab is very open and we collaborate with researchers from a variety of subjects/disciplines.**

URL：https://github.com/EL-CL/

Previous and ongoing researches-1

- 1-A series of studies with ASJP approach
- 2-Studies on senses of basic kernel words in a cross-linguistic perspective
- 3-A serial studies on differentiations between Chinese verbs and nouns
- 4-A series of studies on acoustic distance calculation
- 5-Serial studies on relationship between language and environment

Previous and ongoing researches-1

- 6-Serial Acoustic voice analysis of language/dialects
- 7-Other studies based on large database/dataset and quantitative approaches
- 8-Serial studies based on word-vector calculation(Knowledge Graph)
- 9-Adopting machine learning into linguistic study(classification tasks)
- ……

Previous and ongoing researches-2

- Ongoing researches using deep learning and AI
- 1-Text-To-Speech(EL&CL speakers, specific speakers, comic characters，Chinese dialects, chanting for ancient poem…)
- 2-ASR（IPA recognition, language recognition, foreign accent evaluation, speaker identification…）
- 3-Machine Translation(Manchu-Mandarin automatic translation)
- 4- digital human (Yuen-Ren Chao, Chia-ying Yeh, the first president of Nankai Univ.；Tangwangnese, Tianjin dialect)

I-External differentiations between nouns and verbs: A myth or not?

- **Ran, Qibin（冉启斌） et al.(2023) Phonetic differences between nouns and verbs in their typical syntactic positions in a tonal language: Evidence from disyllabic noun–verb ambiguous words in Standard Mandarin Chinese, *Journal of Phonetics*, 98:101241.**
- **夏全胜、高凯、冉启斌（2023）汉语动名兼类词做名词和动词时的声学语音差异，《语言教学与研究》第2期。**

Journal of Phonetics 98 (2023) 101241

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Phonetic differences between nouns and verbs in their typical syntactic positions in a tonal language: Evidence from disyllabic noun–verb ambiguous words in Standard Mandarin Chinese

Qibin Ran [a,c], Kai Gao [b,d], Yuzhu Liang [a,c], Quansheng Xia [b,c,*], Søren Wichmann [e]

[a] School of Liberal Arts, Nankai University, Tianjin, China
[b] College of Chinese Language and Culture, Nankai University, Tianjin, China
[c] Laboratory of Social Science of Tianjin, Tianjin, China
[d] Institute of Forensic Science, Ministry of Public Security, China
[e] University of Kiel, Germany

ARTICLE INFO

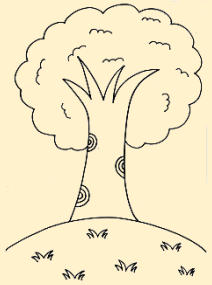
Article history:
Received 8 August 2021
Received in revised form 24 March 2023
Accepted 3 April 2023
Available online 24 April 2023

Keywords:
Noun
Verb
Prosody
Typical syntactic position
Phonetic difference
Information load
Number of syllables

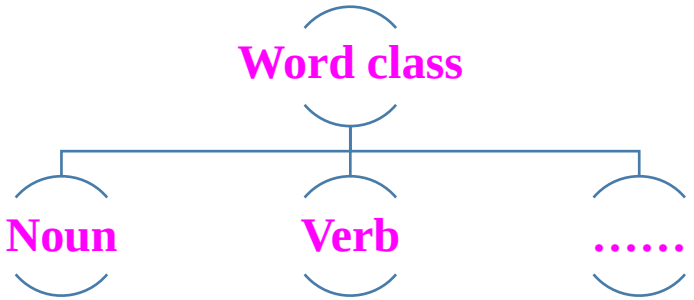
ABSTRACT

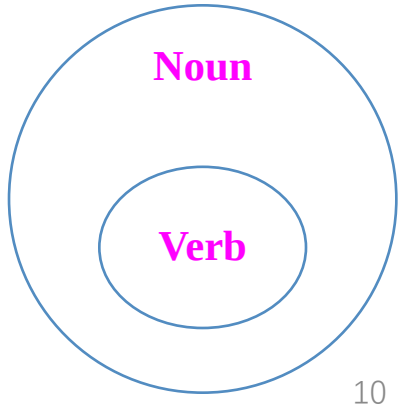
This study investigates how word categories, namely noun and verb, influence acoustic realizations (duration, F0, intensity) in Standard Mandarin Chinese, a language having phonemically distinctive tones and a simple morphological system. Noun-verb ambiguous words were selected and presented in the final positions of typical syntactic contexts in order to avoid the interference of prosodic boundary, syntactic complexity, contextual predictability, tonal environment, F0 range and syllable properties (consonant, vowel, tone, syllable length). Linear mixed models were fitted to duration, and generative additive mixed models were fitted to F0 and intensity. The results showed that phonetic differences between nouns and verbs were still evident in duration, F0 and intensity after lexical frequency, speech rate and some other related factors were taken into consideration in the models. The second syllables of nouns were longer than those of verbs, and both syllables of nouns were higher in F0 and greater in intensity than those of verbs. Since the prosodic boundary, frequency and other factors were controlled for, the phonetic differences between nouns and verbs might be attributed to their differences in information load and number of syllables. This study provided evidence that phonetic differences between nouns and verbs might be driven by the grammatical classes themselves and is not an epiphenomenon of other processes.

- **Nouns and verbs are different in external form/representations**
- **English:**
- **Suffix:**
- **V.→N.: -ness, -ing, -er…**
- **N.→V.: -ize, -en, -ate…**
- **Stress:**
- **re'cord (V.) : 'record(N.)**
- **??:**
- **lock(V.) : lock(N.)**

noun	verb
	
	
	

- **Most previous studies have shown that nouns tend to be longer in duration than verbs in sentences and discourse (Lohmann & Conwell, 2020; Sorensen, Cooper, & Paccia, 1978; Strunk et al., 2020).**
- **Strunk et al. (2020) showed that, among the ten different languages investigated, the grammatical class effect on duration is evident in eight non-tonal languages while it disappears in the two tonal languages, Bora and N‖ng. These findings seem to indicate the noun–verb difference in speech is less likely to be realized in duration in tonal languages compared with non-tonal languages.**
- **Standard Mandarin is a typical tonal language with four tones.**
- **Contour tone**
- **Duration, +F0, intensity**





- **Standard Chinese:**
- **锁suo3 lock(parallel with English)，关guan1；ancient Chinese: 言yan2，雨yu3……**

- **Disyllabic ambiguous words:**
- **爱好 àihào hobby(N)/like(V)**
- **包装 bāozhuāng package(N/V)**
- **工作gong1zuo4，暗示an4shi4，编辑bian1ji，报告bao4gao4，辩论bian4lun4，表现biao3xian4……**

- **Mini pairs:**
- **(1) Noun phrase:**
- **一个+爱好 yīgè + àihào ‘a hobby’**
- **一些+爱好 yīxiē + àihào ‘some hobbies’**
- **几种+爱好 jǐzhǒng + àihào ‘several types of hobbies’**

- **(2) Verb phrase:**
- **能够+爱好 nénggòu + àihào ‘be able to like’**
- **应该+爱好 yīnggāi + àihào ‘should like’**
- **可以+爱好 kěyǐ +àihào ‘can like’**

- **30 mini pairs**
- **28 native Mandarin speakers (14 females, 14 males, mean age = 23 yr, SD = 2.2)**
- **Annotated with Praat: 4 tiers**

- **pitch, duration, intensity**

$$St = 12 * log_2(F0/F_{ref})$$

$$duration\ ratio = D/\mu_d$$

$$z - score = (I - \mu_{st})/\sigma_{st}$$

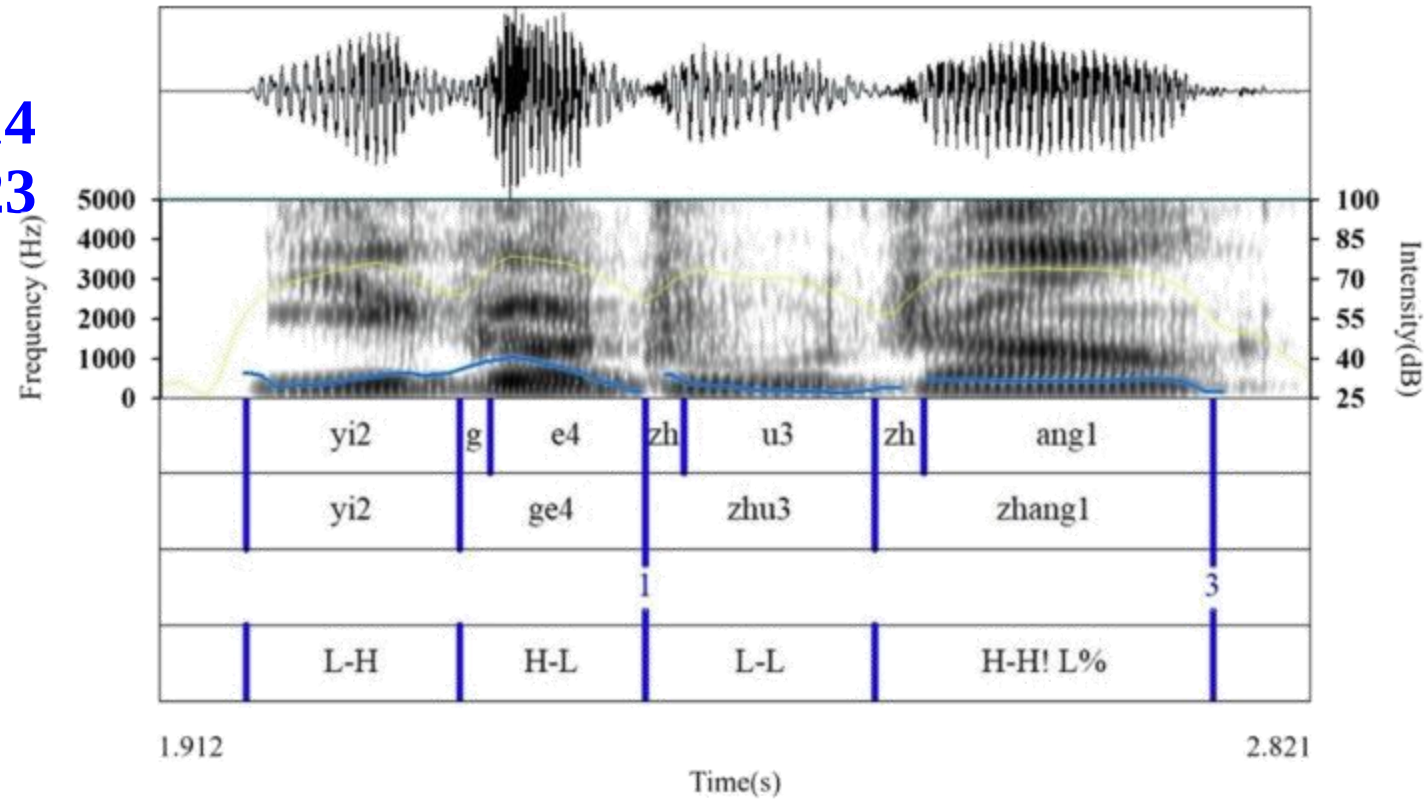


Fig. 1. Example segmentations (the blue and yellow curves displayed illustrate F0 and intensity, respectively).

• Duration:

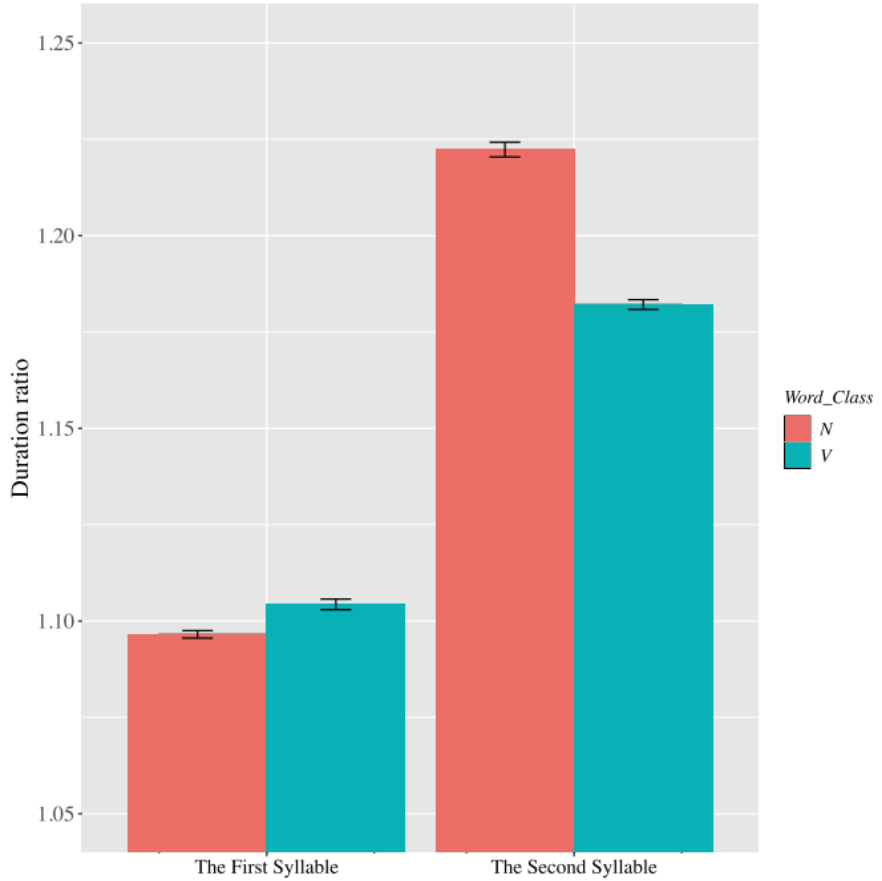


Fig. 2. Duration ratio of nouns (red) and verbs (green). Error bars enclose +/- 1 SE.

Table 3
Model output of LMM model fitted to the duration (duration ratio) (N = 146160; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

R code: duration ~ category * position + frequency ratio + gender + (1| speaker) + (1 + category | word)

Random effects			
	Variance	SD	
Subject (intercept), 28	0.0054	0.074	
Word (intercept), 29	0.0015	0.038	
Category = Verb	0.0011	0.033	
Residual	0.07	0.265	
Fixed effects			
	Coefficient	Std. Error	t
Intercept	1.242e + 00	2.233e-02	55.627***
Category = Verb	−4.274e-02	7.457e-03	−5.732***
Position = 1	−1.238e-01	1.958e-03	−63.247***
frequency ratio	9.719e-04	4.220e-03	0.23
Gender of speaker (male)	−4.139e-02	2.792e-02	−1.482
Category = Verb by Position = 1	4.872e-02	2.769e-03	17.599***

- (histogram) **The second syllables of nouns were longer than those of verbs**

- **Pitch:**

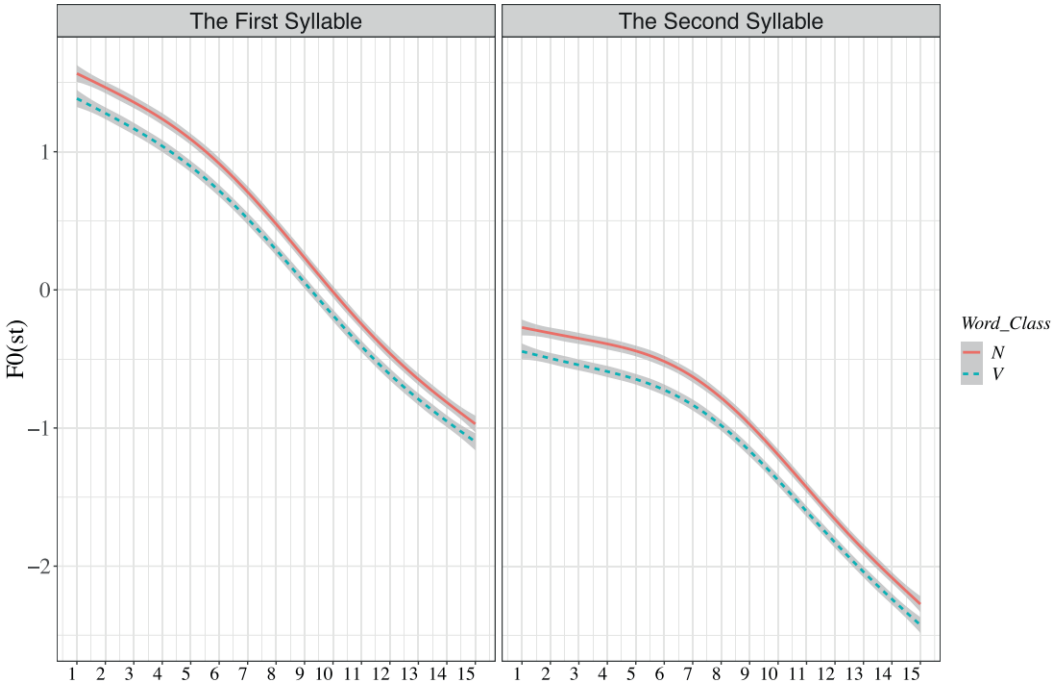


Fig. 3. The GAM smoothers for F0 of nouns (red) and verbs (green). The grey color of F0 contour shows +/- 1 SE.

Table 4

Model output of GAMM model fitted to the F0 (semitone) (N = 146160; $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$).

R code: F0 ~ category + position + frequency ratio + gender + speech rate + s (point, by = category, bs="cr") + s(point, by = position, bs="cr") + s(tone duration, bs="cr") + ti(point, tone duration) + s(point, subject, bs = "fs", xt= "cr", m = 1, k = 15) + s(subject, position, bs ="re") + s(point, word, bs = "fs", xt= "cr", m = 1, k = 15) + s (word, position, bs = "re")

Parametric coeffects	Estimate	Std. Error	t
(Intercept)	−0.826	0.308	−2.68**
Category = Verb	−0.063	0.021	−2.968**
Position = 1	1.438	0.382	3.768***
Frequency ratio	−0.109	0.015	−7.474***
Gender of speaker (male)	−0.024	0.109	−0.217
Speech rate	−0.038	0.16	−0.24
Smooth terms			
	edf	Ref. df	F
s(Point): Category = Noun	1.808	2.249	26.576***
s(Point): Category = Verb	1.001	1.002	55.795***
s(Point): Position = 2	4.564	5.514	19.118***
s(Point): Position = 1	3.025	3.878	3.165*
s(Tone duration)	8.511	8.904	101.91***
ti(Point, Tone duration)	9.669	11.276	26.111***
s(Point, Subject)	64.909	402	10.611***
s(Subject, Position)	43.433	51	22.013***
s(Point, Word)	103.508	433	431.791
s(Word, Position)	51.181	56	231.24***

- **Syllables of nouns were higher in F0 than those of verbs**

- **Intensity:**

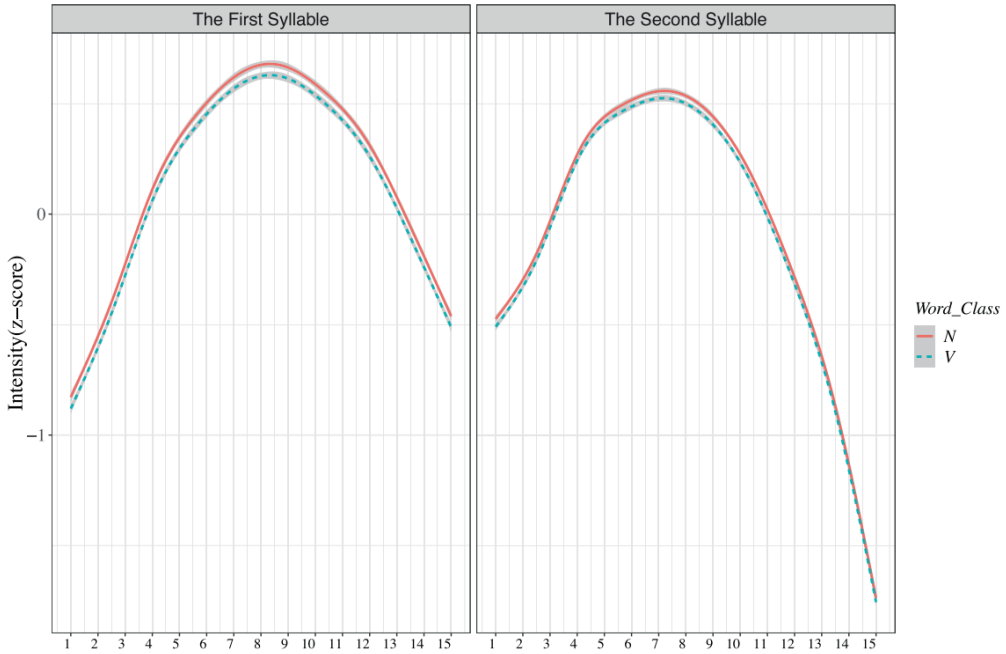


Fig. 4. The GAM smoothers for intensity of nouns (red) and verbs (green). The grey color of intensity contours shows +/- 1 SE.

Table 5

Model output of GAMM model fitted to the intensity (z-score) (N = 146160. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

R code: intensity ~ category + position + frequency ratio + gender + speech rate + s(point, by = position, bs = "cr") + s(duration, bs = "cr") + ti(point, duration) + s(point, subject, bs = "fs", xt = "cr", m = 1, k = 15) + s(subject, position, bs = "re") + s(point, word, bs = "fs", xt = "cr", m = 1, k = 15) + s(word, position, bs = "re")

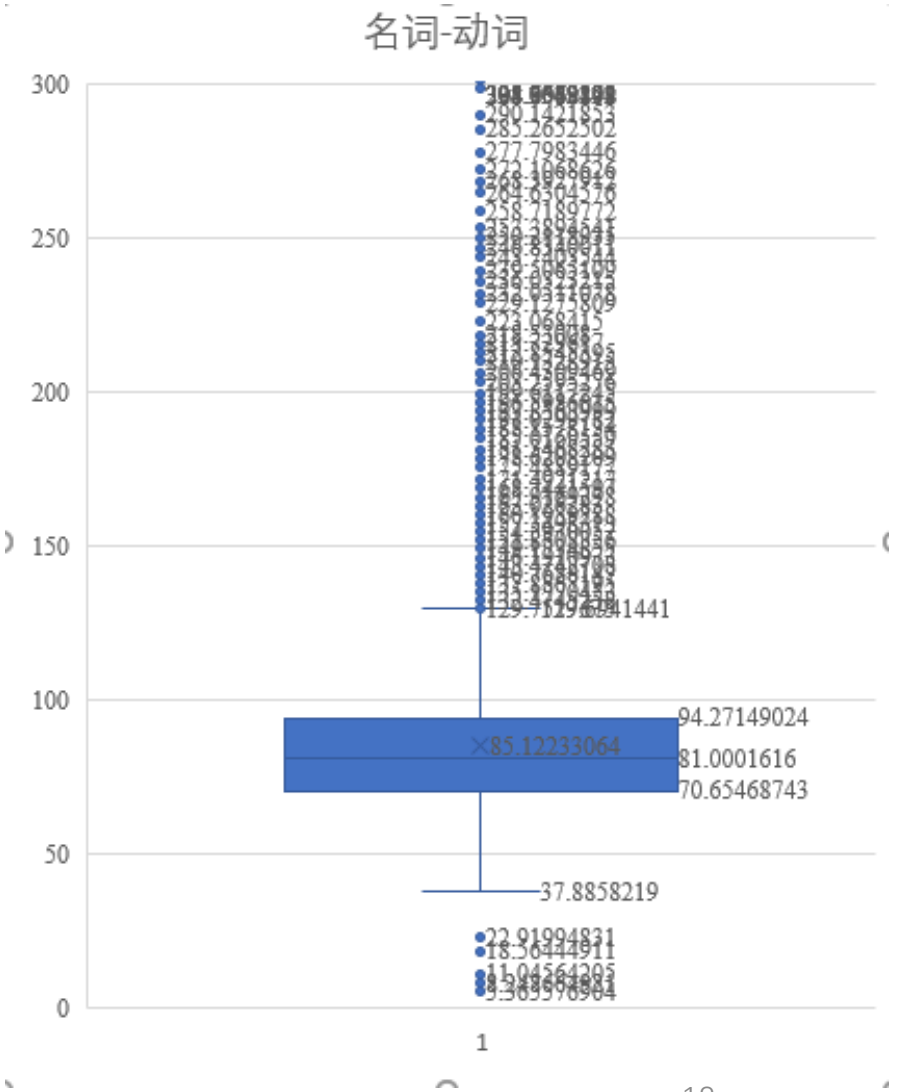
Parametric coeffects

	Estimate	Std. Error	t
(Intercept)	0.0331	0.0784	0.422
Category = Verb	−0.0235	0.00429	−5.473***
Position = 2	−0.228	0.0911	−2.5*
Frequency ratio	−0.159	0.00295	−5.386***
Gender of speaker (male)	−0.0113	0.0384	−0.295
Speech rate	0.2104	0.0415	5.067***
Smooth terms			
	edf	Ref. df	F
s(Point): Position = 1	7.444	7.734	6.47***
s(Point): Position = 2	8.277	8.454	23.21***
s(Duration)	8.856	8.99	339.21***
ti(Point, Duration)	15.407	15.943	58.1***
s(Point, Subject)	203.312	417	219.32***
s(Subject, Position)	49.606	53	86.33***
s(Point, Word)	283.48	433	2498.28*
s(Word, Position)	47.72	56	292.35***

- **Syllables of nouns were greater in intensity than those of verbs**

- **Explanation:**
- **Strunk et al. (2020): nouns are longer than verbs in 8 non-tonal languages**
- **(1)Information load: In Mandarin Chinese, nouns are usually replaced by pronouns or omitted when mentioned a second time. Thus, nouns and verbs in Mandarin should be considered to be inherently different in information load, leading to longer duration of nouns than verbs.**
- **(2)the Smooth Signal Redundancy (SSR) hypothesis (Aylett & Turk, 2004; Turk, 2010): the more predictable or redundant a linguistic element is, the less salient its phonetic realization will be.**
- **(3)When the two grammatical classes are embedded in sentences, nouns need more planning time (Seifart et al., 2018)and are pronounced slower than verbs (Strunk et al., 2020).**
- **(4)Number of syllables: Nouns are typically disyllabic (Ke, 2012; Liu, 1996), while verbs are typically monosyllabic (Chen, 1987; Liu, 1996; Zhang, 1989).**

- **Acoustic distance calculation between nous and verbs**
- **audio1~audio2**
- **20 words*6 times*30 speakers=3,600**
- **It is at a nearly random variations level.**
- **There isn’t essential difference between nouns and verbs??**



• The hierarchy of acoustic distance between different sounds(preliminary)

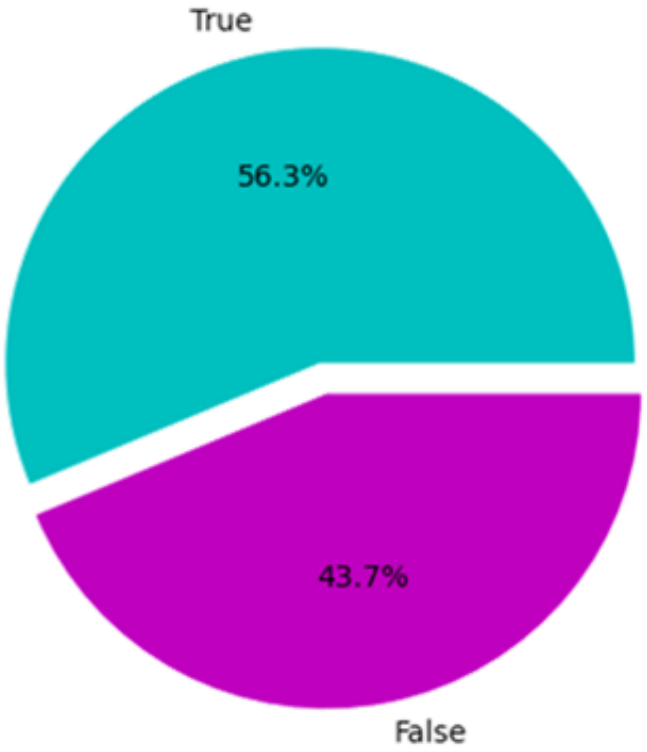
Category	distance（mean）	备注
Sandhi tone3/tone2	65.71	20个词对（仁者-忍者），21名发音人
noun/verb	81.00（双音节）	20对动词-名词（爱好），30名发音人
4 tones	86.79	四声俱全音节，4名发音人
n-l	88.03	55对音节，6名发音人
Neutralized/non-N..d	109.30	14个词对（东西），10名发音人
rhoticized/non-R..d	126.03（双音节）、144.10（单音节）	单、双音节各20对，10名发音人
initials	127.57	声母均值，4名发音人
finals	127.84	韵母均值，4名发音人
vowel	131.05	8个单元音，现代女发音人
inter speaker	158.08	10名发音人，300核心词
Swadesh list1	165.87	207核心词，69种语言
Swadesh list2	206.76	160核心词，50种语言
between languages	215.79	44种语言，各300核心词
between animals	334.68	175种动物，各3种发音

- **Using machine learning to differentiate nouns and verbs**
- **3 acoustic parameters: pitch, duration, intensity**
- **151,200 samples**

- **56~57%:43~42%**

- **A little bit higher than probability level**

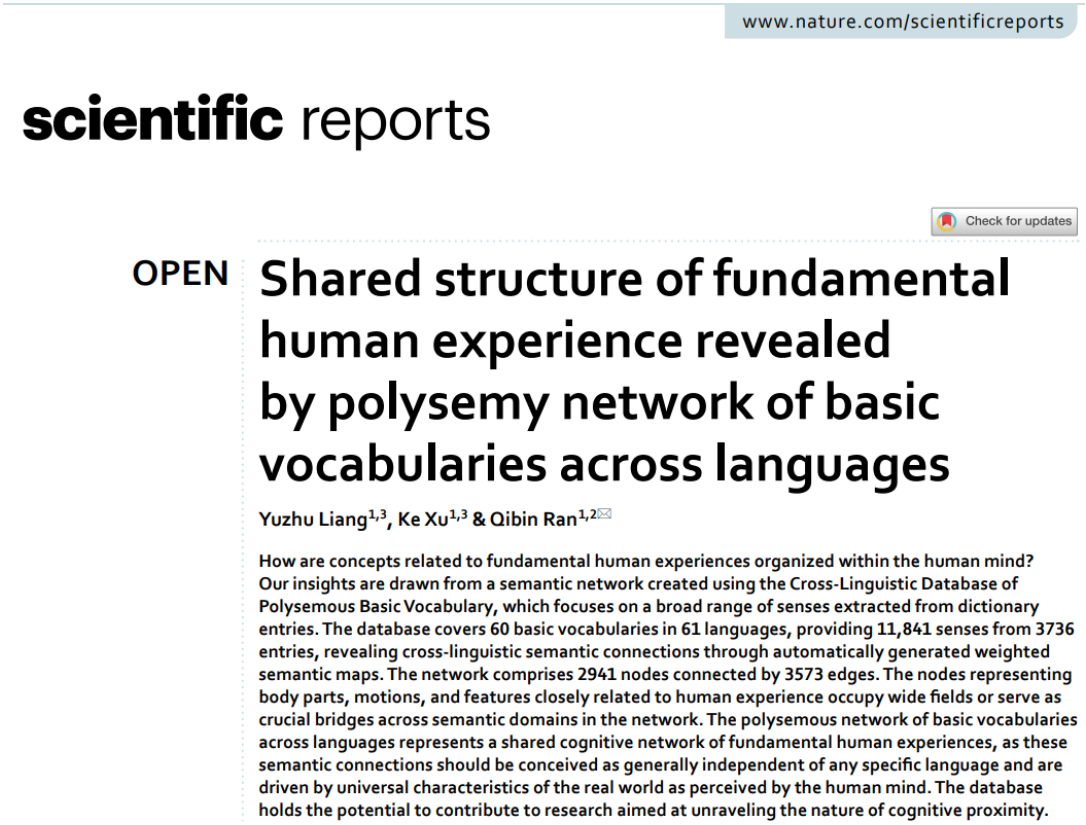
- **So, is there any differences between nouns and verbs??**
- **There isn't essential difference between nouns and verbs?**



- **The external differentiations between nouns and verbs: A myth or not?**

II- Studies on a dataset of cross linguistic sense for kernel words

- **Yuzhu Liang（梁煜珠）, Ke Xu（许可）, Qibin Ran（冉启斌）, (2024). Shared structure of fundamental human experience revealed by polysemy network of basic vocabularies across languages. *Scientific Reports* 14: 5877.**
- **许可、冉启斌（2024）基本核心词词义的跨语言表现与特点——对61种语言数据的分析，《世界汉语教学》第1期**
- **许可、冉启斌、李坤怡（2022）基本核心词词义的数量及其内部关系——对50种语言词义数据的分析探索，《中国语言学研究》第二辑，北京：中国社会科学出版社**
- **许可、李坤怡、冉启斌、黄玮（2022）跨语言基本核心词词义分析——以24种语言基本核心词词义为例，《语文学刊》第5期**


www.nature.com/scientificreports

scientific reports

Check for updates

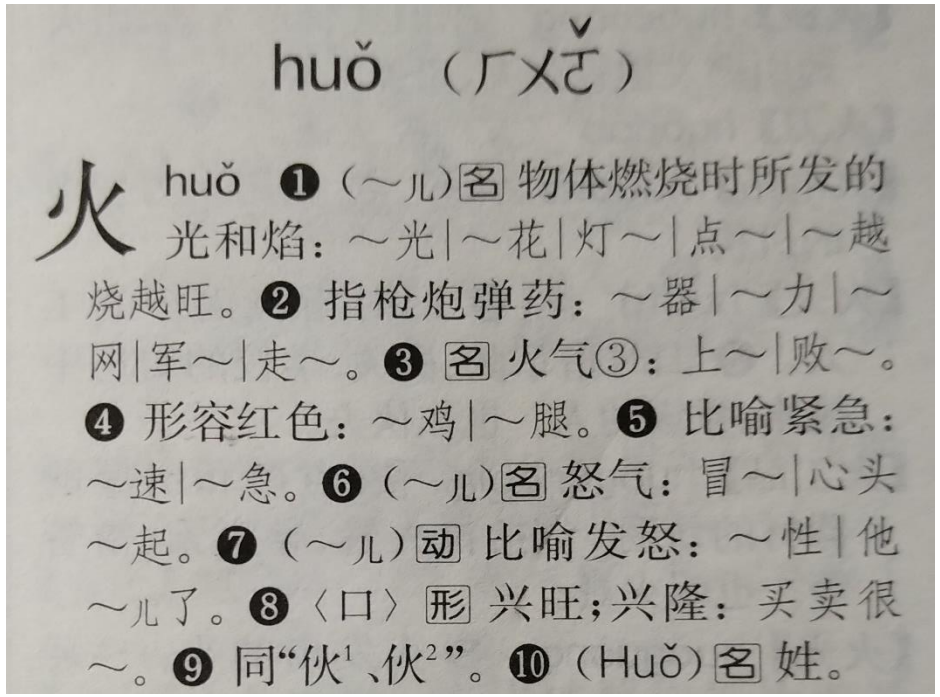
OPEN

Shared structure of fundamental human experience revealed by polysemy network of basic vocabularies across languages

Yuzhu Liang[1,3], Ke Xu[1,3] & Qibin Ran[1,2]✉

How are concepts related to fundamental human experiences organized within the human mind? Our insights are drawn from a semantic network created using the Cross-Linguistic Database of Polysemous Basic Vocabulary, which focuses on a broad range of senses extracted from dictionary entries. The database covers 60 basic vocabularies in 61 languages, providing 11,841 senses from 3736 entries, revealing cross-linguistic semantic connections through automatically generated weighted semantic maps. The network comprises 2941 nodes connected by 3573 edges. The nodes representing body parts, motions, and features closely related to human experience occupy wide fields or serve as crucial bridges across semantic domains in the network. The polysemous network of basic vocabularies across languages represents a shared cognitive network of fundamental human experiences, as these semantic connections should be conceived as generally independent of any specific language and are driven by universal characteristics of the real world as perceived by the human mind. The database holds the potential to contribute to research aimed at unraveling the nature of cognitive proximity.

- **Basic concepts and their derived senses**



huǒ （ㄏㄨㄛˇ）

火 huǒ ❶（～儿）名 物体燃烧时所发的光和焰：～光|～花|灯～|点～|～越烧越旺。❷ 指枪炮弹药：～器|～力|～网|军～|走～。❸ 名 火气③：上～|败～。❹ 形容红色：～鸡|～腿。❺ 比喻紧急：～速|～急。❻（～儿）名 怒气：冒～|心头～起。❼（～儿）动 比喻发怒：～性|他～儿了。❽〈口〉形 兴旺；兴隆：买卖很～。❾ 同"伙[1]、伙[2]"。❿（Huǒ）名 姓。

fire (fīr)

n.

1.
   - **a.** A rapid, persistent chemical change that releases heat and light and is accompanied by flame, especially the exothermic oxidation of a combustible substance: *destruction by fire.*
   - **b.** A specific instance of this change that destroys something: *a house fire.*
   - **c.** A burning fuel: *a cooking fire.*

2. Burning intensity of feeling; ardor or enthusiasm: *a musical performance that had fire.* See Synonyms at **passion**.
3. Luminosity or brilliance, as of a cut and polished gemstone.
4. Liveliness and vivacity of imagination; brilliance: *the fire of an artistic genius.*
5. A severe test; a trial or torment: *went through fire to become a leader.*
6. A fever or bodily inflammation: *tormented by the fire in an infected toe.*
7.
   - **a.** The discharge of firearms or artillery: *heard the fire of cannon.*
   - **b.** The launching of a missile, rocket, or similar ballistic body.
   - **c.** Discharged bullets or other projectiles: *subjected enemy positions to heavy mortar fire; struck by rifle fire.*

8. Intense, repeated attack or criticism: *answered the fire from her political critics.*

v. **fired**, **fir·ing**, **fires**

1 yī ❶ 数 最小的正整数。参看 1212 页〖数字〗。❷ 数 表示同一：咱们是～家人|你们～路走|这不是～码事。❸ 数 表示另一：番茄～名西红柿。❹ 数 表示整个；全：～冬|～生|～路平安|～屋子人|～身的汗。❺ 表示专一：～心～意。❻ 数 表示动作是一次，或表示动作是短暂的，或表示动作是试试的。a）用在重叠的动词（多为单音）中间：歇～歇|笑～笑|让我闻～闻。b）用在动词之后，动量词之前：笑～声|看～眼|让我们商量～下。❼ 数 用在动词或动量词前面，表示先做某个动作（下文说明动作结果）：～跳跳了过去|～脚把它踢开|他在旁边～站，再也不说什么。❽ 数 与"就"配合，表示两个动作紧接着发生：～请就来|～说就明白了。❾ 一旦；一经：～失足成千古恨。❿ 〈书〉副 用在某些词前加强语气：～何速也|为害之甚，～至于此！‖ 注意 "一"字单用或在一词一句末尾念阴平，如"十一、一一得一"，在去声字前念阳平，如"一半、一共"，在阴平、阳平、上声字前念去声，如"一天、一年、一点"。本词典为简便起见，条目中的"一"字，都注阴平。

one (wŭn)

adj.

1. Being a single entity, unit, object, or living being: *I ate one peach.*
2. Characterized by unity; undivided: *They spoke with one voice.*
3.
   a. Of the same kind or quality: *two animals of one species.*
   b. Forming a single entity of two or more components: *three chemicals combining into one solution.*
4. Being a single member or element of a group, category, or kind: *I'm just one player on the team.*
5. Being a single thing in contrast with or relation to another or others of its kind: *One day is just like the next.*
6. Occurring or existing as something indefinite, as in time or position: *He will come one day.*
7. Occurring or existing as something particular but unspecified, as in time past: *late one evening.*
8. *Informal* Used as an intensive: *That is one fine dog.*
9. Being the only individual of a specified or implied kind: *the one person I could marry; the one horse that can win this race.*

n.

1. The cardinal number, represented by the symbol 1, designating the first unit in a series.
2. A single person or thing; a unit: *This is the one I like best. Of her many books, the best ones are the last two.*
3. A one-dollar bill.

pron.

1. An indefinitely specified individual: *She visited one of her cousins.*
2. An unspecified individual; anyone: *"The older one grows the more one likes indecency" (Virginia Woolf).*

• The differences and similarities in the derived senses of basic concepts

stone	shi2 (石)
stone	stone
a piece of stone shaped for a purpose	stone inscription
a hard seed in some fruits	
a unit of weight	
a natural shade of whitish or brownish grey	

go	zou3(走)
English	Chinese
go	go
leave	run
intend to be or do something	move or work
pass into a specified state	become
turn out	leave
be harmonious, complementary, or matching	die
function	visit
be regularly kept or put in a particular place	pass through
use a toilet	leak
	change

listen	ting1(听)
English	Chinese
listen	listen
accept one's opinion	accept one's opinion
pay attention to	at one's convenience
	govern or judge

- **Deriving directions and sense network**

- **If we collect enough senses from a vast amount of languages, we can build a large cross-linguistic sense network.**

- **Investigations:**
- **How many derived senses for different basic concepts in a cross-linguistic perspective?**
- **What are the universals and differences of deriving directions in a large dataset?**
- **Do the derived senses reflect the cognitive significance of different languages?**
- **Are there differences of deriving depth in different languages?**
- **What are the most prominent sense units(including basic senses and derived senses from a cross-linguistic perspective?**
- **What are the most prominent connections in the big network of sense units?**
- **……**

- **Before we started the research, we reviewed a vast amount of literature.**
- **Talmy (1985,2000); Anderson(1978)words for body parts; Comrie(2011) numeral words**
- **伍铁平（1989）、蒋栋元（2002）、温凌云（2007）、蒋绍愚（2008）、黄树先（2012）……**
- **Max-Plank Institute: Wilkins(1996), Fortson IV(2004), Goddard & Wierzbicka(2008, Newman(2009, Urban(2010, Pericliev(2015)……**
- **Miller et al.， WordNet (Miller 1995, Fellbaum 2012)**
- **https://wordnet.princeton.edu/; http://lope.linguistics.ntu.edu.tw/cwn/; http://compling.hss.ntu.edu.sg/cow/**
- **Chinese Concept Dictionary, CCD, 2001**
- **List, Mayer, Terhalle & Urban，CLICS (Rzymski et al. 2020): colexification: https://www.clics.com/**
- **These studies can not fully meet our requirements.**

60 basic kernel words(selected from Swadesh list)

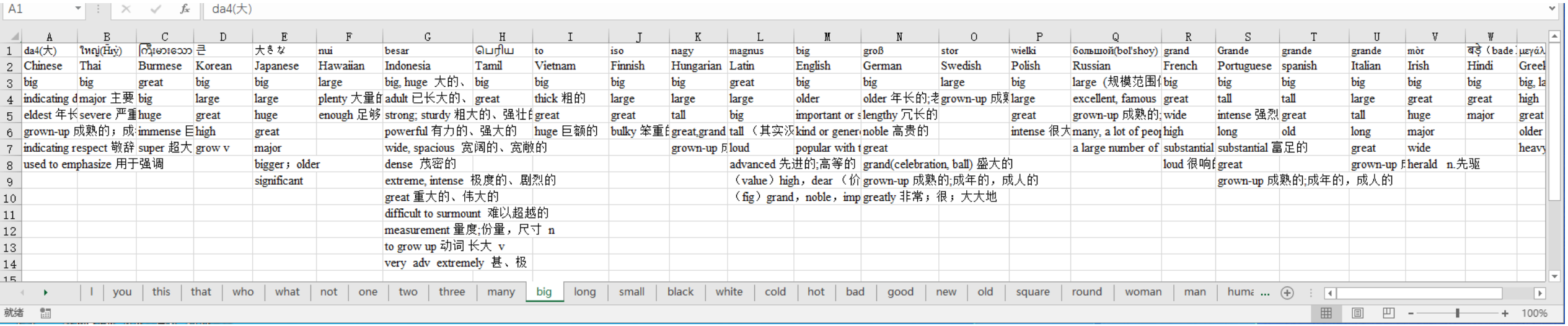
Part of speech	Sense category	word
noun（25）	Natural objects	water、fire、stone、tree、earth/soil、rain、sky、sun、moon、night
	Living being	human、bird、dog、person、man、woman
	Body parts	head、eye、ear、mouth、hand、foot、heart、bone、blood
verb（11）	Action/motion	eat、drink、say、see/look、hear/listen、stand、sleep、give、go、come、die
adj.（14个）	Quantity/size	many、long、big、small
	Property/quality	new、old、good、bad
	Color	black、white
	Shape	square、round
	Temperature	cold、hot
pro-noun（6个）	Personal pronoun	I、you
	demonstrative pronoun	this 、that
	interrogative pronoun	who、what
num.（3）	Numeral	one、two、three
adv.（1）	negator	not

- about 5 years
- 61 languages

- Medium sized authoritative dictionaries

- URL: https://github.com/EL-CL/CLD_Polysemous_Basic_Vocabulary-

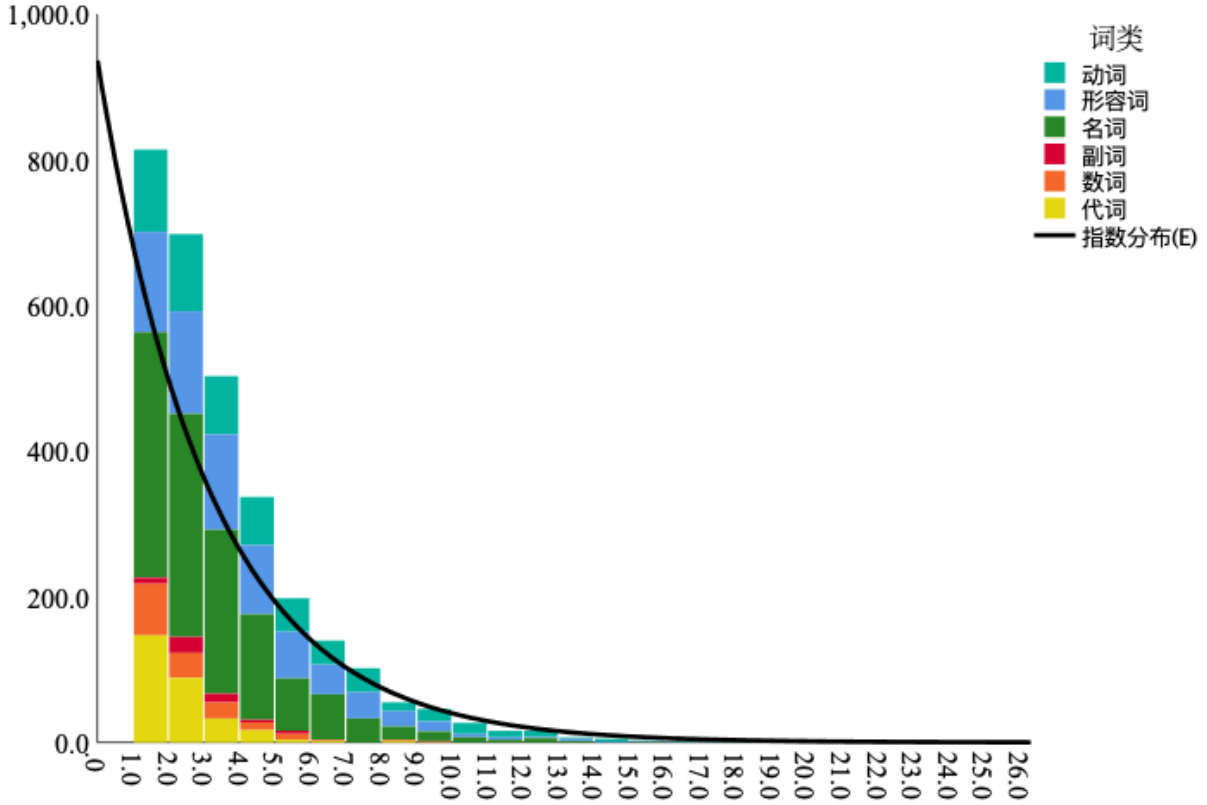
	A	B	C	D	E	F	G	H	I
1	序号	整理时间	language	语言（汉语拼音音序）	词典	作者	出版社	出版时间	
2	1	20190614	Arabic	阿拉伯语	商务阿汉词典				梁增珠
3	2	20190614	Irish	爱尔兰语	Irish-English Dictionary with a	O'Reilly, Edward, and John O'Donc	James Duffy		
4	3	20190614	Polish	波兰语	Collins Polish English Dictior	www.collinslanguage.com			董振渡团队
5	4	20190614	Persian	波斯语					
6	5	20190614	German	德语	Oxford Duden German Diction	J.B.Sykes、Werner Scholze-Stubenrecht		2008	
7	6	20190614	Russian	俄语	现代俄汉双解词典	外语教学与研究出版社		1992	董振渡团队
8	7	20190614	French	法语	FRENCH-ENGLISH Dictionary	Barron's Educational Series Inc., U.S.		2016	董振渡团队
9	8	20190614	Finnish	芬兰语	Isalla Finnish-English dictionary				董振渡团队
10	9	20190614	Korean	韩语	DONG-A'S PRIME KOREAN-	동아（斗山东亚）		1998	董振渡团队
11	10	20190614	Chinese	汉语	现代汉语词典（第7版）	商务印书馆		2016	冉启斌
12	11	20190614	Czech	捷克语	Velký česko-anglický slovník	(英文翻译：Big Czech-English Dict	Leda	2013	薛施惠
13	12	20190614	Latin	拉丁语	Collins Latin Dictionary & Gu	世界图书出版公司北京公司		2013	何佳怡
14	13	20190614	Manchu	满语					马晓筱
15	14	20190614	Burmese	缅甸语	缅汉词典	北京大学东方语言文学系缅甸语教	商务印书馆		董振渡团队
16	15	20190616	Norwegian	挪威语	Norwegian English dictionary		Oslo : Universitetsforlag	1990	许可
17	16	20190614	Portuguese	葡萄牙语	Oxford Portuguese Dictionary	Sinda López Fuentes & Ana Franken	Oxford University Press		董振渡团队
18	17	20190614	Japanese	日语	A STANDARD JAPANESE –	TSUNETA TAKEHARA M. A. Ph. D	東京大阪 賓文館		董振渡团队
19	18	20190614	Swedish	瑞典语	The Standard Swedish-English	Cassell Publishers Ltd Villiers House	首次出版者：Holt Rinehart & Winston		董振渡团队
20	19	20190614	Slovak	斯洛伐克语					薛施惠
21	20	20190615	Swahili	斯瓦西里语	A standard Swahili-English di	Inter-territorial Language (Swahili) C	Nairobi, Kenya ; Oxford	1939	许可
22	21	20190614	Tamil	泰米尔	Tamil-English, English-Tamil (	Victor	Hippocrene Books	2004	董振渡团队
23	22	20190614	Thai	泰语	THAI-ENGLISH DICTIONARY			2003	董振渡团队
24	23	20190614	Uighur	维吾尔语					马晓筱
25	24	20190614	Hungarian	匈牙利语	Hungarian-English Dictionary		匈牙利科学院		董振渡团队
26	25	20190614	Spanish	西班牙语	Oxford Spanish Dictionary	Gulinhetti Jamian、Roy Russell、Cin	OUP Oxford；4	2008	马乐
27	26	20190615	Hebrew	希伯来语					梁增珠
28	27	20190614	Greek	希腊语	Collins Greek-English Diction	HarperCollins Publishers		2003	董振渡团队
29	28	20190614	Hawaiian	夏威夷语	Hawaiian Dictionary	Mary Kawena Pukui; Samuel H. Ells	University of Hawai'i Pre	1986	董振渡团队
30	29	20190614	Italian	意大利语	Oxford-Paravia Italian Dictionary		Oxford University Press	2010	董振渡团队
31	30	20190614	Hindi	印地语	Hindi-English English-Hindi Dictionary		HIPPOCRENE BOOKS, INC.171 Madison A		董振渡团队
32	31	20190614	Indonesian	印度尼西亚语	A Comprehensive INDONESIA	Allan M. Stevens and A. Ed. Schmidt	Ohio University Press		董振渡团队
33	32	20190614	English	英语	Oxford Intermediate Learner's I	商务印书馆		2010	马小芳
34	33	20190614	Vietnamese	越南语	现代越汉词典（第2版）	雷航 李宝红 主编	外语教学与研究出版社		董振渡团队
35	34	20190820	Scottish Gaelic	苏格兰盖尔语	The Gaelic-English dictionary	Colin Mark.	London ; New York : Ro	2004	许可
36	35	20200903	Turkish	土耳其语	土耳其语汉语词典	周正清 周近金 主编	商务印书馆	2008	王帅
37	36	20200903	Mongolian	蒙古语	新蒙汉词典	《新蒙汉词典》编委会编	商务印书馆	1999	王帅
38	37	20200926	Malay	马来语	Kamus umum bahasa malaysia	Yang Kuil Yee,Chan Meow Wah	The World Book Co.,Sdn	1988	王帅
39	38	20201020	Urdu	乌尔都语	乌尔都语汉语词典	孔菊兰 主编	高等教育出版社	2014	王帅
40	39	20201009	Lingao	临高语	临高汉语词典	中国社会科学院民族研究所主编 刘	四川民族出版社	2000	许可
41	40	20201020	Zhuang Language	壮语	壮汉词汇	广西壮族自治区少数民族语言文字	广西民族出版社	1984	许可
42	41	20201027	Bouyei	布依语	布依汉词典	中国社会科学院民族研究所主编 吴	民族出版社	2002	许可
43	42	20201104	Li	黎语	黎汉词典	中国社会科学院民族研究所主编 郑	四川民族出版社	1993	许可
44	43	20201208	Khmer	高棉语	Cambodian-English Dictionary	Robert K. Headley, Ruth Chim, and	Dunwoody Press	1997	许可
45	44	20201216	Lao	老挝语	Lao-English Dictionary	William L. Patterson and Mario E. Sa	Dunwoody Press	1995	丁俊
46	45	20201212	Lithuanian	立陶宛语	Lithuanian dictionary	Bronius Piesarskas, Bronius Sveecv	Routledge	1995	王帅
47	46	20201220	Dutch	荷兰语	Prisma Handwoordenboek Eng	Prue Ganjano, Fokko Veldman	Uitgeverij Unieboek	2010	许可
48	47	20201220	Hausa	豪萨语	A Hausa-English Dictionary	Paul Newman	Yale University Press	2007	王帅
49	48	20210323	Malagasy	马尔加什语	A new Malagasy-English dicti	J. Richardson	The London Missionary	1885	黄玮
50	49	20210107	Danish	丹麦语	Danish Dictionary:Danish-Eng	Anna Garde	Routledge	1995	王帅
51	50	20210602	Kyrgyz	柯尔克孜语	Kyrgyz-English Dictionary	Karl A. Krippes	Dunwoody Press	1998	王帅
52	51	20210616	basque	巴斯克语	CBS-Morris English-Basque/ E	Morris Academy Press	Mikel Morris	2010	千爽
53	52	20210609	Telugu	泰卢固语	A TELUGU-ENGLISH DI	New York: Oxford University Press	Gwynn, J. P. L. (John P	1991	王帅
54	53	20210719	icelandic	冰岛语	Íslensk-ensk orðabók / Con	Sverrir Hólmarsson; Sanders, Ch	Munn	1989	史晴琳
55	54	20210610	Armenian	亚美尼亚语	A Comprehensive Dictionary	Mesrob G. Kouyoumdjian	Atlas Press	1970	千爽
56	55	20210613	Croatian	克罗地亚语	Veliki hrvatsko-engleski rjeeni	Z. Bujas	Globus	2008	王帅
57	56	20210726	Bengali	孟加拉语	Samsad Bengali-English dic	Biswas, Sailendra	Sahitya Samsad	2000	史晴琳
58	57	20210628	maori	毛利语	A Dictionary of the Maori Lan	Herbert W. Williams	Government Printer	1957	史晴琳
59	58	20210906	Pashto	普什图语	A Dictionary of the Pukhto,	Raverty, H. G.	Literary Licensing, LLC	1867	千爽
60	59	20210906	Shan	掸族语	http://sealang.net/shan/dictionary.htm				史晴琳
61	60	20210906	Pali	巴利语	http://www.ahandfalo-flenves.org/documents/Concise%20Pali%20English%20Dictionary_Buddhadatta.pdf				史晴琳
62	61	20210906	Marshallese	马绍尔语	Marshallese-English Dictionary	Takaji Abo、Byron W. Bender, Alfr	University of Hawaii Pre	1977	千爽



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	da4(大)	ใหญ่(hày)	ကြီးသော	큰	大きな	nui	besar	பெரிய	to	iso	nagy	magnus	big	groß	stor	wielki	большой(bol'shoy)	grand	Grande	grande	grande	mòr	बड़े (bade	μεγάλ
2	Chinese	Thai	Burmese	Korean	Japanese	Hawaiian	Indonesia	Tamil	Vietnam	Finnish	Hungarian	Latin	English	German	Swedish	Polish	Russian	French	Portuguese	spanish	Italian	Irish	Hindi	Greek
3	big	big	great	big	big	large	big, huge 大的、	big	big	big	big	great	big	big	large	big	large (规模范围(	big	big	big	big	big	big	big, la
4	indicating d	major 主要	big	large	large	plenty 大量的	adult 已长大的、	great	thick 粗的	large	large	large	older	older 年长的;老	grown-up 成熟	large	excellent, famous	great	tall	tall	large	great	great	high
5	eldest 年长	severe 严重	huge	great	huge	enough 足够	strong; sturdy 粗大的、强壮的	great	great	tall	big	important or s	lengthy 冗长的		great	grown-up 成熟的;	wide	intense 强烈	great	tall	huge	major	great
6	grown-up 成熟的；成	immense 巨	high	great		powerful 有力的、强大的	huge 巨额的	bulky 笨重的	great,grand	tall （其实汉	kind or gener	noble 高贵的			intense 很大	many, a lot of peo	high	long	old	long	major		older
7	indicating respect 敬辞	super 超大	grow v	major		wide, spacious 宽阔的、宽敞的			grown-up 成	loud	popular with t	great				a large number of	substantial	substantial 富足的		great	wide		heavy
8	used to emphasize 用于强调			bigger；older		dense 茂密的				advanced 先进的;高等的		grand(celebration, ball) 盛大的					loud 很响的	great		grown-up 成	herald n.先驱		
9				significant		extreme, intense 极度的、剧烈的				（value）high，dear （价		grown-up 成熟的;成年的，成人的						grown-up 成熟的;成年的，成人的					
10						great 重大的、伟大的				（fig）grand，noble，imp		greatly 非常；很；大大地											
11						difficult to surmount 难以超越的																	
12						measurement 量度;份量，尺寸 n																	
13						to grow up 动词 长大 v																	
14						very adv extremely 甚、极																	
15																							

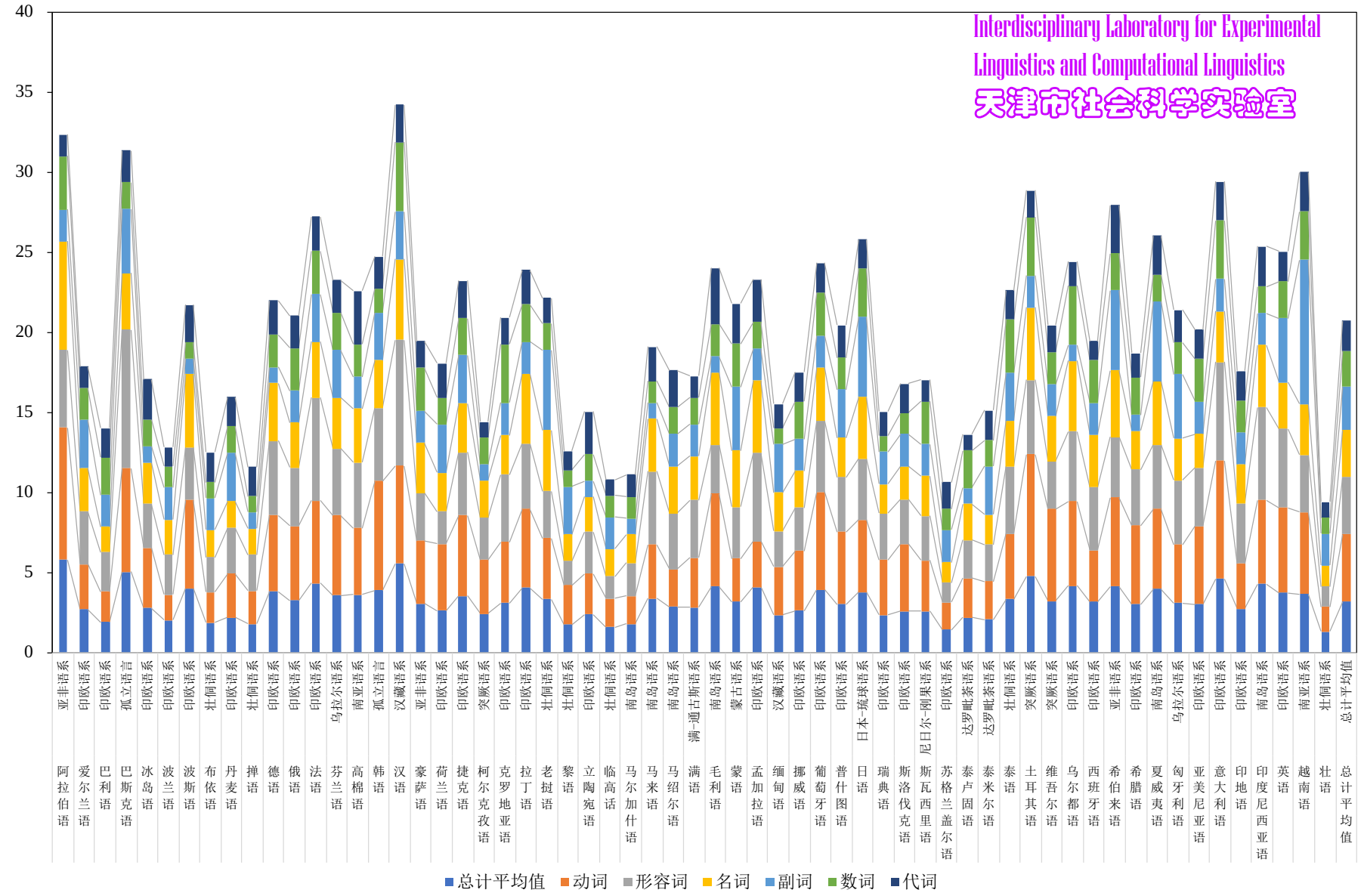
Statistics of the dataset

- The minimum number of senses: 1; the maximum number of senses: Arabic: eye, 22
- the maximum of sense in different word classes:
- Verbs: 20, adjective:15, adv.(not): 9, pronoun:9, numeral:8
- 义项数量分布广，但多数集中在较低值上
- exponential distribution



Stacked histogram of sense number for different word classes

- 词类、语言对义项数量均值有显著影响；不同语言中不同词类的义项数量大小可能不同（F=69.702，df=5，p<0.001；F=3.022，df=60，p<0.001；交互作用：F=1.32，df=300，p<0.001）



Stacked histogram of sense number for different word classes

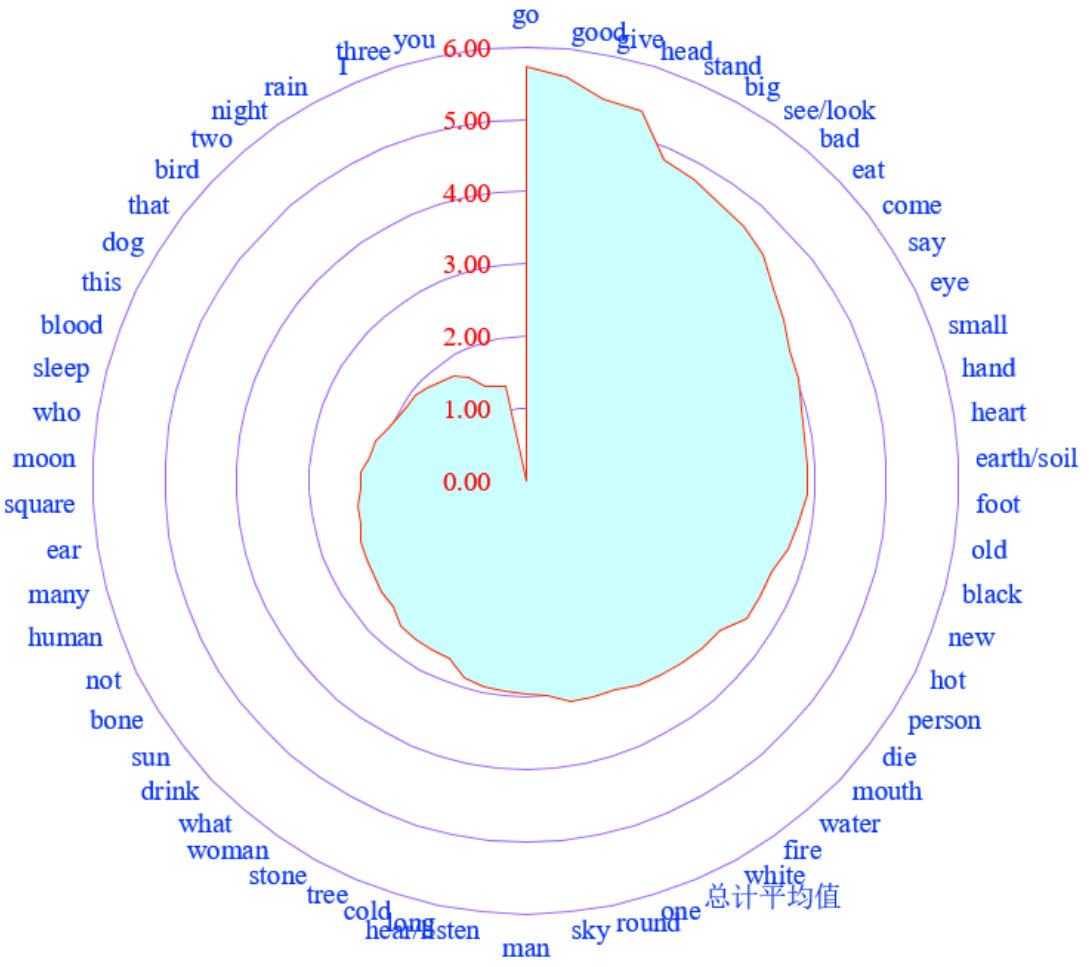
- General descending order: verb > adjective > noun > adverb > numeral > pronoun
- Chinese: adjective > verb> noun > other classes. (including Indonesian, Hindi, etc.)
- Korean: verb > adjective > noun > other classes. (Japanese, Greece, etc.)
- Hawaiian, Latin: verb > noun > adjective > other classes.
- Thai, Burmese, Irish, Tamil: almost equally distributed in different word classes.
- Pronoun
- Lithuanian:I:1, you:1, this:1, that:2, who:8, what:3
- Who→“whoever，anybody（无论谁，任何人）”、“someone（某人）”、“something（某事）”、“anything（任何事）”、what、that、“the thing or things that（……的事）”）
- negative adverbs
- Japanese, Hawaiian, Vietnamese: not→’without’ ‘never’ ‘avoid’ ‘empty’ ‘gratis’

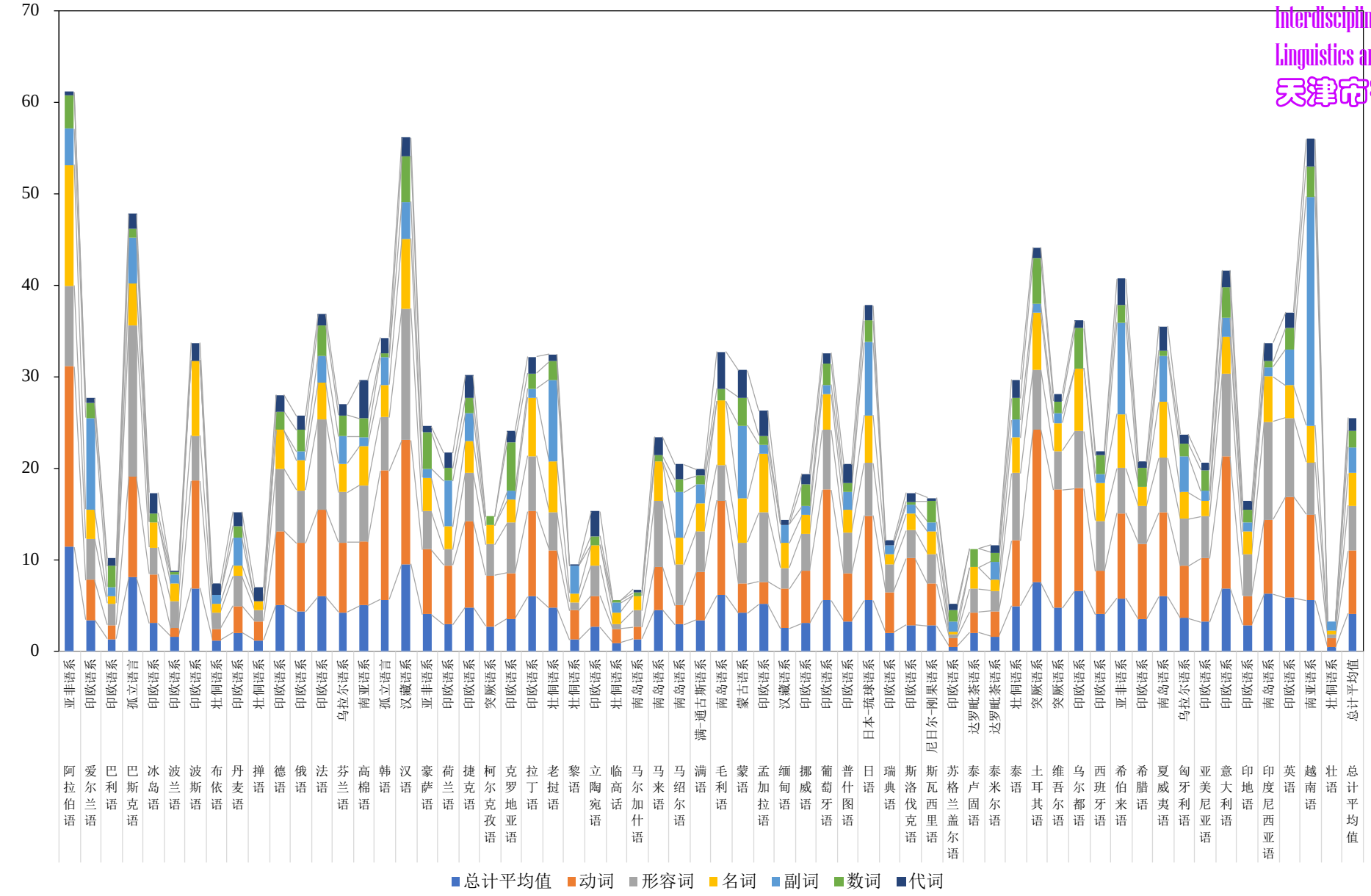
- **The degree of dispersion in the distribution of the number of senses**

	Sum.	verb	adj.	noun	adv.	numeral	pronoun
Mean	3.17	4.14	3.56	2.99	2.58	2.12	1.86
S. D.	2.45	3.24	2.39	2.18	1.43	1.44	1.21
Maximum	21.00	20.00	15.00	21.00	9.00	8.00	9.00
Maximum estimated within the general range	10.51	13.87	10.73	9.53	6.87	6.45	5.49
Minimum	1.00	1.00	1.00	1.00	1.00	1.00	1.00

- **Radar chart of sense mean values of the 60 words**
- **Cross-linguistic data**
- **从基本概念来看**
- **最容易产生出引申意义的基本概念**
- **最容易发生认知延伸的基本概念**

- **具体性？“tree”、“square”义项数量较少。“head”义项数量多**
- **基本性？eat与drink**
- **笼统性？cold等在不少语言中用多种不同级别的概念表示不**



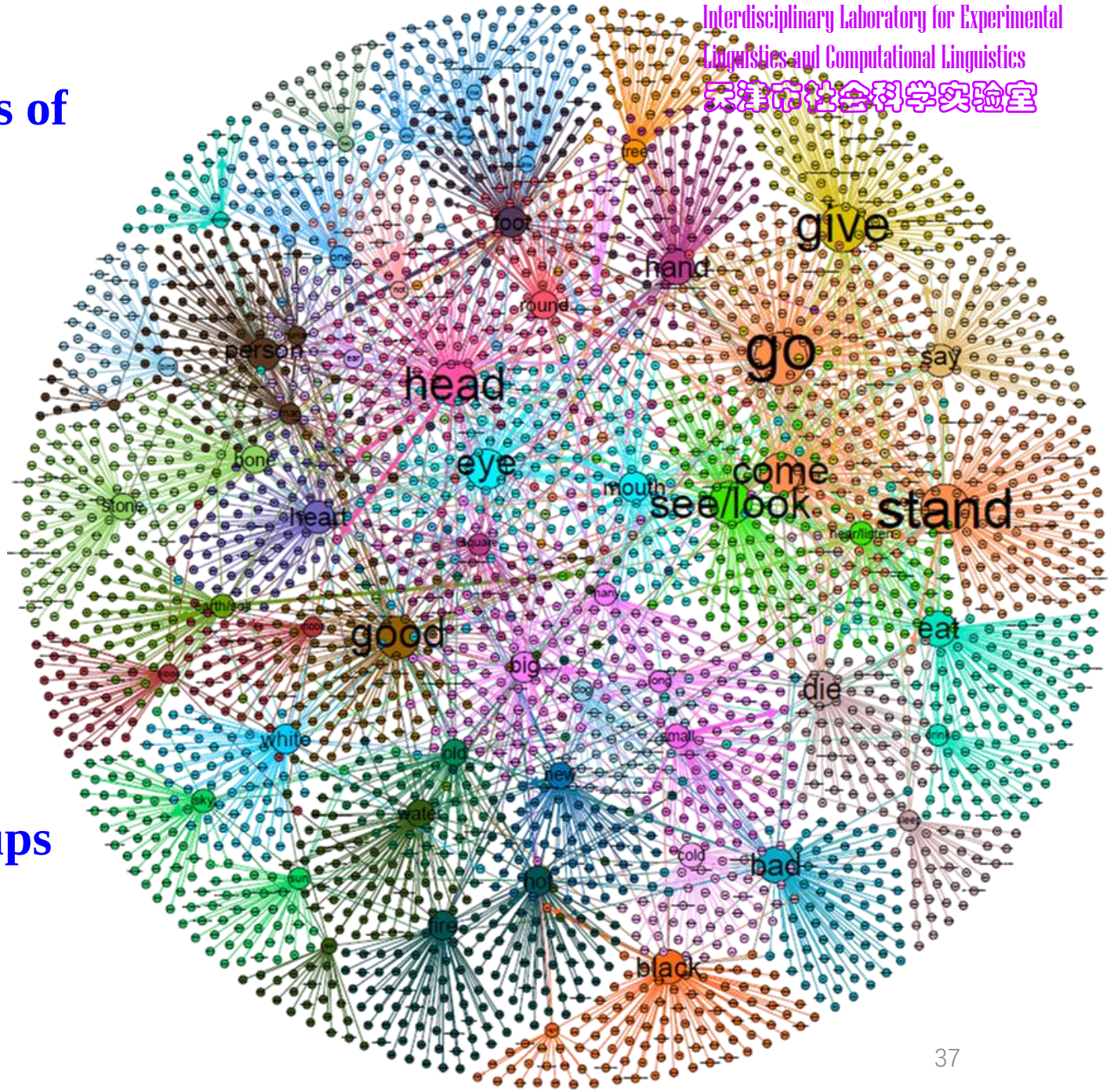


- **Stacked histogram of average extension ability in 61 languages**

- The cross linguistic network atlas of sense unit
- Gephi

- 60 words, in 61 languages
- Sense: 11,841
- Sense unit: 2,941
- Edge(weight) with arrow:3573

- Node size: edges linking a node
- Node color: 34 modularized groups
- Full connection network of 61 languagesdataset.pdf



- **There is extremely rich information in the network atlas.**
- **The 10 most prominent basic concepts:**
- **Go, good, give, head, stand, big, see/look, bad, eat, come, are important basis for cognitive derivation in kernel words, it frequently undertakes the starting point and foundation of derived senses.**
- **The basic concepts that are most prone to cognitive derivation are not very "substantive" concepts, but rather those with "abstract" senses.**
- **"substantive" word: head (indicating that "head" is very important in linguistic cognition of human?)**
- **good: Pollyanna Hypothesis; give:??; stand……**

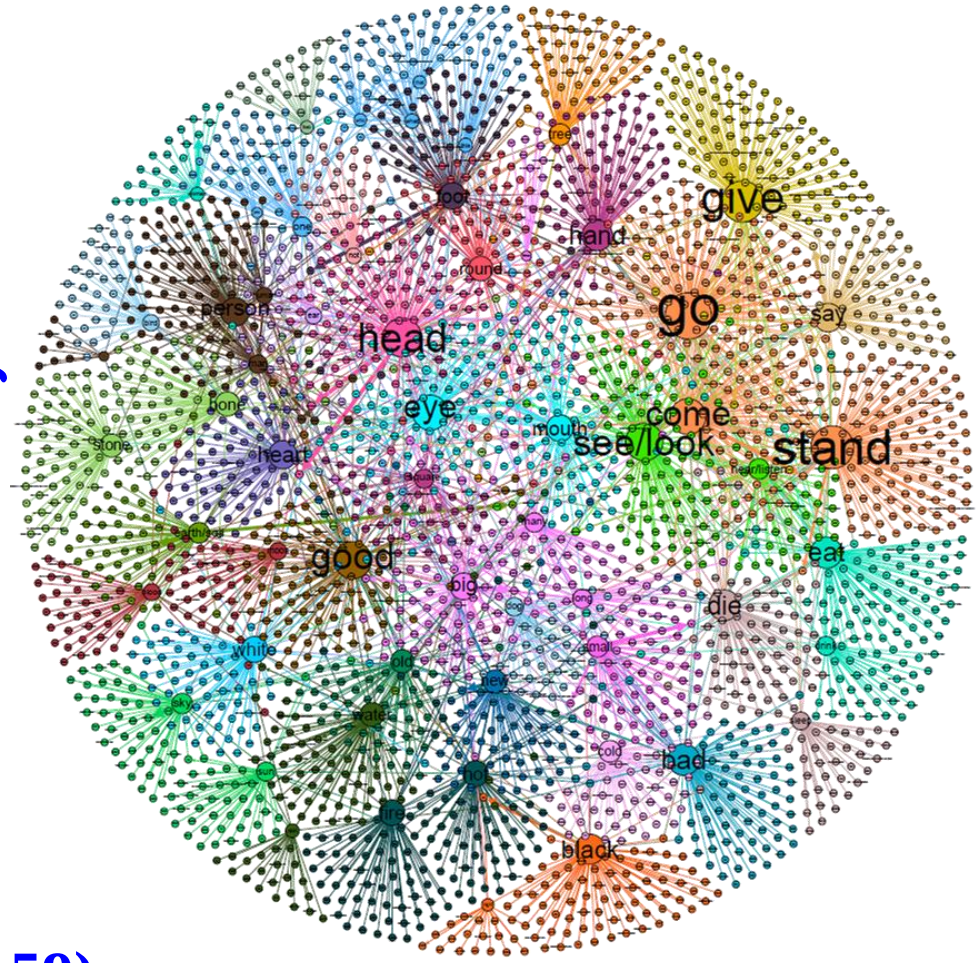
- 20 most frequent connections in deriving senses:
- not → no
- big → great, superior
- sky → paradise, heaven
- head → chief, leader
- foot → leg
- small → little
- head → top
- woman → female
- new → fresh
- earth/soil → ground
- come → reach, arrive
- say → talk, speak
- say → tell
- earth/soil → land
- sun → sunshine, sunlight
- black → dark
- bad →debased, evil, wicked
- hand → arm
- man → husband
- moon → month

- 20 most rare connections in deriving senses:
- person → capital towns Latin
- person → chief, leader
- person → civil rights
- person → flesh
- see/look → have a look
- person → head
- person → life
- see/look → look into
- person → mask Latin
- person → principle
- person → principle money Latin
- person → source of rivers Latin
- see/look → spot
- person → substance, chapter
- person → the world Latin
- person → top Latin
- die → be exhausted
- die → breathe Icelandic
- die → crack
- die → explode Icelandic
- die → split

- The 60 kernel words can be clustered into 10 groups.
- (different clustering methods were used)

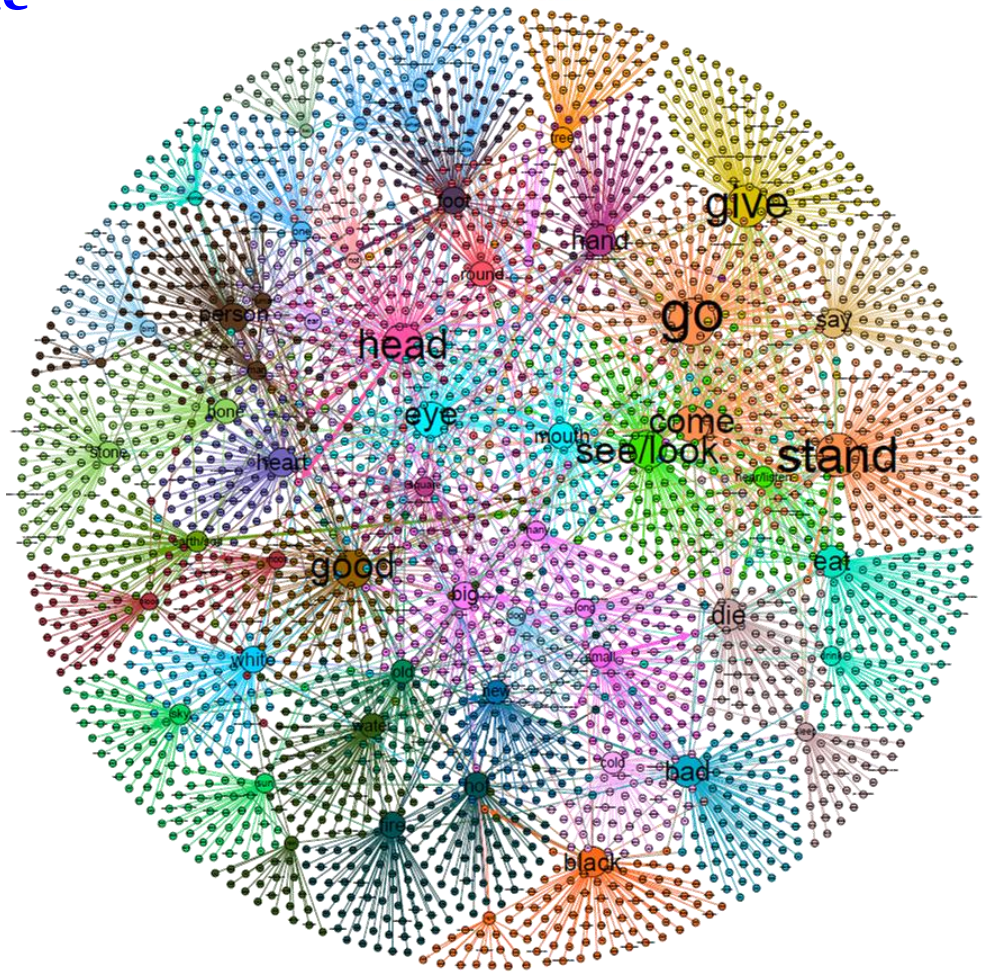
Module	Basic kernel words
1	stand、give、go、hear/listen、see/look、come、say、sleep、die
2	rain、sky、eat、drink、water、moon
3	hand、round、square、ear、earth/soil、white
4	head、mouth、eye
5	stone、tree、heart、foot、cold、bone
6	blood、bird、man、person、you、I、human、woman
7	big、many、long、good、old、not、three
8	dog、bad、black、small、night
9	one、two、that、what、this、who
10	sun、hot、fire、new

- **Convergence and divergence in sense deriving:**
- **Convergence type:**
- **not；big；sky；head；foot；small……**
- **Divergence type:**
- **go；good；give；stand；eat；eye……**
- **The most important hubs or pivots: hand、head、person、die**



- **Distance calculation between 60 basic concepts**
- **In 60 basic concepts, means of shortest path:**
- **Words with maximum means: I(5.00), two(4.71), night(4.63), rain(4.51), ear(4.25)…**
- **Words with minimum means: good(2.56), head(2.59), person(2.63), big(2.64), come(2.76)…**

- **Random traceability**
- **From a random unit, finding the next basic node (60 basic concept), and repeat the process continuously: what basic concepts will be traced/linked easily or frequently? And, what basic concepts will be seldom traced/linked?**
- **PageRank**
- **Most frequently traced/linked:**
- **go(0.018437), give(0.017146), stand(0.016377), head(0.015042), good(0.015042)…**



- **Seldom traced/linked:**
- **You(0.001392, I(0.001553), three(0.001558), who(0.002489, that(0.002760)…**

- **III- Acoustic distance analysis on linguistic elements/language/dialect**

- **Jun Ding1†, Tianheng Wang1†, Ke Xu1, Aleksandr Mitkov1, Lining Wang2*, Gang Peng3*, Qibin Ran(to be published), Acoustic distances of 300 core words imply Indo-European phylogeny and chronology.**

Submitted Manuscript: Confidential
Template revised November 2023

Title: Acoustic distances of 300 core words imply Indo-European phylogeny and chronology

Authors: Jun Ding[1]†, Tianheng Wang[1]†, Ke Xu[1], Aleksandr Mitkov[1], Lining Wang[2]*, Gang Peng[3]*, Qibin Ran[1,4]*

Affiliations:

[1]School of Liberal Arts, Nankai University; Tianjin, 300071, China.

[2]Center for the Protection and Research of Language Resources of China, Beijing Language and Culture University; Beijing, 100083, China.

[3]Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University; Kowloon, Hong Kong SAR.

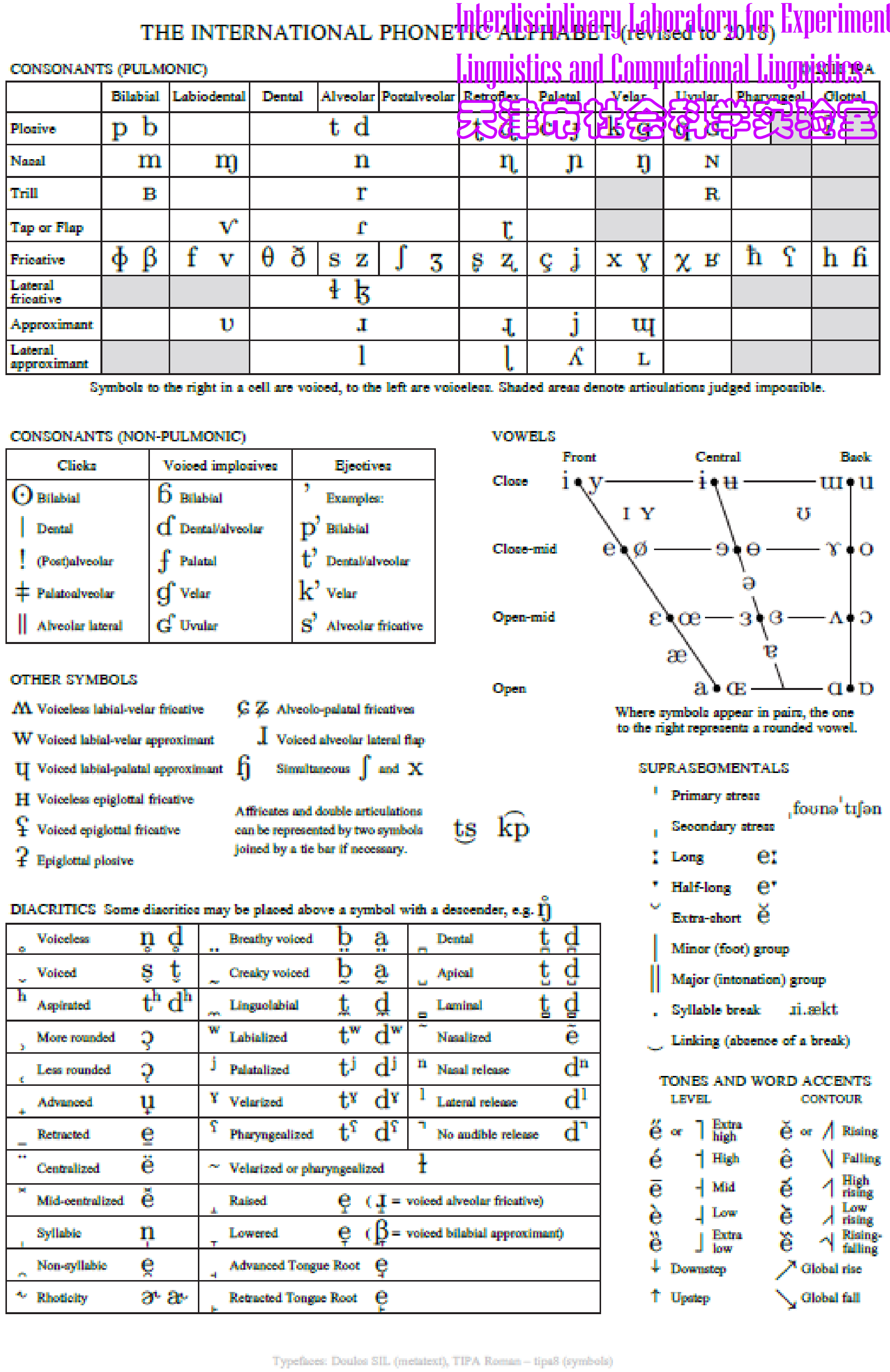
[4]Laboratory of Social Science of Tianjin, Nankai University; Tianjin, 300071, China.

*Corresponding author. Email: wangln@blcu.edu.cn (L.W.); gpeng@polyu.edu.hk (G.P.); ranqibin@126.com (Q.R.)

†These authors contributed equally to this work.

Abstract: Language offers crucial clues to human history; yet can these clues be directly gleaned from speech sounds? Unlike previous investigations into language chronology and phylogeny that primarily rely on historical comparisons, this study explores an acoustic approach to examine the history of Indo-European languages by leveraging synthesized sound samples of 300 core concepts in 42 languages. The acoustic distance-based clustering of these languages aligns well with established classifications, and the acoustic distances effectively reflect language divergence ages. Acoustic results imply that the origin of Indo-European languages (except Anatolian and Tocharian) dates back to 5700–5000 BP. This groundbreaking acoustic approach not only shows potential to unravel language chronology and phylogeny, but also provides fresh methods and perspectives on human prehistory.

One-Sentence Summary: Synthesized sounds from 42 Indo-European languages disclose the age, classification, and origin date of these languages.



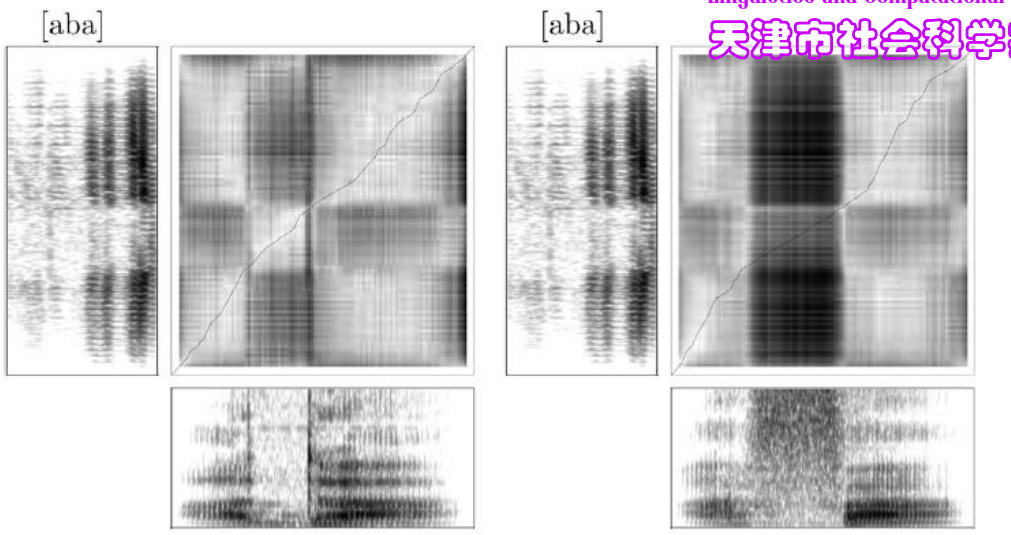
- **Difference between sounds**
- **Distances between sounds are equal or not?**

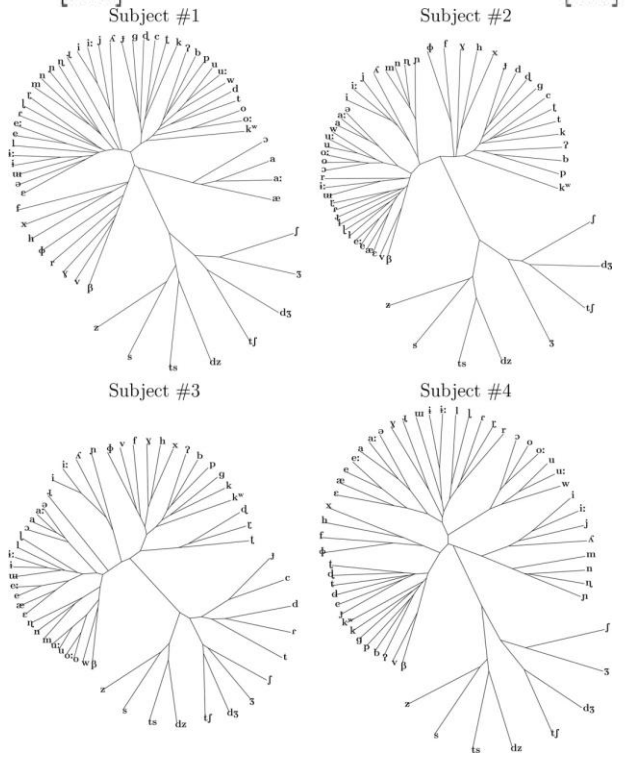
- **D (s,m) > D (ɛ,æ)**
- **D (p,s) > D (n,l)**
- **……**

- **Itakura(板仓1965): DTW**
- **Holmes and Holmes (2001) : (DTW，dynamic time warping algorithm)**
- **Mielke(2012)：VXV**

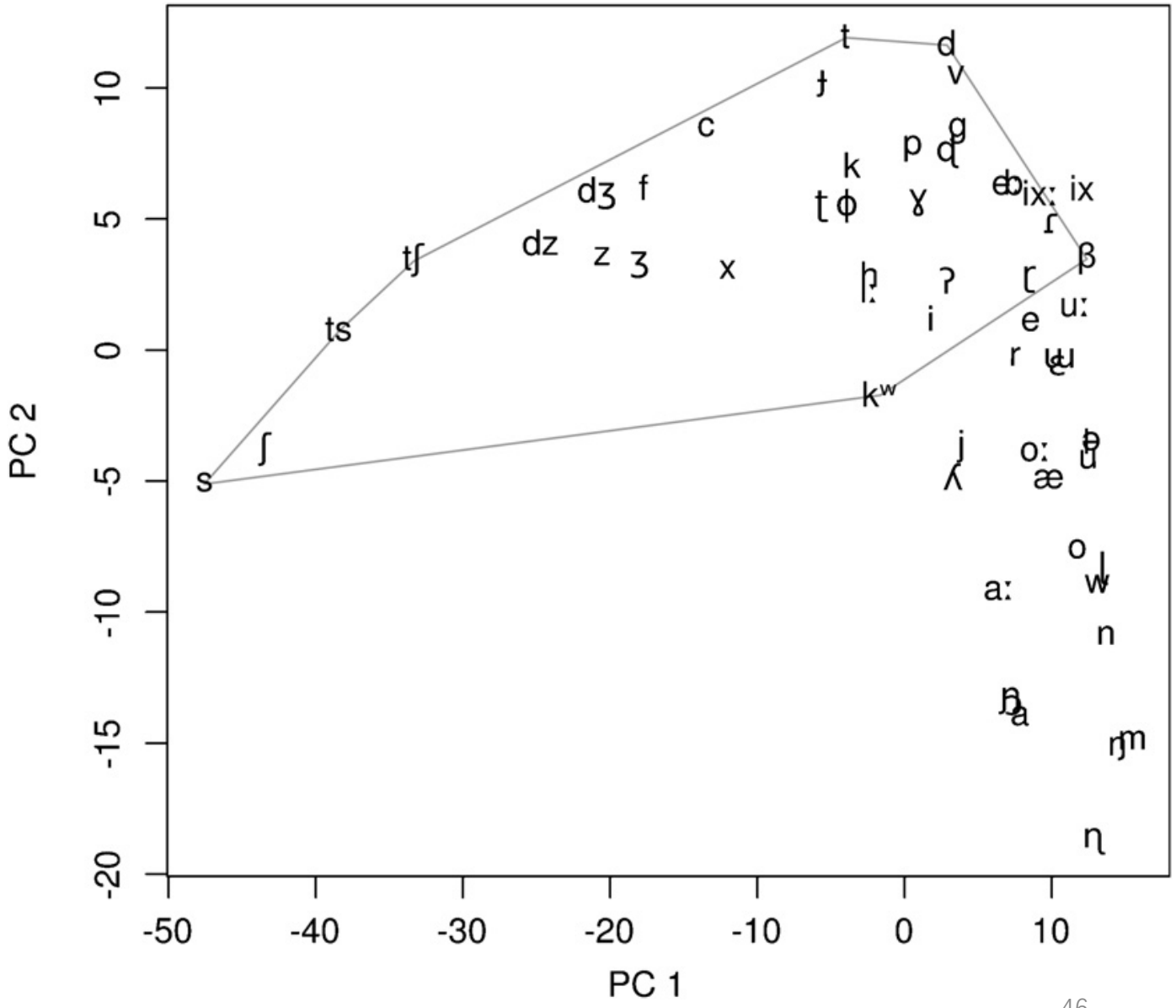
- **J. Mielke(2012)A phonetically based metric of sound similarity, *Lingua*122(2).**

- **Mielke(2012): 58 phonemes (vowel: 17, consonant: 41)**
- **Phylip**

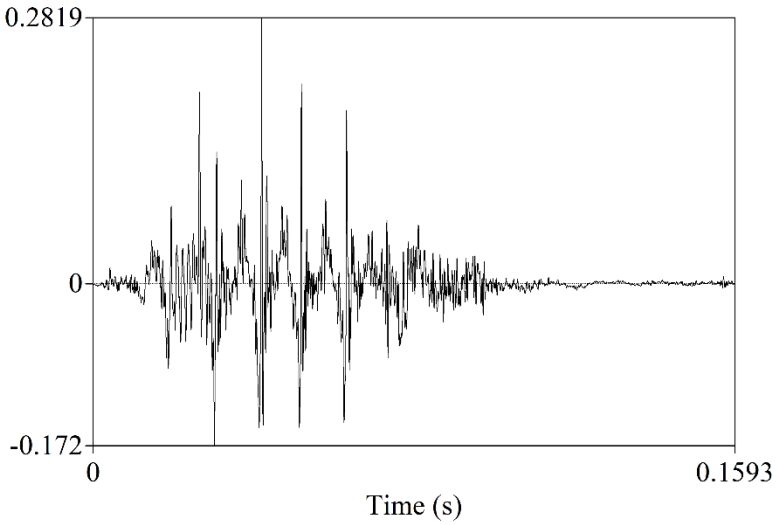




- **PC analysis**
- **2 phonetic dimensions of phonemes**
- **sibilant**
- **sonorant/loundness**



- **Changing a sound file into logfbank spectrogram**

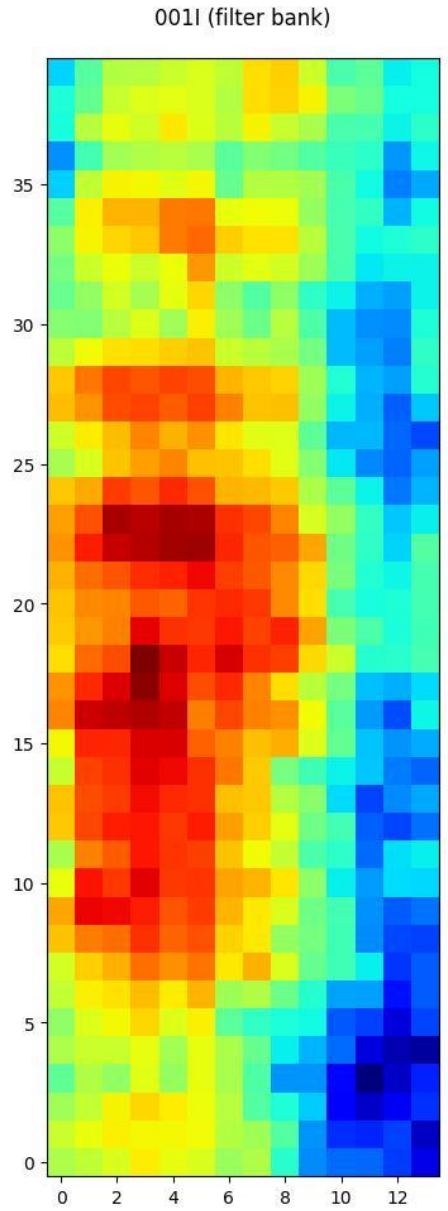


AfrikaansMale.txt - 记事本

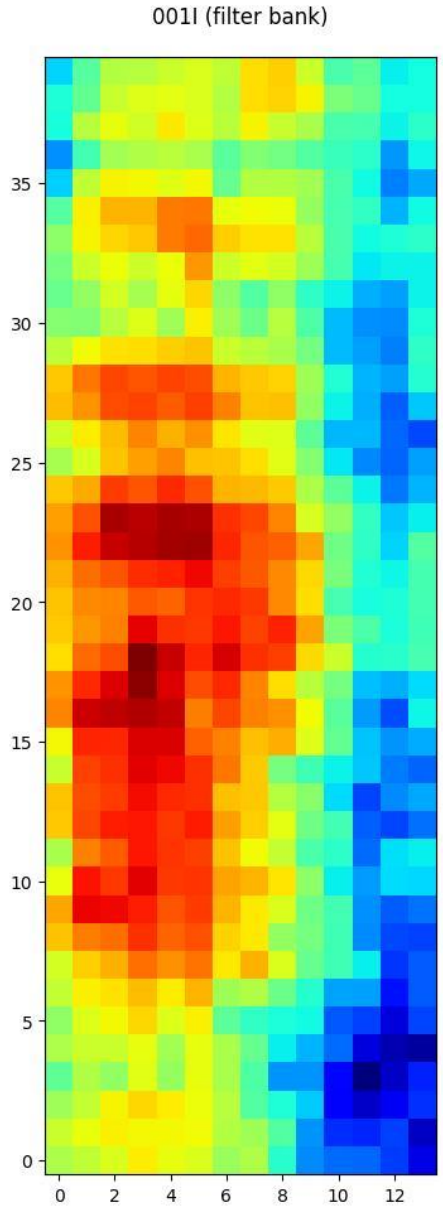
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

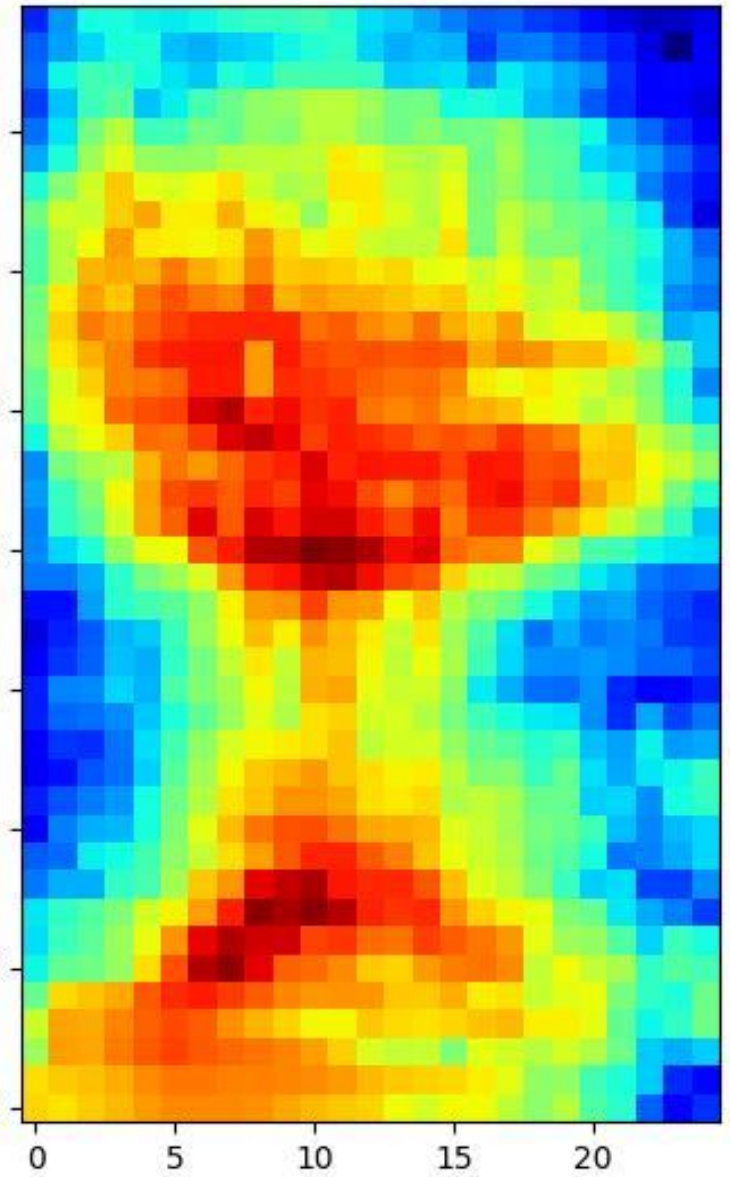
```
AfrikaansMale
001I
[[16.09318872 16.38537773 16.83003579 17.47683358 17.08911863 16.85441997
  15.73627862 16.11543865 13.77538309 11.80183848 11.36335178 11.2798829
  10.67291908  9.34844731]
 [16.58517382 17.10813229 17.43071866 17.28786053 17.31027178 17.24261797
  16.17248206 15.83117861 14.50531471 11.92550699 10.45546413 10.33278323
  10.72898563  8.90095918]
 [15.82529966 16.4834277  17.36693708 17.78335786 17.4806111  17.19161539
  16.26481013 14.52458704 13.62485972 12.69860732  9.75052798  8.96553931
   9.48600653 10.49473751]
 [14.72298115 16.11393345 15.61291548 16.98935248 15.69078029 17.08828438
  16.03738268 14.45094048 11.92037299 11.91672489  9.88481705  8.02566548
   8.96278404 10.29343605]
 [16.05661021 16.55791366 16.60123298 17.07844943 15.91260572 17.0938056
  16.01016829 14.8321726  13.2252715  12.39695091 11.37621658  9.24268275
   8.65956295  8.43519  ]
 [15.54487616 16.89257079 17.32946719 17.8346006  16.9248057  17.44017453
  14.64323941 13.95905894 13.53141577 13.37355062 11.09456808 10.76666433
   9.2163525  10.79778705]
 [16.4820098  17.419895   17.61326176 18.17816624 17.46027706 18.33582932
  15.82245147 16.11213465 14.95319686 13.72892263 12.13627781 12.08509168
  10.0697676  11.15493427]
 [16.7365217  17.91804913 18.37319209 19.36032961 18.84584334 19.3042299
  17.52697959 18.37171566 16.86118357 14.83554119 14.15208485 13.23441042
  10.56745502 11.15495653]
 [18.11195695 19.15372323 19.40038847 20.30760187 19.57353913 19.80908273
  17.86239659 17.50256681 15.62080003 15.09018312 14.1734914  11.8049879
  10.84228499 10.76922011]
 [18.54152722 21.01034634 20.95344386 20.6278906  19.74407471 20.17624332
  18.30591591 17.51937492 16.83569349 15.01115479 14.24508105 11.83882843
  11.12643407 11.47943924]
 [17.15057314 20.77350211 20.22063685 21.09120356 20.2463769  20.27976977
  18.58319707 18.30533252 17.60033072 15.50000647 13.19898895 12.06067805
  12.95104901 12.84717946]
 [16.00552    19.12182162 19.67924697 20.71441759 20.29035037 20.11900052
  18.07272306 17.23956444 16.46996405 14.38013725 13.84908192 11.37722387
  12.96338229 13.21392332]
 [18.03535828 19.96956623 20.59378352 20.70333884 20.23486475 20.64173168
  18.66050151 17.94574558 17.02550565 14.98169702 14.20834192 11.21812716
  10.8239827  11.39608847]
 [18.08515385 19.87385218 20.16557974 20.86785916 20.44791867 20.30797087
  18.16905411 18.03310602 16.16399936 15.43304232 12.92637755 10.78404985
  11.78492508 12.20868754]
 [16.53737842 20.07301907 20.31719053 21.12820701 20.94713208 20.32528508
```

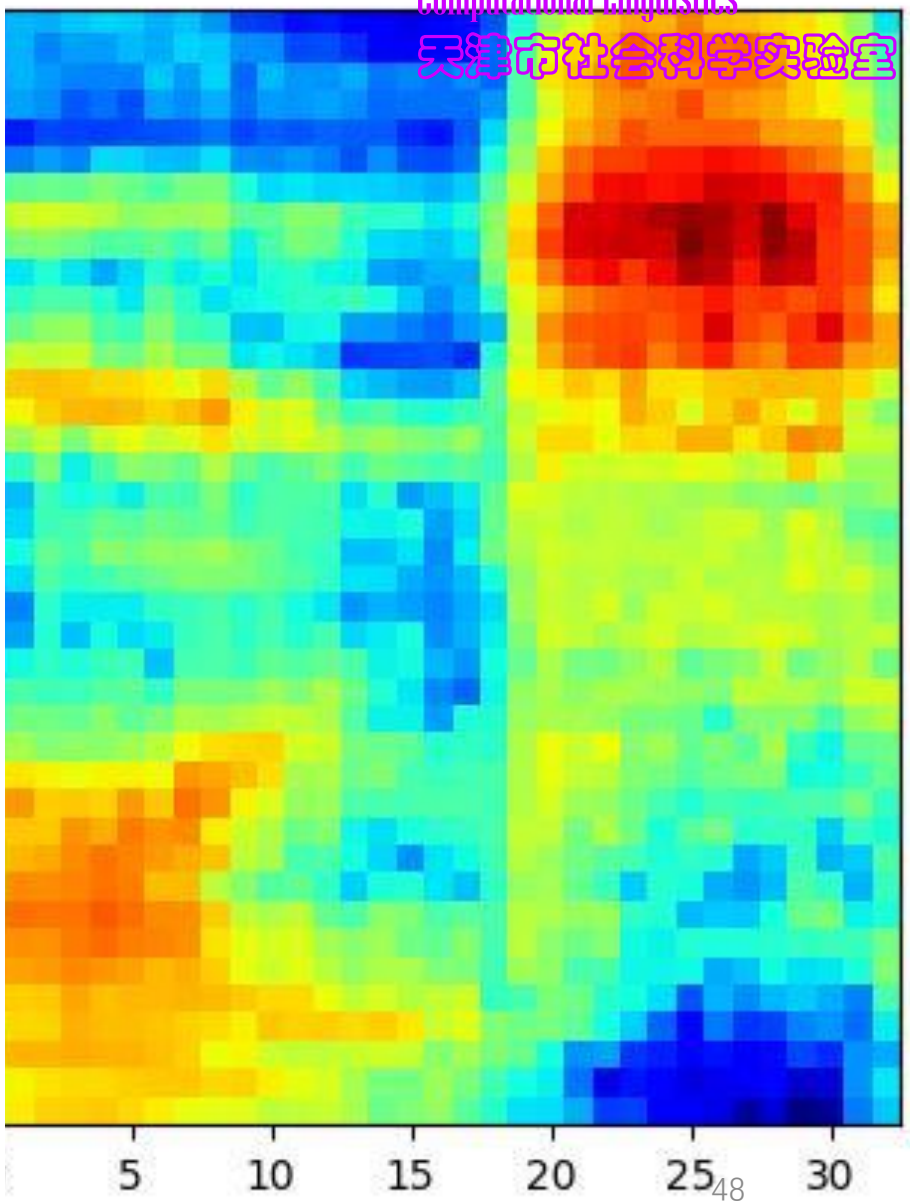
第 1 行，第 1 列　100%　Windows (CRLF)　UTF-8



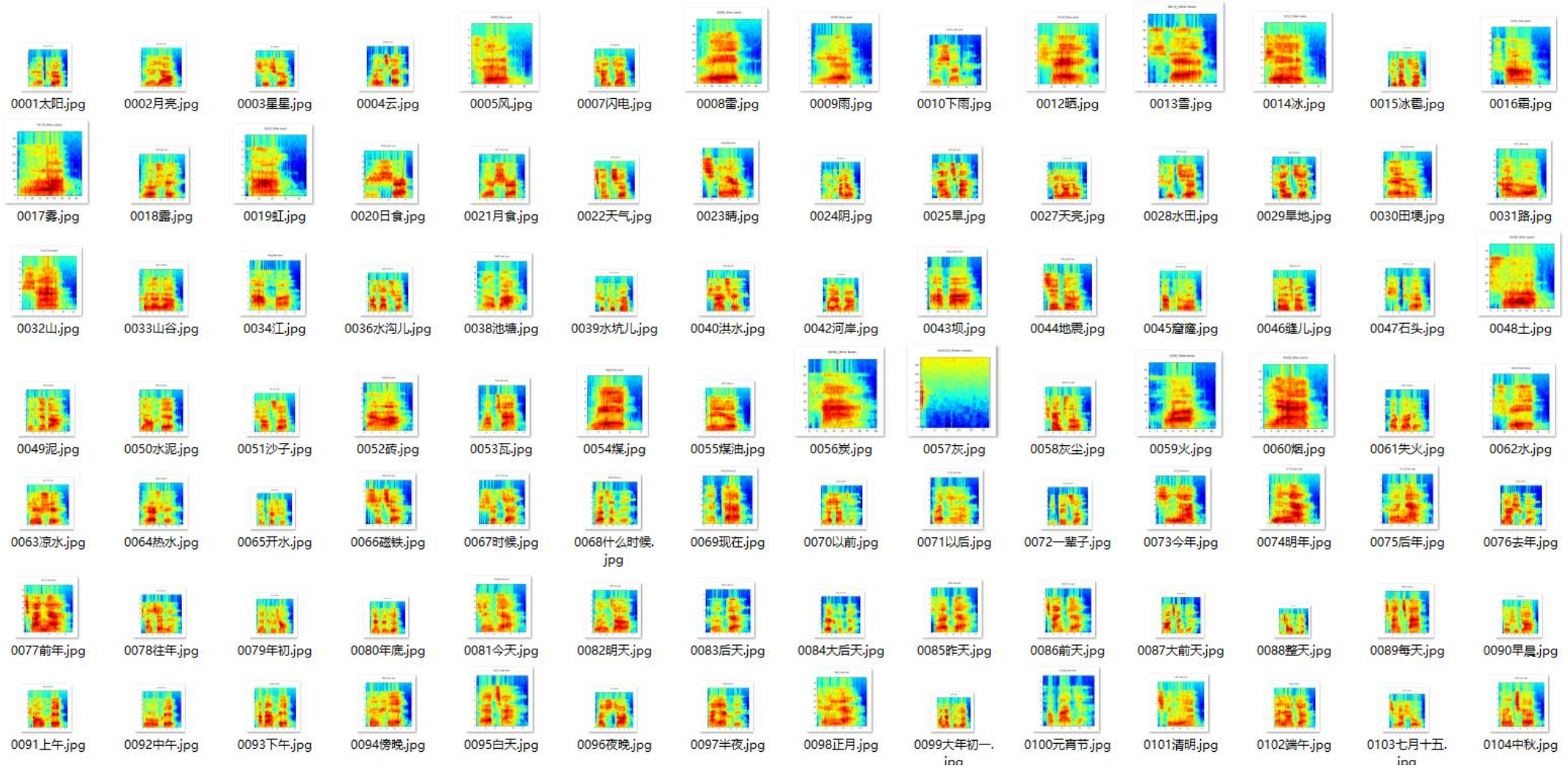
- **Afrikaans**


001l (filter bank)
0
5
10
15
20
25
30
35
0
2
4
6
8
10
12


002you-a (filter bank)
0
5
10
15
20


003we (filter bank)
5
10
15
20
25
30

- **Microsoft TTS engine: Azure**
- **Afrikaans.txt (98/300)**

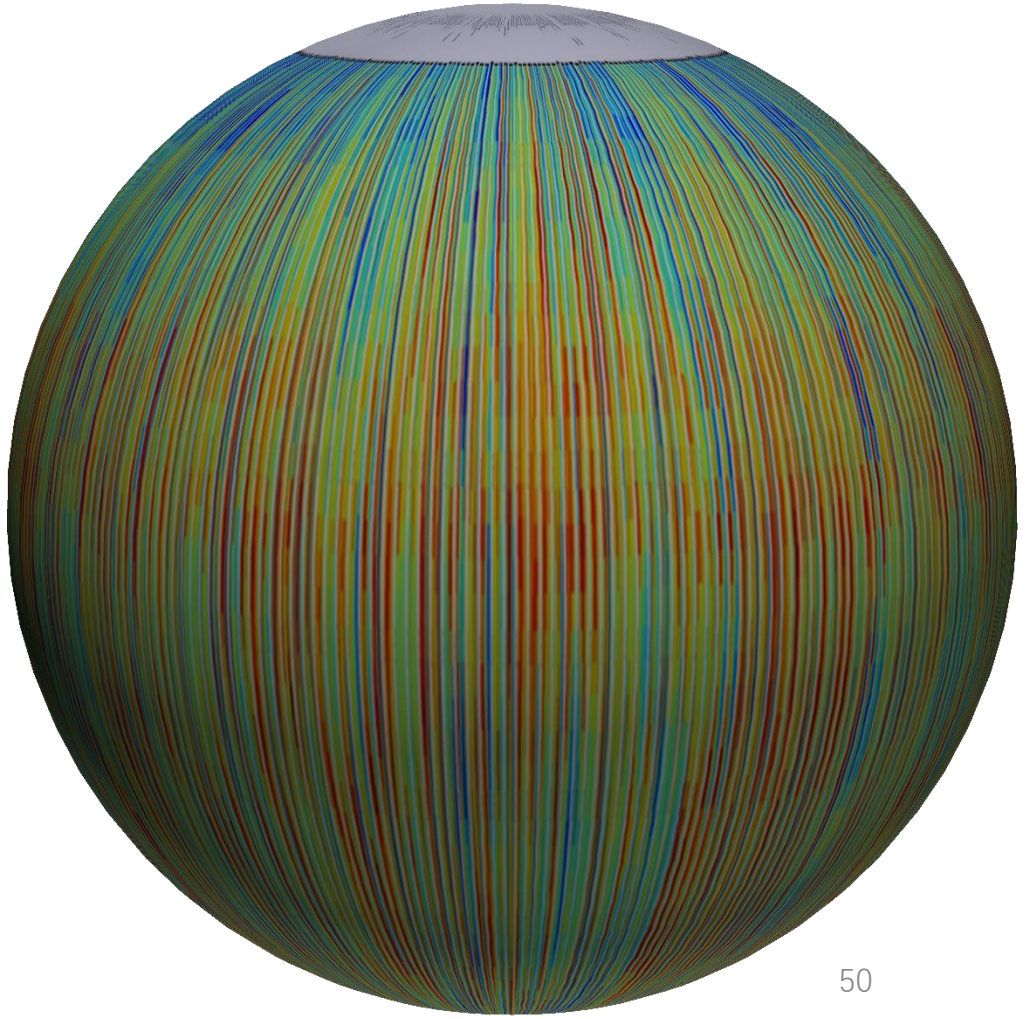


- **Acoustic graph of Affrikaans words(300 kernel words)**

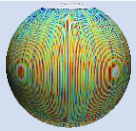
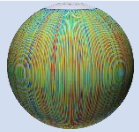
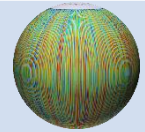
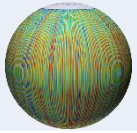
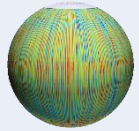
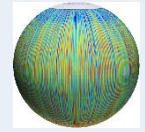
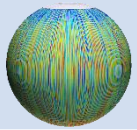
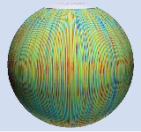
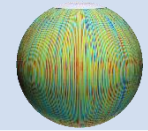
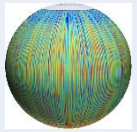
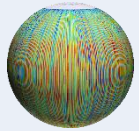
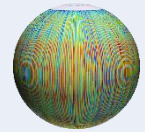
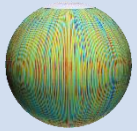
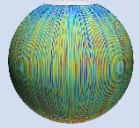
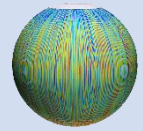
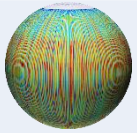
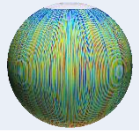
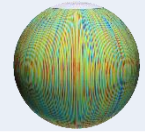
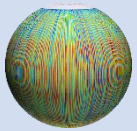
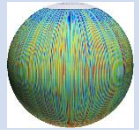
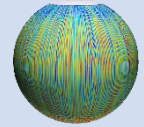
- **Afrikaans.tif**

- **Acoustic sphere of Affrikaans(planet/star)**

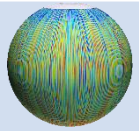
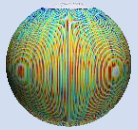
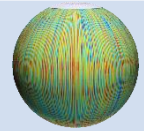
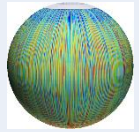
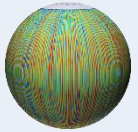
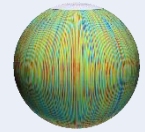
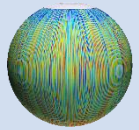
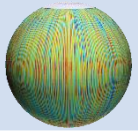
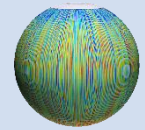
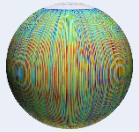
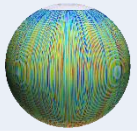
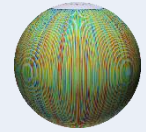
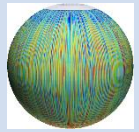
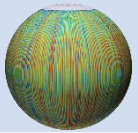
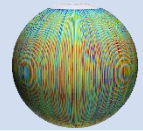
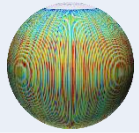
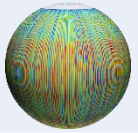
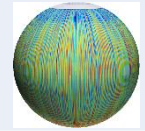
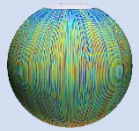
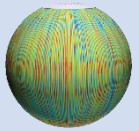
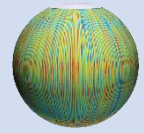
- **Digital acoustic map/atlas of kernel words?**
- **Sound-Meaning spheres of a language**



- **Sound-Meaning spheres of 300 kernel concepts of IE languages (1)**
- **21/42 IE languages**

language	S-M sphere	language	S-M sphere	language	S-M sphere
Afrikaans		Croatian		German	
Albanian		Czech		Greek	
Armenian		Danish		Gujarati	
Bengali		Dutch		Hindi	
Bosnian		English		Icelandic	
Bulgarian		French		Irish	
Catalan		Galician		Italian	

- **Sound-Meaning spheres of 300 kernel concepts of IE languages (2)**

language	S-M sphere	language	S-M sphere	language	S-M sphere
Latvian		Persian		Slovak	
Lithuanian		Polish		Slovenian	
Macedonian		Portuguese		Spanish	
Marathi		Romanian		Swedish	
Nepali		Russian		Ukrainian	
Norwegian		Serbian		Urdu	
Pashto		Sinhala		Welsh	

- **Using DTW algorithm to calculate distance between any 2 languages**

- **Matrix:**

	Afrikaans	Albanian	Armenian	Bengali	Bosnian	Bulgarian	Catalan	Croatian	Czech	Danish	Dutch	English	French	Galician	German	Greek	Gujarati	Hindi	Icelandic	Irish	Italian	Latvian	Lithuanian	Macedonian	Marathi	Nepali	Norwegian	Pashto	Persian	Polish	Portuguese	Romanian	Russian	Serbian	Sinhala	Slovak	Slovenian	Spanish	Swedish	Ukrainian	Urdu	Welsh
Afrikaans	0	2.89698	2.5839	3.31269	2.72788	3.21626	3.42708	3.11769	3.13706	3.1248	2.39854	2.92295	3.20711	3.06355	2.96842	3.2829	3.38965	3.32825	2.78161	3.29388	2.78707	2.91771	2.42463	3.02028	3.38074	2.74221	3.08284	3.0781	2.75698	2.92199	3.08182	3.11684	2.95482	3.00787	2.6711	3.22228	3.22487	3.13396	2.93652	2.89344	3.15865	2.63658
Albanian	2.89698	0	2.51514	3.07541	2.71241	3.07501	3.24008	2.9823	3.06345	3.02877	2.95324	3.15623	2.97111	2.75508	3.05627	3.08167	3.18602	3.14874	2.82342	2.94543	2.70805	2.95778	2.44836	2.91738	3.0967	2.60925	3.14266	2.82585	2.66513	2.87149	2.89631	2.9637	2.89432	2.94461	2.39638	3.07621	3.08475	2.9921	3.04673	2.83581	2.93098	2.60691
Armenian	2.5839	2.51514	0	2.74626	2.65466	2.72676	2.7791	2.71198	2.73963	2.69382	2.70474	2.8085	2.64406	2.52504	2.76498	2.70554	2.84155	2.80186	2.5478	2.61731	2.66175	2.78698	2.50856	2.65652	2.70785	2.55543	2.65197	2.54255	2.54704	2.77874	2.67152	2.70255	2.6878	2.67606	2.41622	2.72038	2.789	2.70784	2.65832	2.64747	2.66106	2.59524
Bengali	3.31269	3.07541	2.74626	0	2.91203	3.24768	3.41575	3.22396	3.17612	3.42981	3.28348	3.34311	3.26728	3.16156	3.30298	3.28415	3.1005	3.00796	3.12838	3.34865	3.24661	3.14221	3.56145	3.11365	3.09301	3.53393	3.43764	3.21892	2.95253	2.97947	3.16945	3.19123	3.08536	3.1437	2.77715	3.26869	3.3462	3.16456	3.25462	2.99879	2.89473	2.76416
Bosnian	2.72788	2.71241	2.65466	2.91203	0	2.46399	2.98946	2.04743	2.54536	2.79364	2.85944	2.9947	2.76056	2.70529	2.93489	2.91653	3.01133	2.96479	2.71702	2.84051	2.69976	2.78006	2.41475	2.16052	2.93442	2.69799	2.83838	2.75328	2.72784	2.44696	2.81684	2.82742	2.30634	1.72486	2.60866	2.41213	2.44933	2.82004	2.86653	2.2856	2.91829	2.73059
Bulgarian	3.21626	3.07501	2.72676	3.24768	2.46399	0	3.33625	2.57843	2.88586	3.32811	3.24133	3.23687	3.18776	3.10661	3.30412	3.21255	3.36339	3.31023	3.08337	3.29652	2.74976	3.0341	2.45292	2.45924	3.34282	2.86294	3.35179	3.15929	2.96094	2.68129	3.17479	3.15678	2.70843	2.61329	2.79468	2.89811	2.85184	3.10304	3.19464	2.68197	3.11618	2.68548
Catalan	3.42708	3.24008	2.7791	3.41575	2.98946	3.33625	0	3.28793	3.30544	3.57602	3.34464	3.34112	3.18973	3.24131	3.37635	3.3921	3.52592	3.43603	3.18983	3.5249	3.58351	3.18078	3.56814	3.20165	3.5172	3.99002	3.49323	3.43831	3.05183	3.077	2.88454	3.09347	3.18561	3.25045	2.97017	3.42384	3.45113	3.81119	3.33998	3.11848	3.21562	3.82405
Croatian	3.11769	2.9823	2.71198	3.22396	2.04743	2.57843	3.28793	0	2.71526	3.17324	3.16894	3.2183	3.09301	3.04738	3.23178	3.20387	3.34627	3.30768	3.02194	3.20467	2.7792	3.07024	2.44308	2.45108	3.26139	2.79902	3.27762	3.04461	2.90762	2.63973	3.13845	3.05919	2.64487	2.10031	2.74778	2.66744	2.58787	3.07798	3.13724	2.50839	3.09939	2.71331
Czech	3.13706	3.06345	2.73963	3.17612	2.54536	2.88586	3.30544	2.71526	0	3.17715	3.16411	3.18062	3.14312	3.09032	3.31951	3.231	3.3632	3.31532	3.04891	3.22431	2.87232	3.00672	2.55201	2.78848	3.29395	2.81614	3.26095	3.07006	2.94711	2.65939	3.13967	3.09932	2.74742	2.68203	2.75988	2.51921	2.82647	3.05188	3.15617	2.62511	3.07926	2.74751
Danish	3.1248	3.02877	2.69382	3.42981	2.79364	3.32811	3.57602	3.17324	3.17715	0	3.02325	3.09777	3.30911	3.20462	3.10227	3.369	3.4869	3.3747	2.80265	3.41321	2.75982	3.1036	2.55559	3.12023	3.48117	2.74708	3.0792	3.25652	2.88028	2.95425	3.21158	3.19237	3.06295	3.13586	2.71737	3.31926	3.34703	3.17042	2.91021	2.97043	3.19394	2.63449
Dutch	2.39854	2.95234	2.70474	3.28348	2.85944	3.24133	3.34464	3.16894	3.16411	3.02325	0	2.93171	3.19164	3.09728	2.81383	3.2116	3.34964	3.31627	2.88043	3.22604	2.84423	3.04242	2.52907	3.06773	3.32897	2.83788	3.01633	3.14067	2.87708	3.00103	3.07437	3.17313	3.05286	3.08535	2.78417	3.23329	3.17987	3.0964	2.92219	2.99518	3.14072	2.73792
English	2.92295	3.15633	2.8085	3.34311	2.9947	3.23687	3.34112	3.2183	3.18062	3.09777	2.93171	0	3.18189	3.14445	3.06794	3.33818	3.37476	3.33287	2.99617	3.19018	2.86754	3.1608	2.59489	3.15883	3.31227	2.90237	3.12721	3.21576	2.98422	3.03964	3.10316	3.22683	3.1194	3.15652	2.9016	3.26638	3.34129	3.20712	3.13277	3.09807	3.23586	2.74322
French	3.20711	2.97111	2.64406	3.26728	2.76056	3.18776	3.18973	3.09301	3.14312	3.30911	3.19164	3.18189	0	3.0099	3.27736	3.24446	3.36805	3.32849	3.01391	3.32317	2.5876	3.00472	2.47921	3.01941	3.32053	2.72131	3.36357	3.11688	2.86051	2.91402	3.00952	2.97468	2.99439	3.04527	2.67403	3.29754	3.16577	2.99398	3.19321	2.8924	3.10267	2.58835
Galician	3.06355	2.75508	2.52504	3.16156	2.70529	3.10661	2.94131	3.04738	3.03027	3.20462	3.09728	3.14445	3.0099	0	3.19907	3.13	3.30622	3.21965	2.82372	3.15524	2.44	2.93493	2.42382	2.95724	3.27296	2.65725	3.29571	3.02921	2.74021	2.90191	2.92585	2.89078	2.92903	2.98385	2.52695	3.195	3.19123	2.86782	3.06748	2.86119	3.04506	2.571
German	2.96842	3.05627	2.76498	3.30298	2.93489	3.30412	3.37635	3.23178	3.31951	3.10227	2.81383	3.09728	3.27736	3.24446	3.19907	0	3.28286	3.39105	3.35256	2.98919	3.27153	2.88714	3.09461	2.55781	3.17058	3.37002	2.87434	3.17092	3.21116	2.94725	3.06123	3.21259	3.1836	3.14671	3.16686	2.84898	3.31181	3.24288	3.16333	3.02852	3.09213	3.20183	2.78386
Greek	3.2829	3.08167	2.70554	3.28415	2.91653	3.21255	3.3921	3.20387	3.231	3.369	3.2116	3.33818	3.24446	3.13	3.28286	0	3.37442	3.27563	3.09797	3.28285	2.85921	3.07208	2.51632	3.14264	3.31586	2.8137	3.3931	3.15807	2.93183	3.03339	3.12582	3.16707	3.12204	3.1566	2.791	3.34014	3.29993	3.05439	3.26962	3.0398	3.13931	2.72095
Gujarati	3.38965	3.18602	2.84155	3.1005	3.01133	3.36339	3.52592	3.34627	3.3632	3.4869	3.34964	3.37476	3.36805	3.30622	3.39105	3.37442	0	2.73869	3.24698	3.46119	2.95049	2.81874	2.60789	3.24456	3.69654	2.51363	3.57263	3.23366	3.05473	3.07782	3.30467	3.31589	3.21294	3.27218	2.90715	3.42031	3.45487	3.30716	3.14869	2.98118	2.86996
Hindi	3.32825	3.14874	2.80186	3.00796	2.96479	3.31023	3.43603	3.30768	3.31532	3.37747	3.31627	3.33287	3.32849	3.21965	3.23795	3.38078	3.34964	3.31627	2.88043	3.22604	2.84423	3.04242	2.52907	3.06773	3.32897	2.83788	3.01633	3.14067	2.87708	3.00103	3.07437	3.17313	3.05286	3.08535	2.78417	3.23329	3.17987	3.0964	2.92219	2.99518	3.14072	2.73792
Icelandic	2.78161	2.82342	2.5478	3.12838	2.71702	3.08337	3.18983	3.02194	3.04891	3.22431	2.88043	3.22604	3.01391	3.32317	2.5876	3.00472	2.47921	3.01941	3.32317	0	2.88714	3.09461	2.55781	3.17058	3.37002	2.87434	3.17092	3.21116	2.94725	3.06123	3.21259	3.1836	3.14671	3.16686	2.84898	3.31181	3.24288	3.16333	3.02852	3.09213	3.20183	2.78386
Irish	3.29388	2.94543	2.61731	3.34865	2.84051	3.29652	3.5249	3.20467	3.22431	3.41321	3.22604	3.19018	3.32317	3.15524	3.27153	3.28285	3.46119	3.38624	3.06527	0	2.80186	3.02704	2.47558	3.09961	3.42176	2.74132	3.54726	3.17582	2.82794	2.94975	3.10429	3.19301	3.08474	3.15672	2.64236	3.34082	3.32242	3.10582	3.23023	3.00373	3.18387	2.54996
Italian	2.78707	2.70805	2.66175	2.84661	2.69976	2.74976	2.58351	2.7792	2.87232	2.75982	2.84423	2.86754	2.5876	2.44	2.88714	2.85921	2.95049	2.89536	2.76764	2.80186	0	2.91017	2.55785	2.73494	2.96308	2.59641	2.81732	2.73801	2.78472	2.82867	2.57785	2.51662	2.78295	2.76702	2.58662	2.8505	2.88141	2.47366	2.82247	2.71107	2.75986	2.55425
Latvian	2.91771	2.95778	2.78698	3.14221	2.78006	3.0341	3.18078	3.07024	3.00672	3.1036	3.04242	3.1608	3.00472	2.93493	3.09461	3.07208	3.21874	3.21684	2.92679	3.02704	2.91017	0	2.24011	2.93039	3.15606	2.91729	3.06457	2.99334	2.88446	2.87103	3.05286	3.08535	2.78414	3.23329	2.90715	3.42031	3.45487	3.30716	3.14869	2.98118	2.86996
Lithuanian	2.44463	2.44836	2.50856	2.56145	2.41475	2.45232	2.56814	2.44308	2.55201	2.55559	2.52907	2.59489	2.47921	2.42382	2.55781	2.51632	2.60789	2.59973	2.41744	2.47558	2.55785	2.24011	0	2.42561	2.5609	2.47243	2.46614	2.46217	2.49525	2.55352	2.50542	2.54076	2.44034	2.36032	2.42514	2.49001	2.47666	2.51669	2.52042	2.39932	2.47666	2.51669	2.50042
Macedonian	3.02028	2.91738	2.65652	3.11365	2.16052	2.45924	2.45934	2.45108	2.45201	2.55201	2.55559	2.52907	2.54976	2.42382	2.55785	2.52614	2.60789	2.59973	2.41744	2.47558	2.55785	2.24011	0	2.42561	2.5609	2.47243	2.46614	2.46217	2.49525	2.55352	2.50542	2.54076	2.44034	2.36032	2.42514	2.49001	2.47666	2.51669	2.52042	2.39932	2.47666	2.51669	2.50042
Marathi	3.38074	3.0967	2.70785	3.09301	2.93442	3.34282	3.5172	3.26139	3.29395	3.12023	3.06773	3.15883	3.01941	2.95724	3.17058	3.14264	3.32456	3.171	2.93766	3.09961	2.72494	2.93039	2.42561	0	3.20381	2.75607	3.17515	2.99666	2.85432	2.61609	3.01415	3.01291	2.50909	2.15653	2.69184	2.7086	2.74751	2.97625	3.05395	2.44777	3.00116	2.68033
Nepali	3.74221	3.60925	3.54255	3.53393	3.39952	3.45108	3.99002	3.27762	3.30901	3.48117	3.32897	3.31227	3.32053	3.27296	3.31586	3.31227	3.37442	3.27563	3.09797	3.28285	2.85921	3.07208	2.51632	3.14264	3.31586	2.8137	3.3931	3.15807	2.93183	3.03339	3.12582	3.16707	3.12204	3.1566	2.791	3.34014	3.29993	3.05439	3.26962	3.0398	3.13931	2.72095
Norwegian	3.08284	3.14266	2.65197	3.43764	3.03838	3.35179	3.43428	3.04461	3.07062	3.74708	3.07924	3.1608	3.00472	2.93493	3.09461	3.07208	3.21874	3.21684	2.92679	3.02704	2.91017	0	2.24011	2.93039	3.15606	2.91729	3.06457	2.99334	2.88446	2.87103	3.05286	3.08535	2.78414	3.23329	2.90715	3.42031	3.45487	3.30716	3.14869	2.98118	2.86996
Pashto	3.0781	2.82585	2.54255	3.21892	2.75328	3.15929	3.43831	3.04461	3.07006	3.25652	3.14067	3.23176	3.11688	3.02921	3.21116	3.15807	3.23366	3.21684	2.92679	3.02704	2.91017	0	2.24011	2.93039	3.15606	2.91729	3.06457	2.99334	2.88446	2.87103	3.05286	3.08535	2.78414	3.23329	2.90715	3.42031	3.45487	3.30716	3.14869	2.98118	2.86996
Persian	2.75698	2.66513	2.54704	2.95253	2.72784	2.96094	3.05473	2.74021	2.88028	2.95425	2.88446	2.95637	2.65372	2.93018	2.60974	0	2.90095	2.85643	2.87881	2.81311	2.85784	2.57436	2.92201	2.93985	2.67663	2.15332	3.1747	2.77958	2.58564	2.84114	2.84124	0	2.16746	2.25649	2.74267	2.21835	2.77101
Polish	2.92199	2.87149	2.77874	2.97947	2.44696	2.68129	3.077	2.63973	2.65939	2.95425	3.00103	3.03964	2.9108	2.87042	3.01591	3.02362	3.07782	3.09322	2.9108	3.14529	2.69654	2.73494	2.55352	2.61609	3.01415	3.01291	2.50909	2.15653	2.69184	2.7086	2.74751	2.97625	3.05395	2.44777	3.00116	2.68033	2.74394	2.54341	2.72248	2.59424	2.67421	2.79512	2.69638
Portuguese	3.08182	2.89631	2.67152	3.16945	2.81684	3.17479	3.88454	3.13845	3.13967	3.21158	3.07437	3.10316	3.00952	2.97468	2.99439	3.04527	2.67403	2.97354	2.91402	3.00952	2.97468	2.99439	3.04527	2.67403	2.97354	2.91402	3.00952	2.97468	2.99439	3.04527	2.67403	2.97354	2.91402	3.00952	2.97468	2.99439	3.04527	2.67403	2.97354	2.91402	3.00952	2.97468
Romanian	3.11684	2.9637	2.70355	3.19123	2.82742	3.10678	3.70843	3.05919	3.09932	3.31258	3.17313	3.22683	2.97468	2.89078	3.10316	3.22683	3.1194	3.15652	2.90191	3.15924	2.97354	2.91402	3.00952	2.97468	2.99439	3.04527	2.67403	2.97354	2.91402	3.00952	2.97468	2.99439	3.04527	2.67403	2.97354	2.91402	3.00952	2.97468	2.99439	3.04527	2.67403	2.97354
Russian	2.95482	2.89432	2.6878	3.08536	2.30634	2.61329	2.79468	2.64487	2.74742	2.88028	2.95425	3.21158	3.19237	3.06295	3.13586	2.71737	3.31926	3.34703	3.17042	2.91021	2.97043	3.02362	2.42514	2.49001	2.47666	2.51669	2.52042	2.39932	2.47666	2.51669	2.50042	2.54076	2.44034	2.36032	2.42514	2.49001	2.47666	2.51669	2.52042	2.39932	2.47666	2.51669
Serbian	3.00787	2.94461	2.67606	3.1437	1.72486	2.61329	3.25045	2.10031	2.68203	3.13586	3.08525	3.15652	3.04527	2.98385	3.16686	3.1566	3.27218	3.25364	2.98264	3.15652	2.76702	2.90061	2.36032	2.42514	2.49001	2.47666	2.51669	2.52042	2.39932	2.47666	2.51669	2.50042	2.54076	2.44034	2.36032	2.42514	2.49001	2.47666	2.51669	2.52042	2.39932	2.47666
Sinhala	2.6711	2.39638	2.41622	2.77715	2.60866	2.79468	2.97017	2.74778	2.82647	3.34703	3.17987	3.34129	3.16577	3.19123	3.24288	3.29993	3.45487	3.38148	3.07773	3.32242	2.88141	3.02536	2.47666	2.74751	3.43187	2.89167	3.40025	3.17069	2.93985	2.67663	3.15332	3.1747	2.77958	2.58564	2.84114	2.84124	0	2.16746	2.25649	2.74267	2.21835	2.77101
Slovak	3.22228	3.07621	2.72038	3.26869	2.41213	2.89811	3.42384	2.66744	2.51921	3.31926	3.22329	3.26638	3.29754	3.195	3.31181	3.34014	3.42031	3.36382	3.09174	3.34082	2.8505	2.98832	2.49001	2.7086	3.39194	2.82762	3.36685	3.18159	2.92201	2.40166	3.15737	3.1762	2.63482	2.58741	2.79815	0	2.84124	3.15676	3.23628	2.57205	3.15282	2.74403
Slovenian	3.22487	3.08475	2.789	3.3462	2.44933	2.85184	3.45113	2.58787	2.82647	3.34703	3.17987	3.34129	3.16577	3.19123	3.24288	3.29993	3.45487	3.38148	3.07773	3.32242	2.88141	3.02536	2.47666	2.74751	3.43187	2.89167	3.40025	3.17069	2.93985	2.67663	3.15332	3.1747	2.77958	2.58564	2.84114	2.84124	0	2.16746	2.25649	2.74267	2.21835	2.77101
Spanish	3.13396	2.9921	2.70784	3.16456	2.82004	3.10304	2.81119	3.07798	3.05188	3.17042	3.0964	3.20712	2.98398	3.26782	3.16333	3.05439	3.30716	3.21577	2.99827	3.10582	2.47366	3.10383	2.51669	2.97625	3.24087	2.77898	3.18648	3.08131	2.88969	2.9185	2.49935	2.97139	3.02779	3.02931	2.6952	3.15676	3.16746	0	3.13349	2.9275	3.04986	2.67466
Swedish	2.93652	3.04673	2.65832	3.25462	2.86653	3.19464	3.33998	3.13724	3.15617	3.91021	3.92219	3.13277	3.19321	3.06748	3.02852	3.26962	3.39425	3.28761	2.71286	3.23023	2.82347	3.10761	2.52042	3.05395	3.36748	2.85392	3.60446	3.12768	2.86837	2.94843	3.12752	3.16566	3.03704	3.07304	2.77559	3.23628	3.25649	3.13349	0	2.99819	3.12428	2.73324
Ukrainian	2.89344	2.83581	2.64747	2.99879	2.2856	2.68197	3.11848	2.50839	2.62511	2.97043	2.99518	3.09807	2.8924	2.86119	3.09213	3.0398	3.14869	3.12619	2.8321	3.00373	2.71107	2.89625	2.39932	2.44777	3.0869	2.78239	3.05345	2.87754	2.7526	2.47219	2.90586	2.95224	2.27519	2.36146	2.63896	2.57205	2.74267	2.9275	2.99819	0	2.99956	2.67417
Urdu	3.15865	2.93098	2.66106	2.99473	2.81829	3.11618	3.31562	3.09939	3.07926	3.19394	3.14072	3.23586	3.10567	3.04506	3.20183	3.12931	2.98118	2.56263	2.9817	3.18387	2.75986	3.07417	2.49657	3.00116	2.99419	2.46776	3.24004	2.97584	2.75925	2.93353	3.05319	3.08839	3.01596	3.06279	2.67193	3.15282	3.21835	3.04986	3.12428	2.93956	0	2.66003
Welsh	2.63658	2.60691	2.59524	2.76416	2.73059	2.68548	2.82405	2.71331	2.74751	2.63449	2.73792	2.74322	2.58835	2.571	2.78386	2.72095	2.96996	2.90753	2.64597	2.54996	2.55425	2.89522	2.58967	2.68033	2.74394	2.54341	2.72248	2.59424	2.67421	2.79512	2.69638	2.70291	2.75506	2.71825	2.46672	2.74403	2.77101	2.67466	2.73324	2.67417	2.66003	0

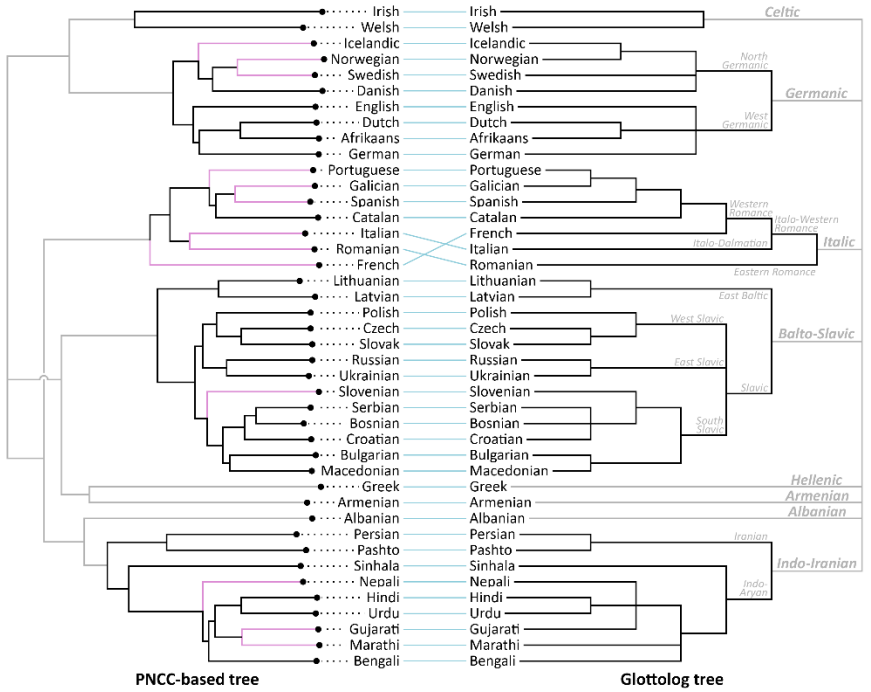
- **DTW-D_by-max_CMN_BFCC.csv**

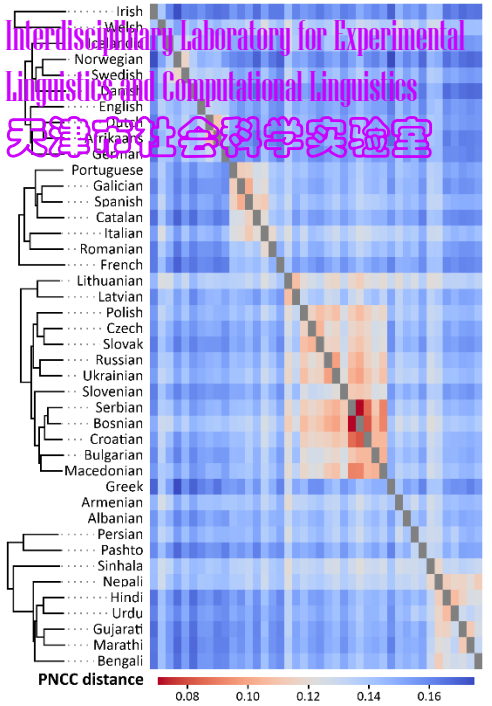
- **New phylogenetic tree is very similar to the traditional classification.**

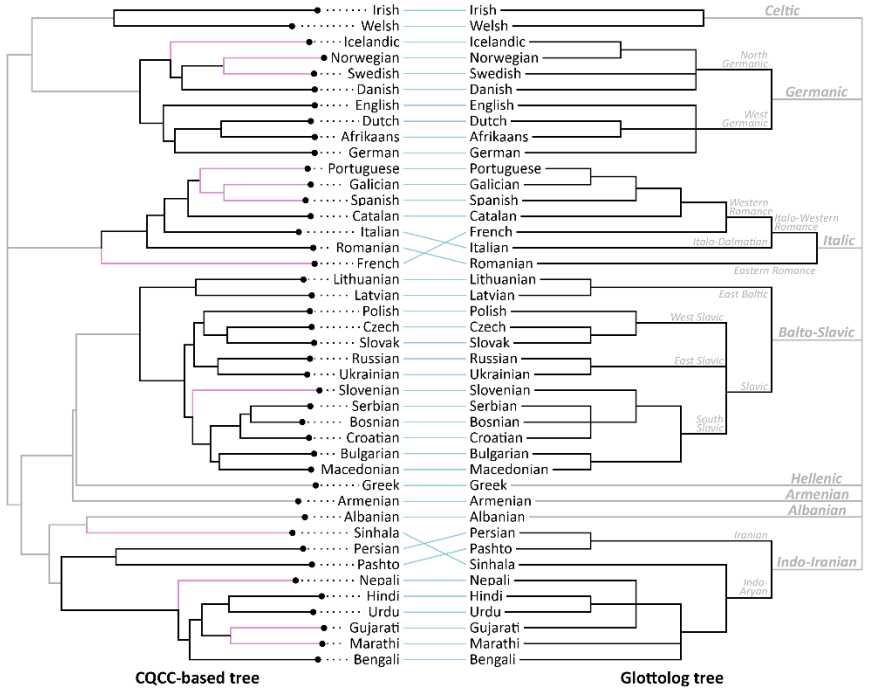
- **This is really interesting.**
- **We tested more than 10 acoustic features.**

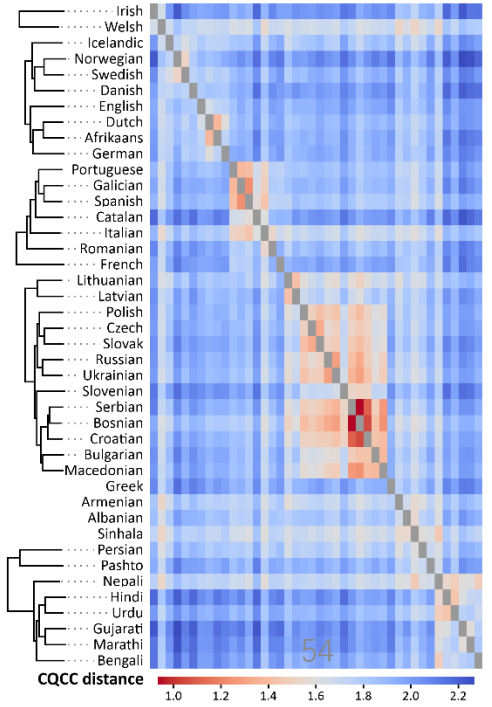
- **×writing system**
- **×transcription**

- **Language history is hided/embedded in speech sounds, or in everyone's mouth!**

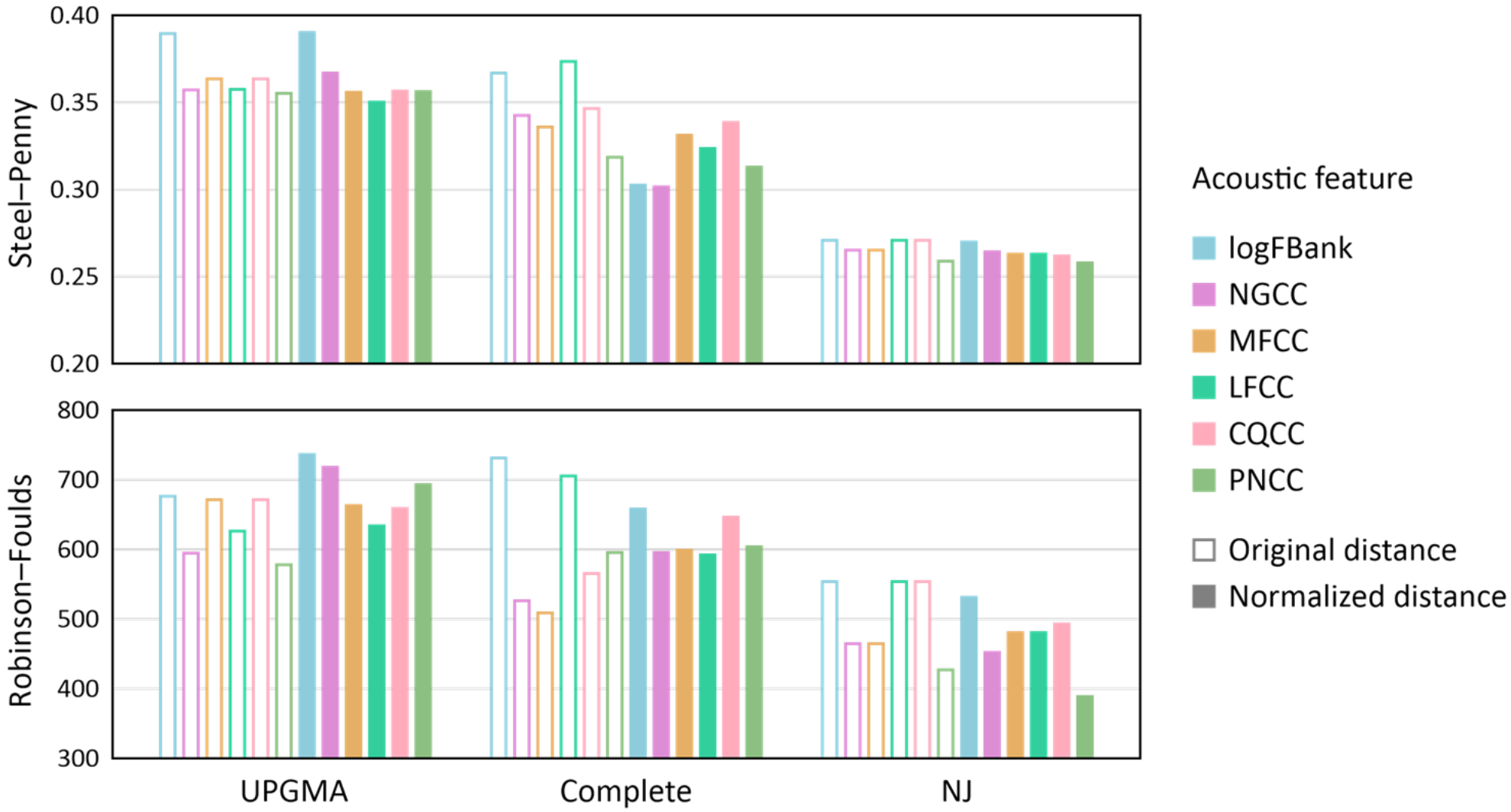






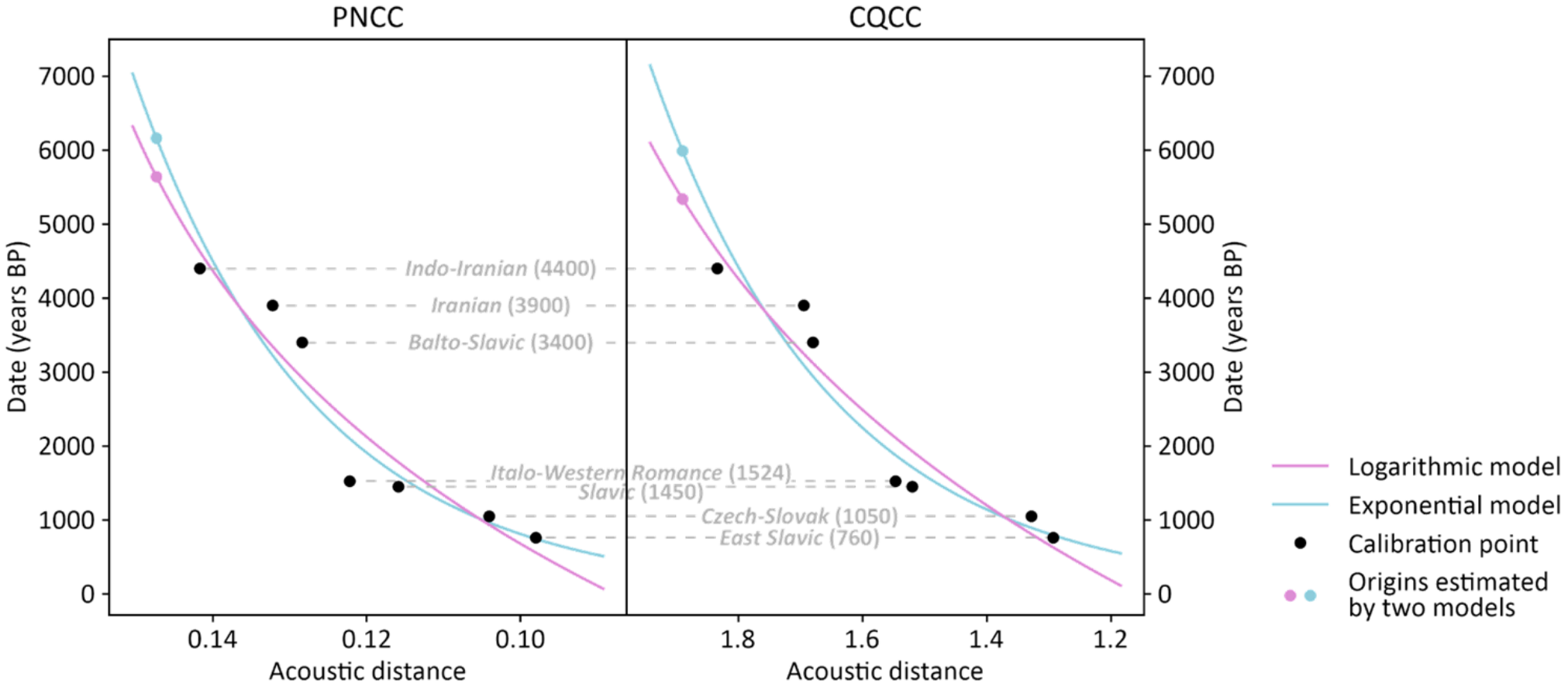


- **Different acoustic features**

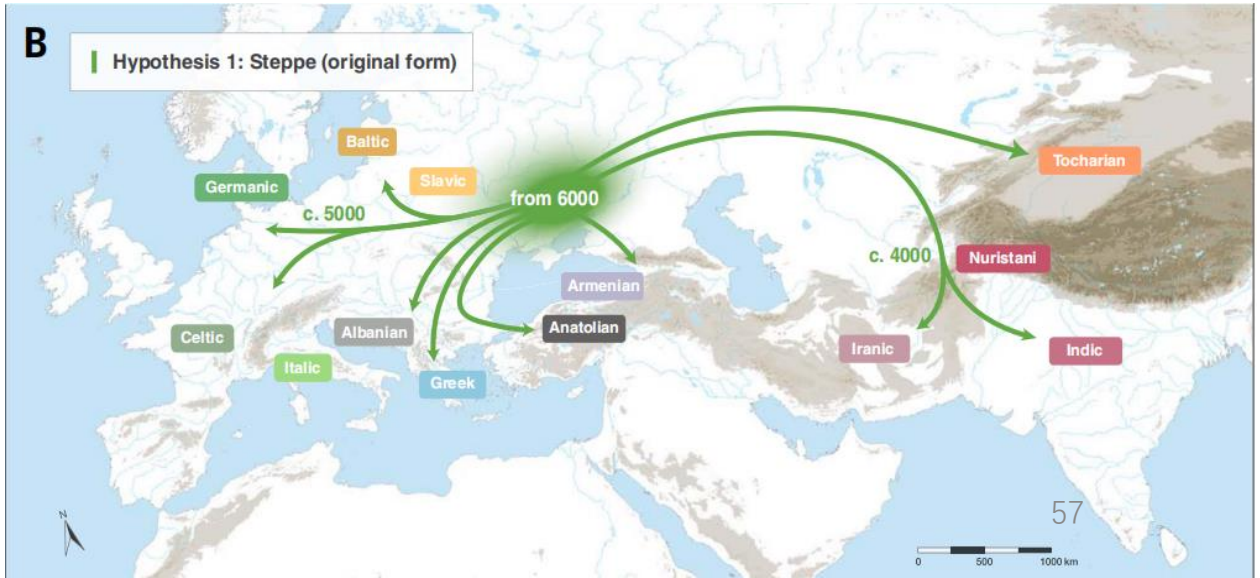


- **Shorter the column is, better the acoustic feature/method is.**

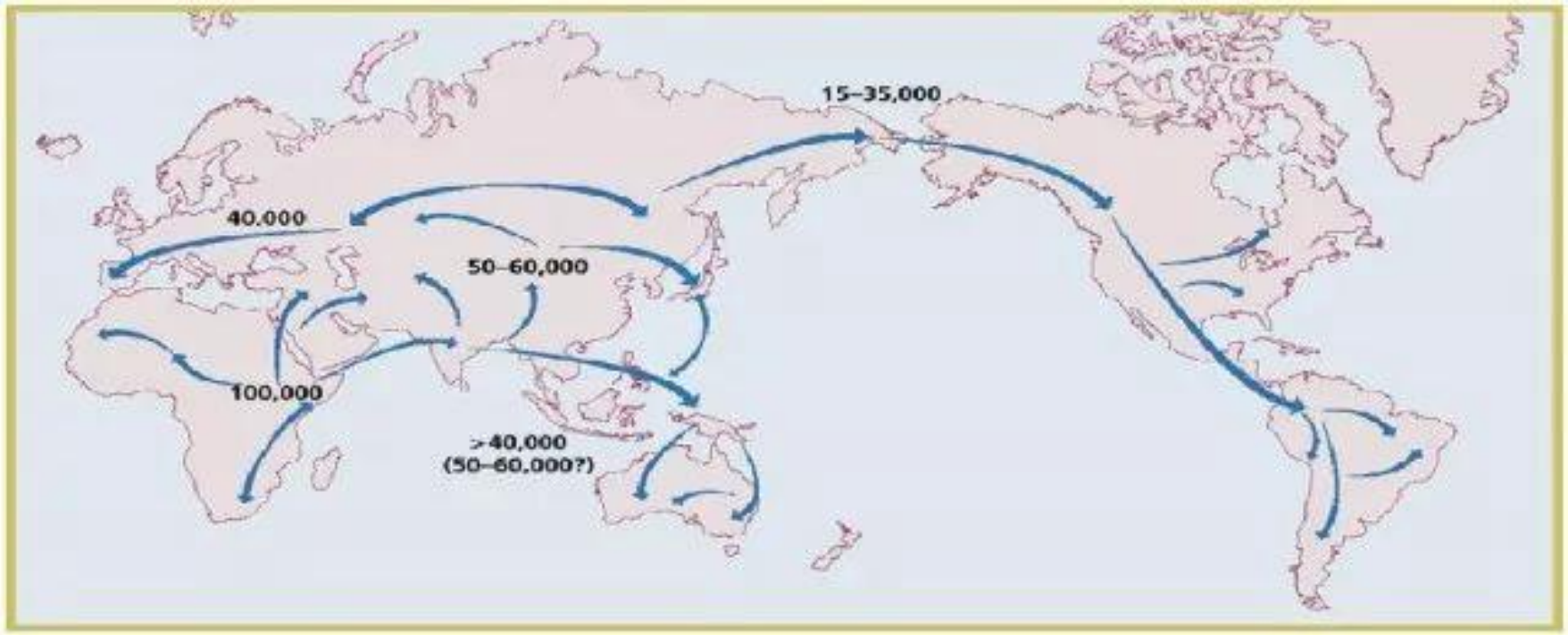
- **Fitting chronology!**
- **The acoustic distance matrix, can not only be used to construct the phylogenetic tree, but also can be used to fit I-E language chronology.**



- **The acoustic distance-based clustering of these languages aligns well with established classifications, and the acoustic distances effectively reflect language divergence ages.**
- **Acoustic results imply that the origin of Indo-European languages (except Anatolian and Tocharian) dates back to 5,700–5,000 BP.**
- **2 hypothesis:**
- **Anatolian hypothesis: 8,500BP, agricultural diffusion**
- **√ Kurgan hypothesis (grassland hypothesis): 6,500BP-4,500BP**
- **This is really interesting and meaningful.**
- **It is beyond expectations.**



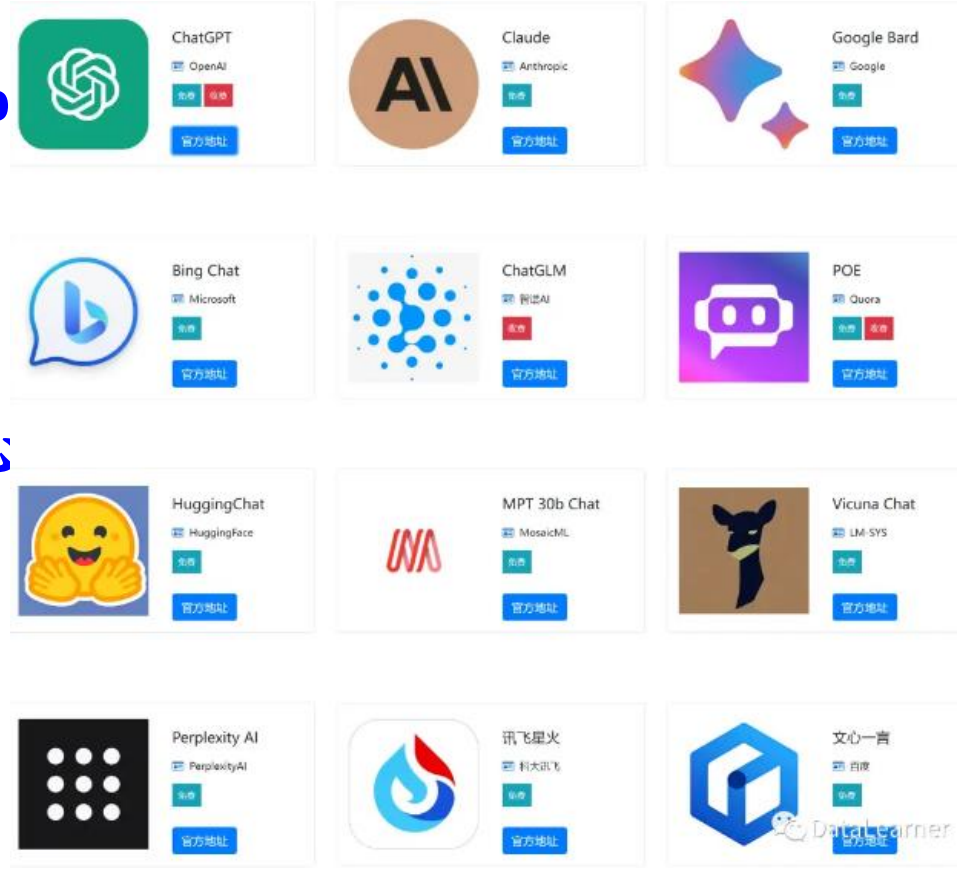
- **Future work:**
- **Simulating and reconstructing the migration route of ancestors of modern human from the 'homeland' in Africa**
- **Global Recordings Network(https://globalrecordings.net/en/)**
- **6,527 language varieties**



Underhill, Shen, Lin et al. (2000)

The age of Artificial Intelligence

- We are in the age of AI developing rapidly.
- Accurate automatic speech recognition，Text-to Speech(speech synthesis)……
- Image recognition, Face recognition……
- Prompt Engineering:
- Large Language Model(LLM，ChatGPT，文心一言等)……
- Generate images, Generate videos……
- transportation、shopping、financial service、education, health care……



IV-Engineering attempts using deep learning and AI technology in our lab

- 1-Speech synthesis(Text-To-Speech)
- LingSound speakers: Lingsheng, Lingzhi, Junsheng, Junzhi
- specific speakers, comic figures, Chinese dialects, chanting for ancient poem…
- Pingwu(Sichuan) dialect; Tianjin dialect;

项目	链接	简介
语音合成	语音合成	支持包括实验室自研男声、女声；名人声音；天津话、四川话等方言的合成
国际音标识别	国际音标识别	支持国际音标的流式识别
语种识别	语种识别	支持汉语普通话、英语、法语、德语、西班牙语、日语、加泰罗尼亚语、卢旺达语等8种语言的识别。
汉语学习者发音评分	汉语学习者发音评分	对汉语学习者的口语进行评估
声纹识别	声纹识别	识别已注册声音
满汉互译系统	满汉互译系统	支持满语、汉语双向翻译（初步测试）

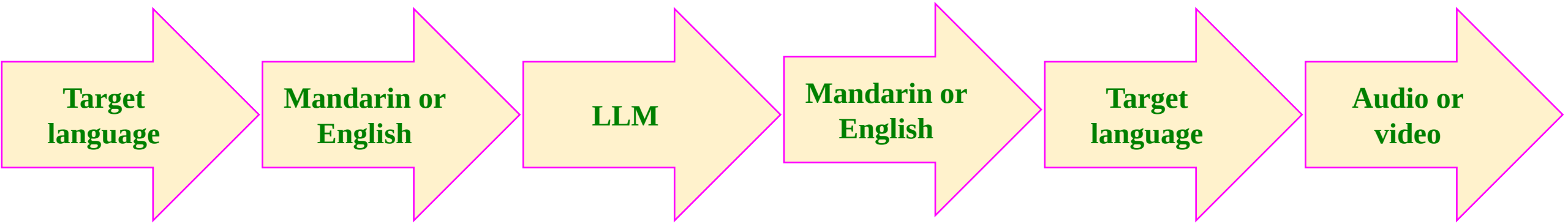
- **2-Automatic Speech Recognition(IPA ASR, language recognition, foreign accent evaluation, speaker identification…)**
- **demo**
- **https://www.ranqibin.com/language-identification**

- **3-Machine Translation(Manchu-Mandarin bidirectionally automatic translation)**
- **Endangered language, Paul Georg von Möllendorff's transcription**
- **https://manchumt.app.ranqibin.com/**

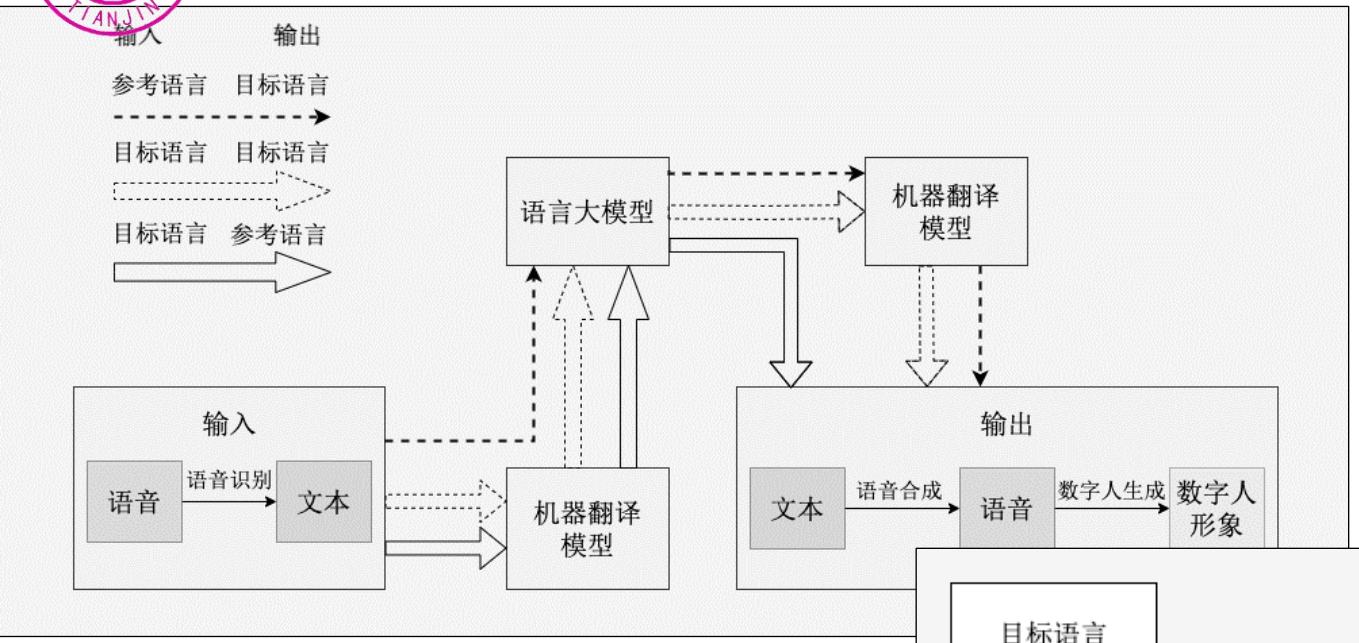
- **Engineering attempts using deep learning and AI in our lab**
  - **4-digital human (Yuen-Ren Chao, Chia-ying Yeh, the first president of Nankai Univ.; Tangwangnese, Tianjin dialect)**

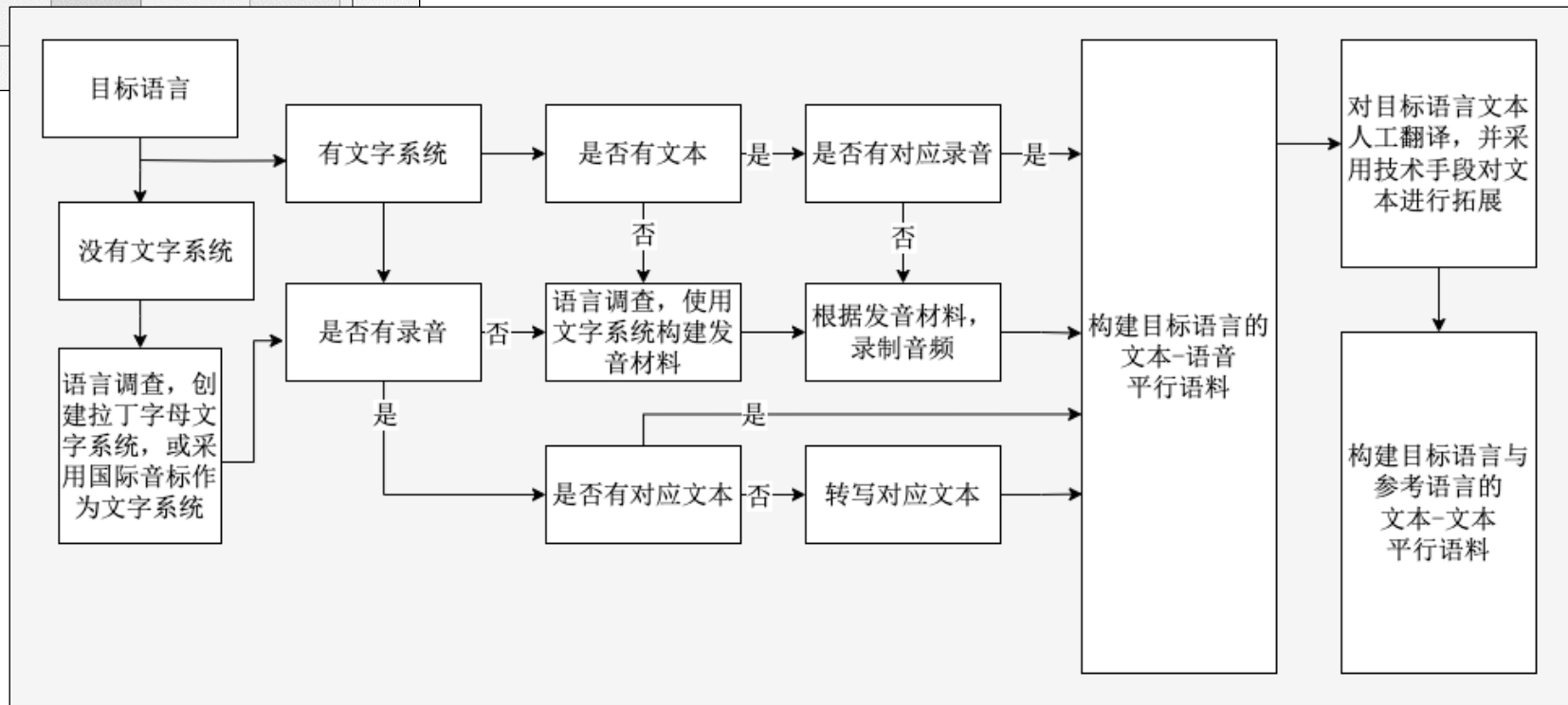
- **Based on the AI technologies displayed above, my team is engaged in 2 large projects:**
- **1-AI-driven Language preservation(especially for endangered languages or dialects)**

- **A huge project: full-chain integrated solution**



• A variety of methods and approaches





- **2-AI-driven Reconstruction of history/language/oral culture**
- **Technology of digital human**

- Synthesizing oral culture audios/videos
- Chanting of Weidong Zhang（张卫东）
- Chanting of official Mandarin（国子监官韵吟诵）
- ancient Chinese prose（古文）
- Chinese poetry（诗词）

- 《论语·学而》：子曰：“学而时习之，不亦说乎？有朋自远方来，不亦乐乎？人不知而不愠，不亦君子乎？”
- The Analects of Confucius; Mencius
- Original sound and synthesized sound

- ☆ **Reconstructing historic audio and video**
- **Hu Shih (胡适, 1891-1962), Yuan-Ren Chao(赵元任, 1892-1982)**
- **New Year's speech, 1935**
- 新年前的两日，我正在作长途的旅行。寂寞的旅途是我最欢迎的，因为平常某日有应作的事，有不能不见的客，很少有整天可以自由用来胡思乱想的。这两天在火车上，因为要替《大公报》写新年的第一篇星期论文。我一路上想的是“我盼望我们这个国家在这新开始的一年里可以做到的几件什么事?”……
- **Yuan-Ren Chao:**
- **Widen history window.**
- **Ethical standards for artificial intelligence issued by Ministry of Science and Technology.**
  （https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html）

- **Essential mechanism of Large Language Model:**
- **3 basic technology:**
- **(1)Semantic Vector Representation of Text(Word embedding)**
- **(2)Next token prediction**
- **(3)Transformer(attention mechanism)**

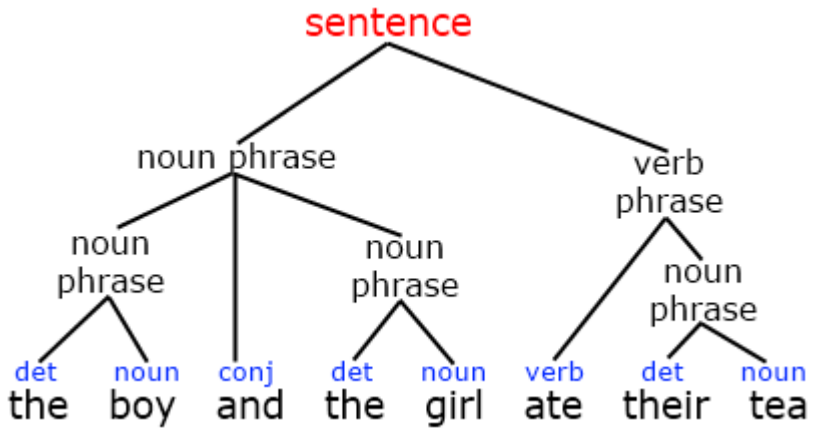
- **Discrete Continuous Metric Space (Semantic Space)**

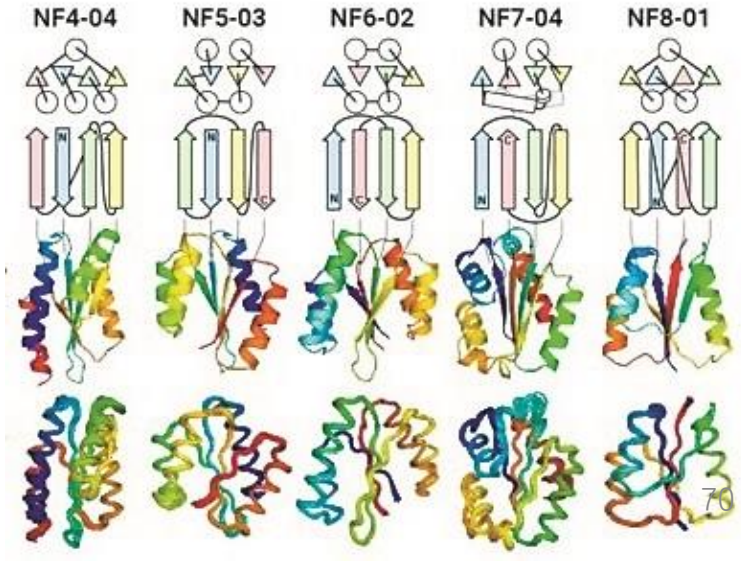
- **Conditional probability distribution of word vectors on high-dimensional space manifolds**

- **Chomsky vs. Hinton**
- **“basically high-tech plagiarism”**

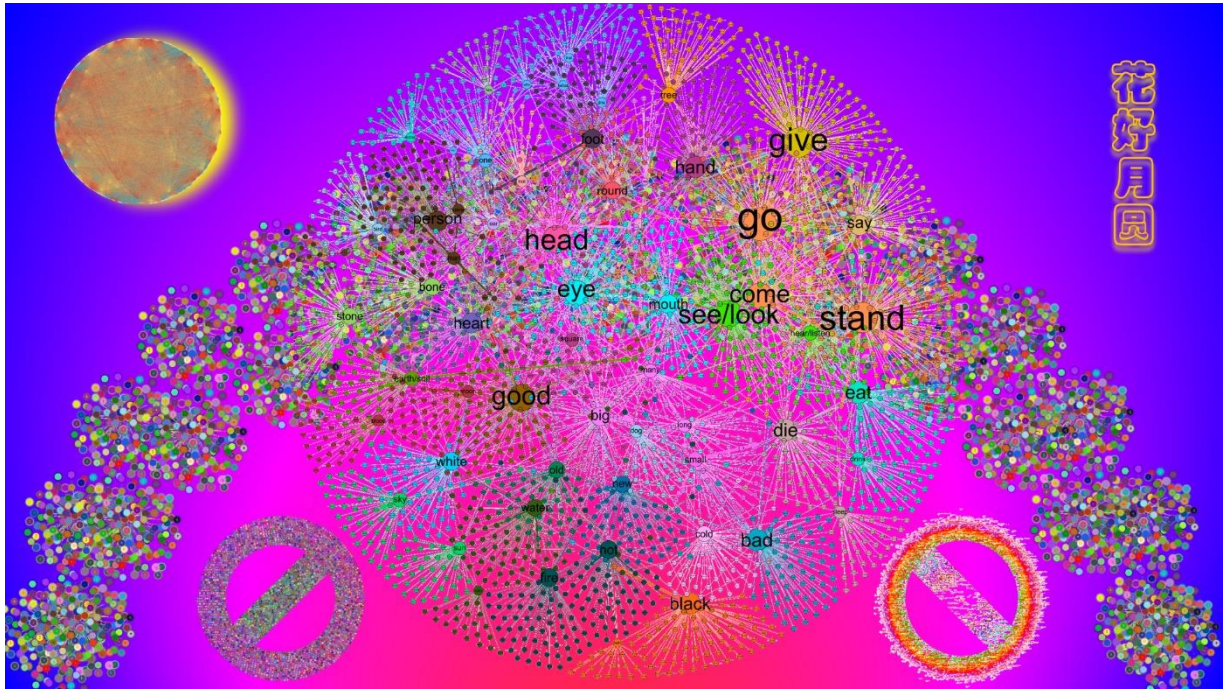
- **AIGC(Artificial Intelligence Generated Content)**
- **Generative linguistics (Transformational-generative grammar)**

- AlphaFold 3
- “AlphaFold is a model that is capable of high-accuracy prediction of complexes containing nearly all molecular types present in the Protein Data Bank.”
- amino acids: 20
- protein: Proteins have complex and diverse spatial conformations. Their primary structure is a linear arrangement of amino acids, while secondary structures such as alpha helices and beta folds are further folded to form tertiary structures. Multiple subunit combinations form quaternary structures. This precise structure endows proteins with specific functions.
- AlphaFold inspired Large Language Model
- Interesting!





- **AI: Artificial Intelligence**
- **‘AI’: Attractive, Interesting**
- **If we weren't all so interested in ourselves, life would be so uninteresting we couldn't endure it. (Arthur Schopenhauer, 1788 - 1860)**
- **AI is ‘AI’.**
- **Fascinating Aspects of Language**
- **Happy Chinese spring festival!**



香港城市大學
City University of Hong Kong

南開大學
Nankai University

Thanks for listening.

Looking forward to your comments and suggestions.

Qibin Ran
1@ranqibin.com