

On Language & Computation

(...and three proposals)

By Ekaterina (Kat) Vylomova

vylomovae@unimelb.edu.au

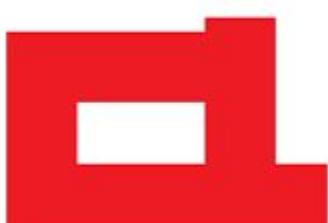


Natural Language Processing and Computational Linguistics



Jason Eisner (JHU):

NLP: is the **art of solving engineering problems** that need to analyze (or generate) natural language text. Here, the metric of success is not whether you designed a better scientific theory or proved that languages X and Y were historically related. Rather, the **metric is whether you got good solutions on the engineering problem.**



ACL

Founded in 1962

CL is analogous to computational biology or any other computational fill-in-the-blank. It develops **computational methods to answer the scientific questions of linguistics.**

Natural Language Processing and Computational Linguistics



Jason Eisner (JHU):

NLP: is the **art of solving engineering problems** that need to analyze (or generate) natural language. The metric of success is not whether you came up with a good theory or proved that language is related to some other field. Rather, the **metric is on the engineering problem**. Over 95%* of current research in ACL (relies on SOTA, leaderboards; task-specific and industry-oriented research)



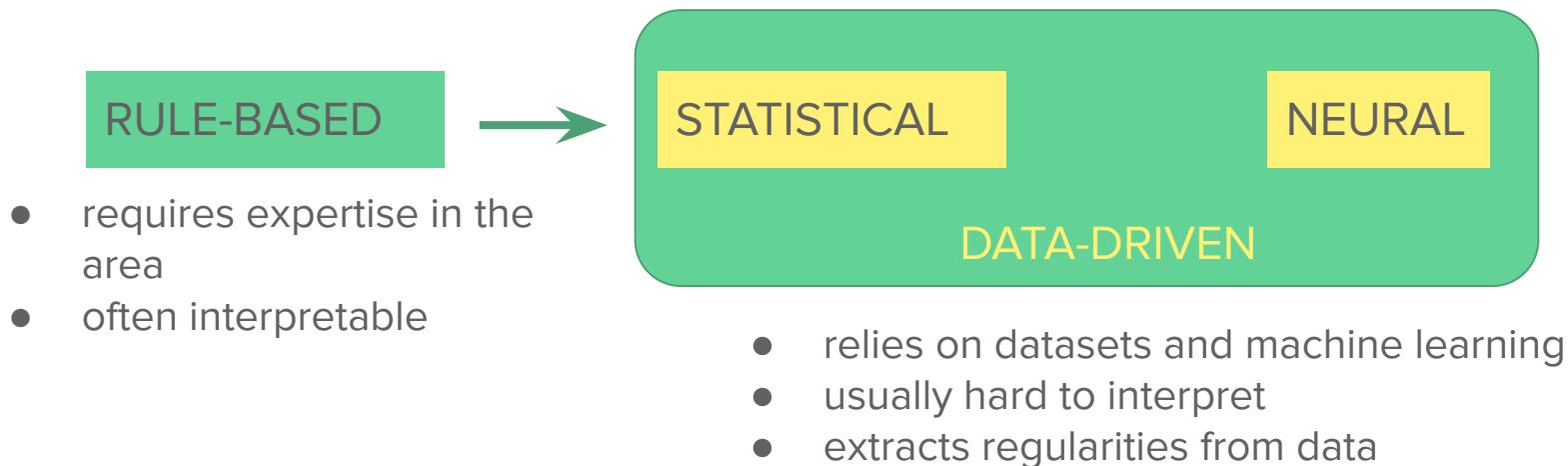
ACL

Founded in 1962

CL is analogous to computational biology or any other computational science. It develops **computational methods to answer the scientific questions of linguistics**.

* a rough estimate from ACing and reviewing

A *Very* Brief History of NLP



A *Very* Brief History of NLP

RULE-BASED



STATISTICAL

NEURAL

- requires expertise in the area
- often interpretable

- relies on datasets and machine learning
- usually hard to interpret
- extracts regularities from data

“Every time I fire a linguist, the performance of our speech recognition system goes up.”*

* (often attributed to) Frederick Jelinek, *the pioneer of speech recognition*



Three Whales on Modern NLP

Draft saved

Recipients

Subject

From Sign to Cosine: on Language and Compu

Sine 

GMail



You

Which sentence is more likely "From sign to cosine: on language and computation" or "From sine to cosine: on language and computation"?



ChatGPT

The more likely sentence would be "From sign to cosine: on language and computation." In mathematical contexts, the term "sign" typically refers to the sign function, which yields -1, 0, or 1 depending on the sign of its argument. In contrast, "sine" and "cosine" are trigonometric functions. So, "sign to cosine" seems to follow a progression of mathematical functions more smoothly than "sine to cosine."

ChatGPT



Which sentence is more likely "From sign to cosine: on language and computation" or "From sine to cosine: on language and computation"?

Show drafts  



The more likely sentence is "From sine to cosine: on language and computation."

Here's why:

- Sine and cosine are well-established trigonometric functions used in mathematics and computer graphics.
- Sign, while a mathematical concept, doesn't directly connect to the realm of computation in the same way.

While it's possible the title uses "sign" in a more abstract way, referring to the symbolic nature of language, "sine" makes a clearer connection to the computational aspect mentioned in the subtitle.

Gemini

Draft saved

Recipients

Subject

From Sign to Cosine: on Language and Compu

Sine 

GMail



Which sentence is more likely "From sign to cosine: on language and computation" or "From sine to cosine: on language and computation"?

Show drafts  



The more likely sentence is "From sine to cosine: on language and computation."

Here's why:

- Sine and cosine are well-established trigonometric functions used in mathematics and computer graphics.
- Sign, while a mathematical concept, doesn't directly connect to the realm of computation in the same way.

While it's possible the title uses "sign" in a more abstract way, referring to the symbolic nature of language, "sine" makes a clearer connection to the computational aspect mentioned in the subtitle.



You

Which sentence is more likely "From sign to cosine: on language and computation"?

Why is “Sign” less preferred?

Gemini



ChatGPT

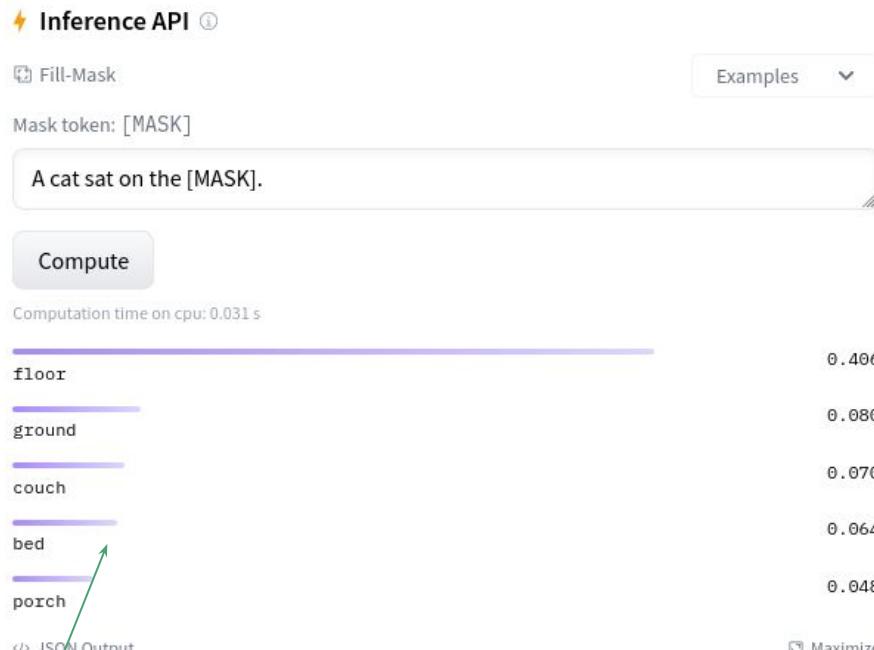
The more likely sentence would be "From sign to cosine: on language and computation." In mathematical contexts, the term "sign" typically refers to the sign function, which yields -1, 0, or 1 depending on the sign of its argument. In contrast, "sine" and "cosine" are trigonometric functions. So, "sign to cosine" seems to follow a progression of mathematical functions more smoothly than "sine to cosine."

ChatGPT

[https://huggingface.co/google-bert/bert-base-uncased?text=A+cat+sat+on+the+\[MASK\]](https://huggingface.co/google-bert/bert-base-uncased?text=A+cat+sat+on+the+[MASK])

Whale#1: Language Models

revolutionalised
modern NLP!



Models that assign probabilities to upcoming strings (words or characters), or sequences of strings in general.

Vocabulary Words

Their probability (given the context)

N-gram Language Models

Predicting the most probable next character based on n previous characters (order)

Generating “Harry Potter” character-by-character:

Order=4, Prompt= “Harr”

“Harry Potion off his eyes for the Philosophers Stone - J.K. Rowling up the”

“Harry, who’s better and passed they saw than unting. “I’ve doors sleep. Se”

Order=10, Prompt= “Harry saw ”

“Harry saw his scared white face look down at the ground. Hagrid could see a hug”

“Harry saw his scar. “I might get lucky again.””

Many English words, partially interpretable syntax. Nothing beyond character collocations!

Whale#2: Distributional Semantics

“You shall know a word by the company it keep” (Firth, 1957)

A meaning of a word is a distribution of its contexts.



J.R. Firth

<https://jramkiss.github.io/2019/08/21/word-embeddings/>

<https://arize.com/blog-course/embeddings-meaning-examples-and-how-to-compute/>

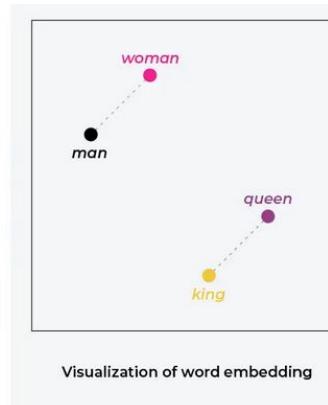
Whale#3: Vector Space Representations

$$X = \begin{matrix} & I & like & enjoy & deep & learning & NLP & flying & . \\ I & \left[\begin{matrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{matrix} \right] \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{matrix}$$

Word co-occurrence:
sparse, big, not generalisable



word	Word embedding						
man	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
woman	0.7	0.3	0.8	-0.7	0.1	-0.5	-0.4
king	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
queen	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9



The output of LM task (prediction-based):
dense, real-valued, better
generalisation

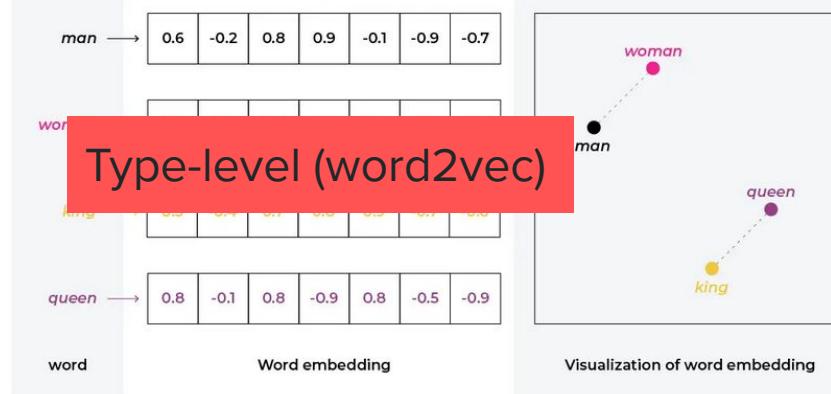
<https://jramkiss.github.io/2019/08/21/word-embeddings/>

<https://arize.com/blog-course/embeddings-meaning-examples-and-how-to-compute/>

Whale#3: Vector Space Representations

$$X = \begin{matrix} & I & like & enjoy & deep & learning & NLP & flying & . \\ I & \left[\begin{array}{ccccccc} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{array} \right] \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{matrix}$$

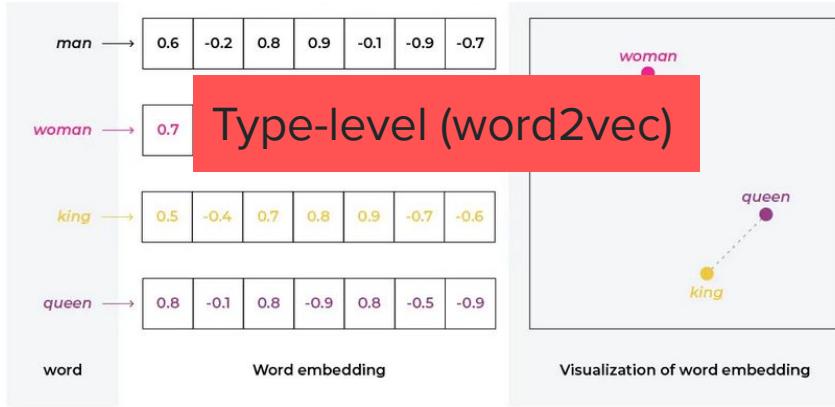
Word co-occurrence:
sparse, big ($V \times V$), not
generalisable



Type-level (word2vec)

The output of LM task (prediction-based):
dense ($V \times \text{dim}$), real-valued, better
generalisation

Contextualising: Resolving Polysemy



Surrounding contexts:

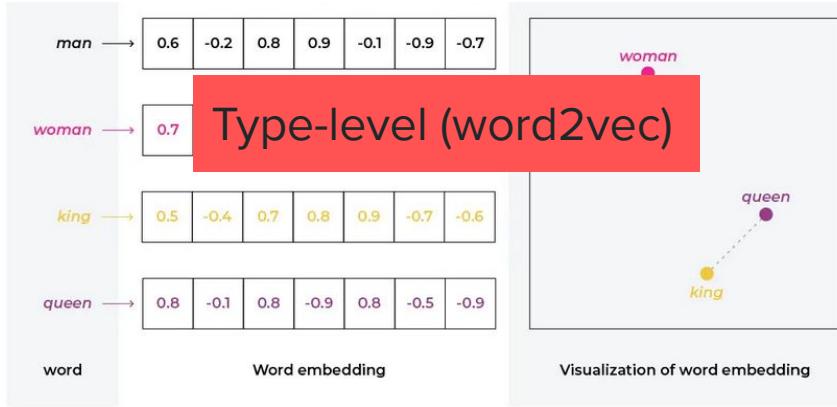
- The plays, presented in celebration of the 100th anniversary of O'Neill's birth, **may** go on to New York City....
- Microsoft browser bug **may** access private files.
- The \$100,000 award, to be presented to Taylor in **May** by the president of Israel,

↓ **language model (DistilBERT, BERT, GPT-2)**

Contextualized word embeddings:

- **may**: <-0.581083, 0.029549, -0.092858, -0.063634, -0.547897, 0.004871, ... >
- **may**: <0.744181, -1.465255, 0.104348, -1.066828, 0.30615, -0.023076, ...>
- **may**: <-1.087184, -0.137186, -0.637199, -0.422636, -0.638509, 0.32379, ... >

Contextualising: Resolving Polysemy



Surrounding contexts:

- The play—presented in celebration of the 100th anniversary of
- M
- T
- p

Token-level, or contextualised
(BERT, GPT, ...)

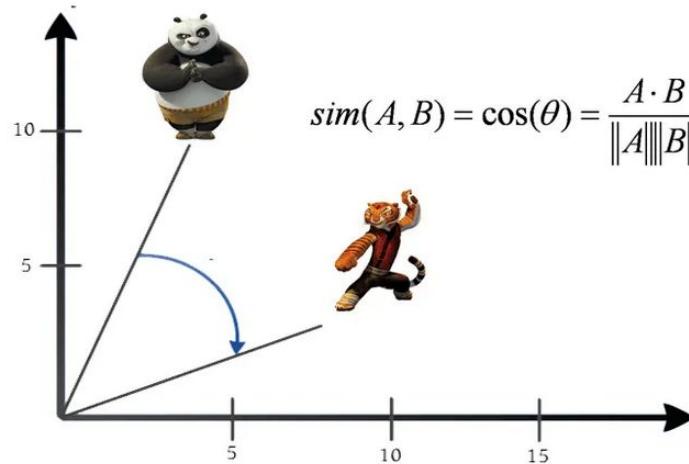
↓ language model (DistilBERT, BERT, GPT-2)

Contextualized word embeddings:

- may: <-0.581083, 0.029549, -0.092858, -0.063634, -0.547897, 0.004871, ...>
- may: <0.744181, -1.465255, 0.104348, -1.066828, 0.30615, -0.023076, ...>
- may: <-1.087184, -0.137186, -0.637199, -0.422636, -0.638509, 0.32379, ...>

Measuring Word- & Sentence-level Similarity

Cosine Similarity



Semantic similarity = 0.93

The **boy** was chased by a **dog**
A **dog** was chasing the **boy**

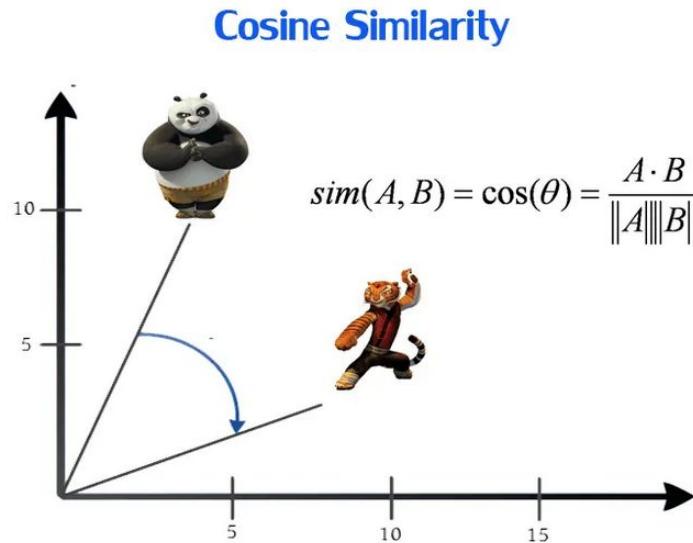
Semantic similarity = 0.68

A **cat** was chasing the **boy**
A **dog** was chasing the **boy**

Semantic similarity = 0.62

A **cat** was chasing the **boy**
The **boy** was chased by a **dog**

Measuring Word- & Sentence-level Similarity



More data → better and more robust models

**Learns many patterns, including spurious correlations and biases
Lacks of interpretability**

Understanding?

Me: Make an image of how a sentence "A cat sat on the mat." is fed into a neural network

Gemini: Sure, here is an image of how the sentence "A cat sat on the mat." is fed into a neural network:



A *Very* Brief History of NLP

RULE-BASED



STATISTICAL

NEURAL

DATA-DRIVEN

- requires expertise in the area
- often interpretable

- relies on datasets and machine learning
- usually hard to interpret
- extracts regularities from data

**Proposal #1: Move from Data-driven to Theory-driven models
(and more interdisciplinary research)**

“The more data you have, the less important it is.”
on
system goes up. ”

* (often attributed to) Frederick Jelinek, *the pioneer of speech recognition*



A Multidimensional Framework for Evaluating Lexical Semantic Change



Lexical Semantic Change

“Semantic change deals with change in meaning, understood to be a change in the concepts associated with a word...” (Campbell, 1999)

Narrowing: superordinate (general) → subordinate (specific)

meat ‘all food’ → *meat ‘edible flesh’*

Widening: subordinate (specific) → superordinate (general)

dogge ‘of specific breed’ *dog ‘all breeds’*

Metaphor: transfer of a name based on associations of similarity

bitraz ‘biting’ → *bitter ‘harsh of taste’*

Metonymy: change based on meaning’s proximity in space or time

ceace ‘jaw’ → *cheek ‘fleshy side face’*

Synecdoche: the meanings are related as whole and part

stobo ‘heated room’ → *stove ‘cooking device’*

Hyperbole: stronger→weaker meaning by overstatement *extonare ‘to strike with thunder’* → *astonish ‘surprise’*

Meiosis: weaker→stronger meaning by understatement

kwalljan ‘to torment’ → *cwellan ‘to kill’*

Degeneration: Positive→negative connotation

cnafa ‘boy servant’ → *knave ‘a tricky deceitful fellow’*

Elevation: Negative→positive connotation

cniht ‘boy, servant’ *knight ‘man honored by sovereign for merit’*

Lexical Semantic Change

“Semantic change deals with change in meaning, understood to be a change in the concepts associated with a word...” (Campbell, 1999)

Narrowing: superordinate’(general) → subordinate (specific)

meat ‘all food’ → *meat ‘edible flesh’*

Widening: subordinate (specific) → superordinate (general)

dogge ‘of specific breed’ *dog ‘all breeds’*

Metaphor: transfer of a name based on associations of similarity

bitraz ‘biting’ → *bitter ‘harsh of taste’*

Metonymy: change based on meaning’s proximity in space or time

ceace ‘jaw’ → *cheek ‘fleshy side face’*

Synecdoche: the meanings are related as whole and part

stobo ‘heated room’ → *stove ‘cooking device’*

Hyperbole: stronger→weaker meaning by overstatement *extonare ‘to strike with thunder’* → *astonish ‘surprise’*

Meiosis: weaker→stronger meaning by understatement

kwalljan ‘to torment’ → *cwellan ‘to kill’*

Degeneration: Positive→negative connotation

cnafa ‘boy servant’ → *knavie ‘a tricky deceitful fellow’*

Elevation: Nega Can NLP tools help to evaluate the changes?

ed by sovereign for merit’

How should the change be measured?

Natural Language Processing (NLP) lacks a unifying framework to integrate multiple dimensions of lexical semantic change

- Detect **Any Change** (Kutuzov et al., 2018; Montanelli & Periti, 2023; Tahmasebi et al., 2019; Tahmasebi & Dubossarsky, 2023)
- Detect specific types of change:
 - **Broadening** (Vylomova et al., 2019)
 - **Metaphor** (Tong et al., 2021)
 - **Hyperbole** (Badathala et al., 2023; Kong et al., 2020; Schneidermann et al., 2023; Tian et al., 2021)
 - **Bleaching** (Luo et al., 2019)
 - **Pejoration** (Dinu et al., 2021)

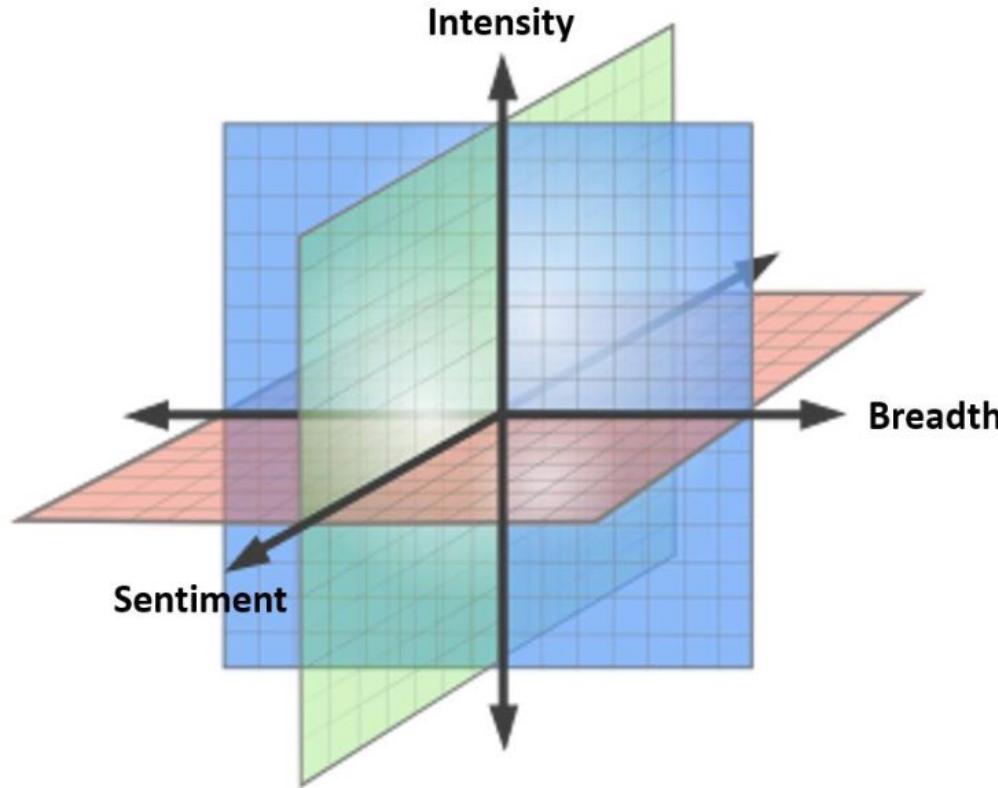
Step 1. Making a more compact model

Dimensions of Lexical Semantic Change

Dimension	Rising	Falling
Sentiment	Elevation (Bloomfield, 1933); Amelioration (Ullmann, 1962)	Degeneration (Bloomfield, 1933); Pejoration (Ullmann, 1962)
Breadth	Widening (Bloomfield, 1933; Ullmann, 1962); Generalization of meaning (Blank, 1999); Horizontal Creep (Haslam, 2016)*;	Narrowing (Bloomfield, 1933; Ullman, 1962); Specialization of meaning (Blank, 1999)
Intensity	Meiosis (Bloomfield, 1933)	Hyperbole (Bloomfield, 1933); Vertical Creep (Haslam, 2016)*

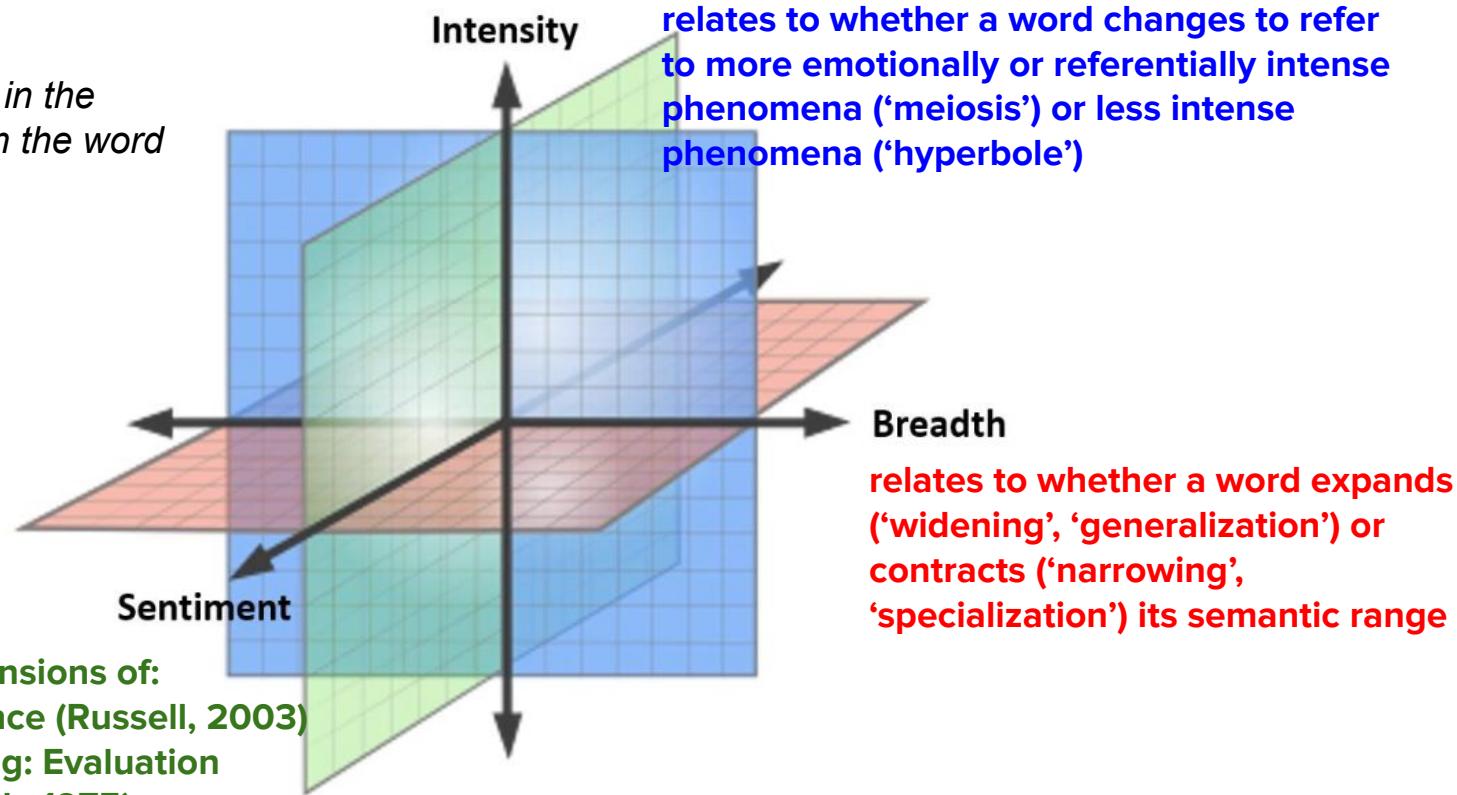
Note: * = specific to harm-related concepts

Three Major Dimensions of Semantic Change



Three Major Dimensions of Semantic Change

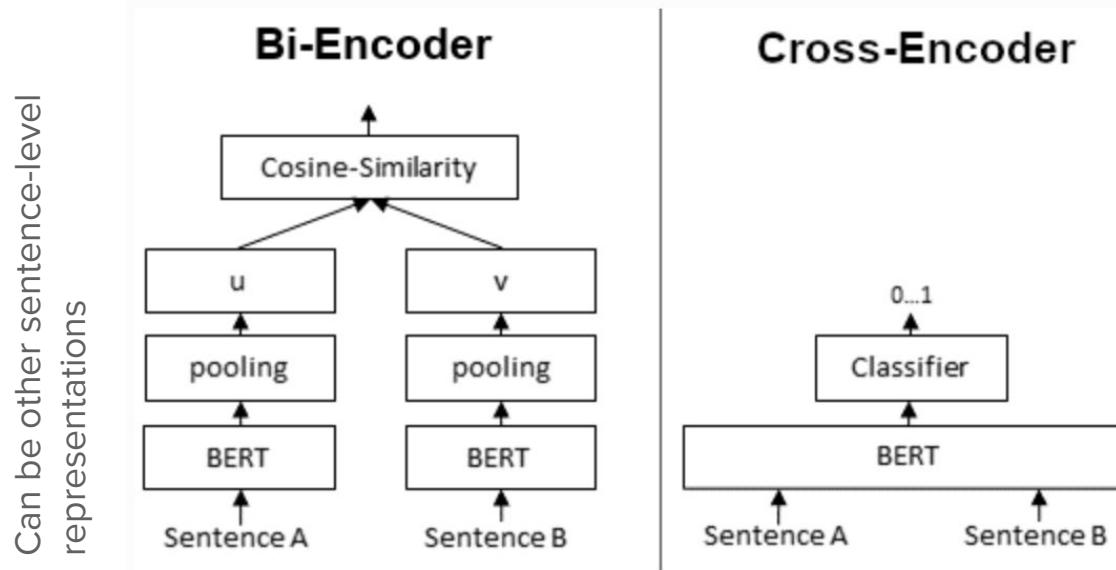
Thematic Content: shifts in the specific contexts in which the word is used



Step 2. Operationalisation of each type of change

Operationalisation: Breadth

Semantic broadening of the target concept was evaluated as average cosine (dis)similarity between sentence level embeddings containing the target term



Operationalisation: Valence and Arousal

Method: Valence and Arousal norms from Warriner et al. (2013), which provide ratings for 13,915 English lemmas collected from 1,827 United States residents, ranging from **low valence** (1: feeling extremely "unhappy", "despaired") to **high valence** (9: feeling extremely "happy", "hopeful") and **low arousal** (1: feeling "calm", "unaroused") to **high arousal** (9: feeling "agitated", "aroused")

Sentiment → valence

Intensity → arousal

Operationalisation: Intensity (Modifier)

How: proportion of adjective modifiers of the target concept that amplify its meaning

Method: intensifiers that reflect adjectives modifying nouns to amplify their meaning “great”, “intense”, “severe”, “harsh”, “major”, “extreme”, “powerful”, “serious”, “devastating”, “destructive”, “debilitating”

Case Study: Concept Creep

“Concept creep” is the gradual semantic expansion of harm-related concepts such as bullying, harassment, mental disorder, prejudice, and trauma (1980–2020s).

It takes two forms:

Intensity

1) **Vertical creep:** “concept’s meaning becomes less stringent, extending to quantitatively milder variants of the phenomenon to which it originally referred” (Haslam, 2016)

traumatic event ‘distressing and abnormal human experience’ (rape) included milder items (business loss)



Breadth

2) **Horizontal creep:** “concept extends to a qualitatively new class of phenomena or is applied in a new context” (Haslam, 2016)

trauma ‘physical wound’ trauma including psychological injuries too

Case Study: Concept Creep

Psychology Abstracts: The corpus comprises abstracts from journals in the field of psychology covering the period of 1930–2019 that were collected from the E-Research and the PubMed databases. In total, there are 871,340 abstracts from 875 journals resulting in 133,082,240 tokens. We mainly focus our study on the 1970–2016 period

General (CoHA+CoCA): The corpus of historical English (CoHA) starts in the 1810s and ends in the early 2000s, comprising 400 million words from 115,000 texts evenly sampled for each decade from fiction, magazines, newspapers, and non-fiction books. The corpus of contemporary American English (CoCA) covers the period from 1990 till 2019 and contains about 1 billion words from 500,000 texts evenly sampled from spoken, TV shows, academic journals, fiction, magazines, newspapers, and blogs

Broadening of “Trauma”, “Addiction”, “Harassment”

Semantic broadening via cosine similarity between sentence-level representations

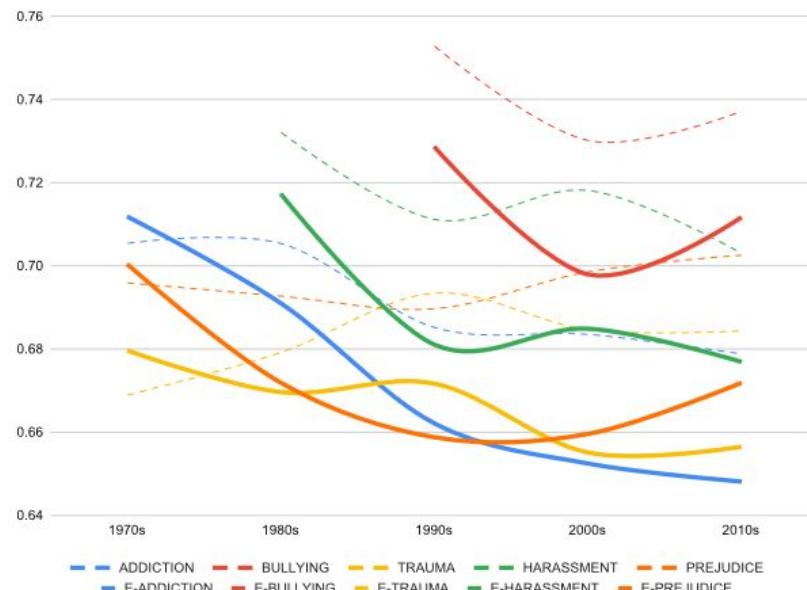
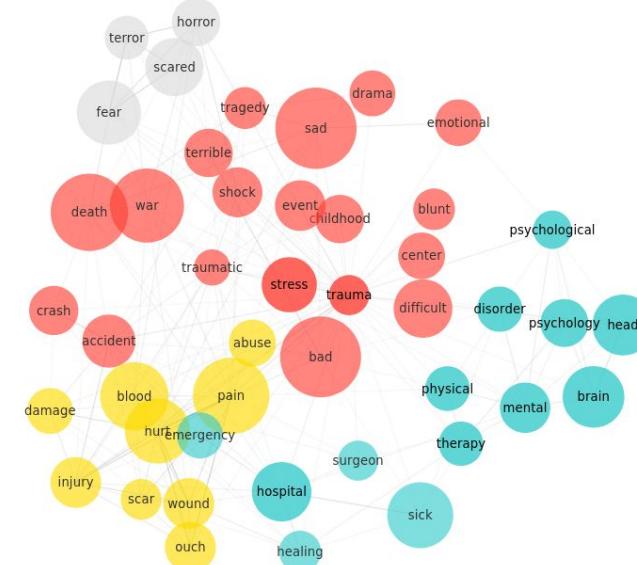
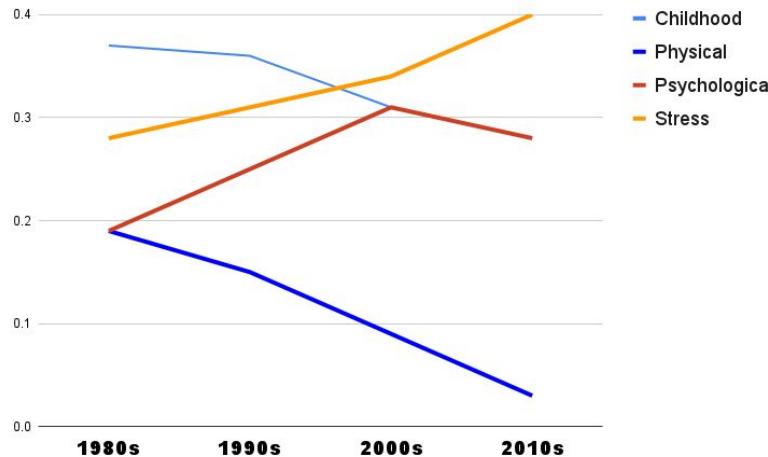


Figure 2.4: Mean cosine similarities (polynomial smoothing) over five decades (psychology abstracts corpus). Bold and dashed lines correspond to epoch-specific (e^*) and global (static) embeddings, respectively.

Concept Creep (Haslam, 2016): Trauma

Childhood/Physical→Emotional/Stress

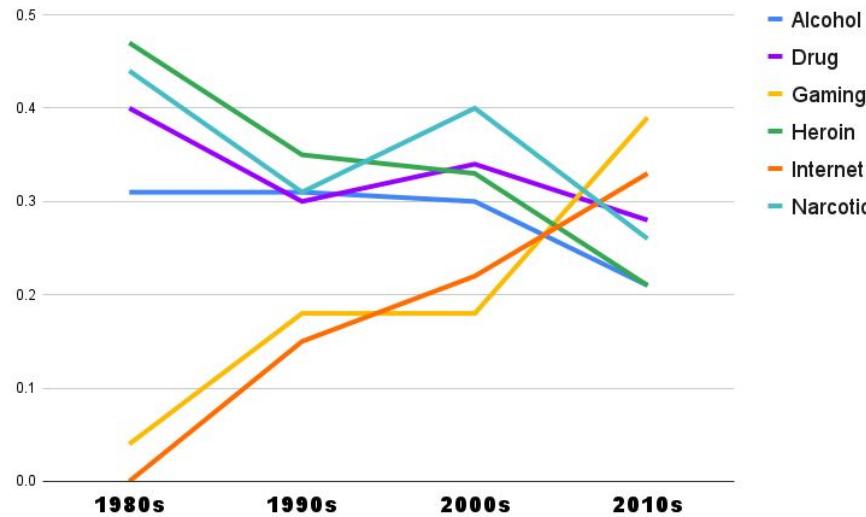


Top Nearest Neighbors (Psychology)

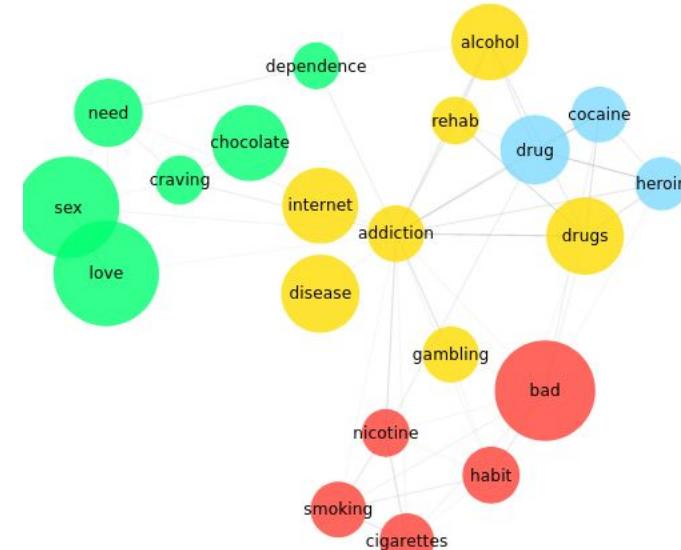
Word Associations

<https://smallworldofwords.org/en/project/visualize>

Concept Creep (Haslam, 2016): Addiction Substance→ Behaviour



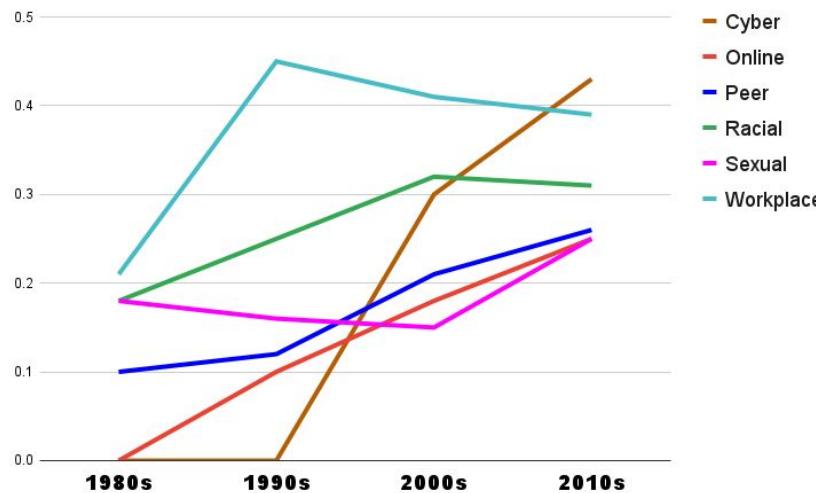
Top Nearest Neighbors (Psychology)



Word Associations

<https://smallworldofwords.org/en/project/visualize>

Concept Creep (Haslam, 2016): Harassment Workplace → Online



Top Nearest Neighbors (Psychology)

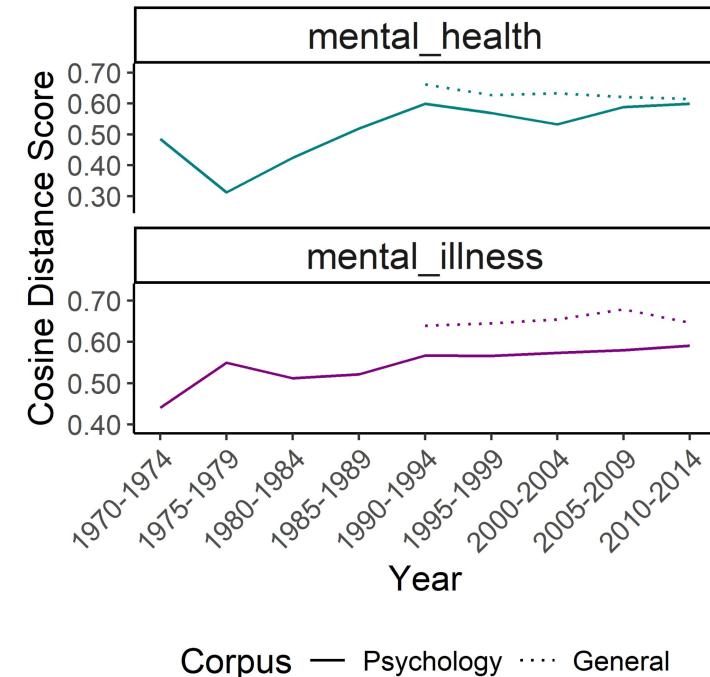


Word Associations

<https://smallworldofwords.org/en/project/visualize>

Broadening of “Mental health”, “Mental illness”

Significant increasing trends for *mental health* and *mental illness* in the psychology corpus, but no trends in the general corpus



Decrease in Valence of “Mental health”

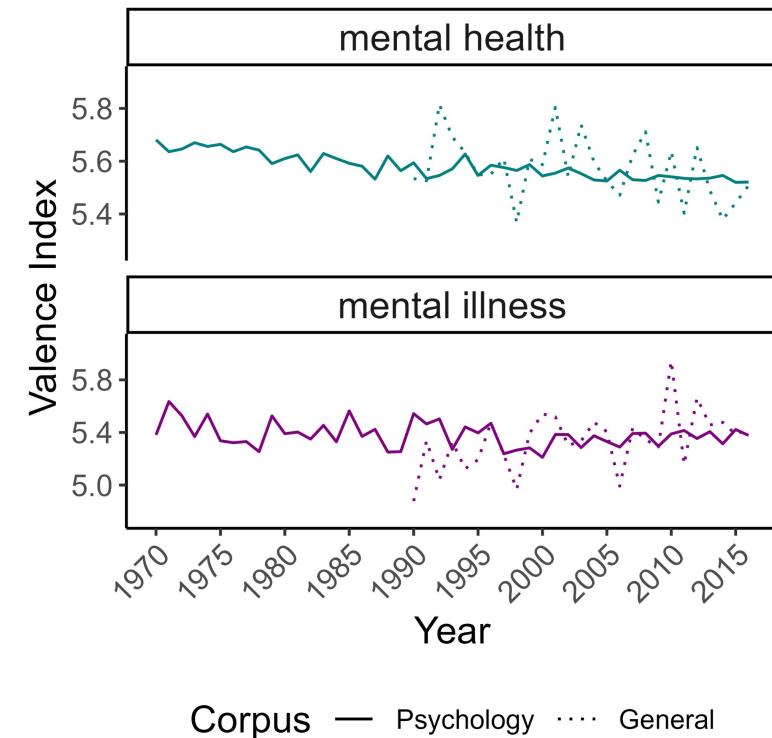
Significant declining trend in the valence of words used in the context of *mental health* in the psychology corpus and no trend in the general corpus

1980s

195 community MH
30 use MH
26 physical MH
20 infant MH
19 child MH
17 need MH
17 health MH
15 preventive MH
14 system MH
14 relate MH
14 poor MH
14 group MH
14 general MH
14 concern MH
13 relationship MH
13 provide MH
13 measure MH
12 influence MH
12 implication MH
11 role MH

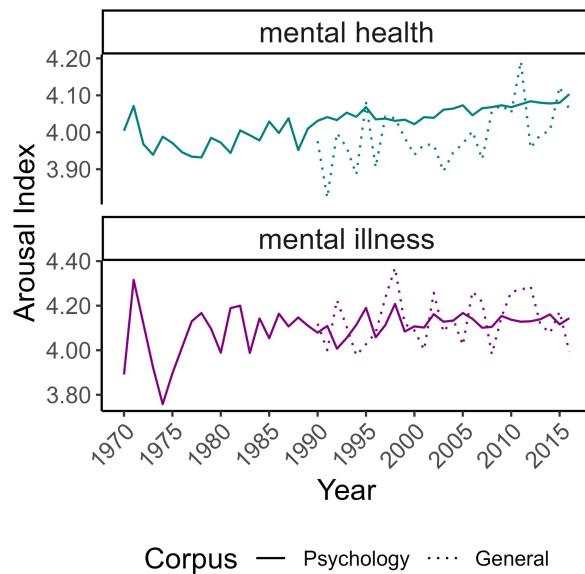
2010s

826 physical MH
597 child MH
516 adolescent MH
495 poor MH
471 community MH
368 use MH
324 improve MH
282 associate MH
281 base MH
260 positive MH
239 health MH
231 youth MH
222 adult MH
221 and MH
214 relate MH
214 maternal MH
210 experience MH
205 risk MH
199 forensic MH
191 school MH

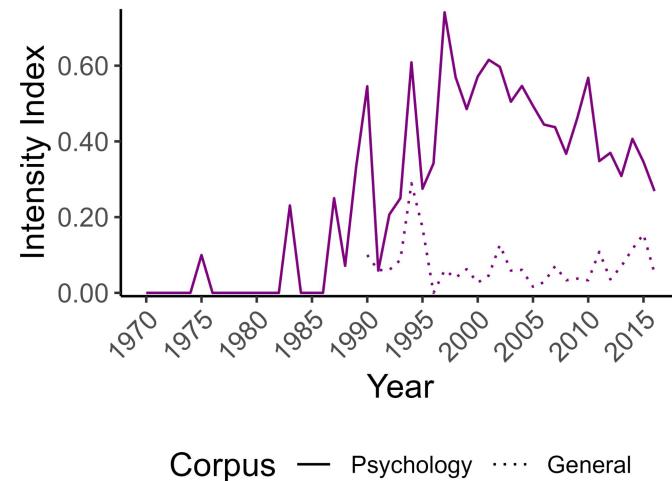


Increase in Arousal in “Mental health” and Rise of Intensifiers in “Mental illness”

Increasing trend in the arousal index “mental health” in psychology and the general corpus.



Proportion of mentions of “mental illness” with one of intensifying adjectives (“great”, “intense”, “severe”, “harsh”, “major”, “extreme”, “powerful”, “serious”, “devastating”, “destructive”, “debilitating”) as amod



Increase in Arousal in “Mental health” and Rise of Intensifiers in “Mental illness”

1980s	1990s	2000s	2010s
18 attitude MI	105 severe MI	262 severe MI	360 severe MI
7 major MI	24 chronic MI	119 people MI	203 people MI
7 chronic MI	22 persistent MI	66 person MI	118 individual MI
6 risk MI	21 person MI	40 individual MI	105 person MI
6 physical MI	21 people MI	36 persistent MI	71 stigma MI
5 cause MI	19 history MI	36 major MI	71 history MI
5 attribution MI	10 retardation MI	34 history MI	61 parent MI
4 severe MI	9 treatment MI	30 attitude MI	58 adult MI
4 perception MI	9 relationship MI	28 diagnose MI	56 attitude MI
4 level MI	9 major MI	25 symptom MI	54 offender MI
4 label MI	9 form MI	25 patient MI	53 experience MI
4 history MI	8 symptom MI	24 stigma MI	43 patient MI
4 conception MI	8 attitude MI	23 treatment MI	42 chronic MI
3 view MI	8 adult MI	20 chronic MI	41 treatment MI
3 symptom MI	8 abuse MI	19 parental MI	38 symptom MI
3 rate MI	7 woman MI	18 prevalence MI	38 associate MI
3 patient MI	7 physical MI	17 suffer MI	36 parental MI
3 opinion MI	7 diagnosis MI	17 adult MI	33 burden MI
3 live MI	6 suffer MI	16 parent MI	33 belief MI
3 demonizing MI	6 stigma MI	16 disorder MI	32 diagnose MI

Operationalisation: Thematic content

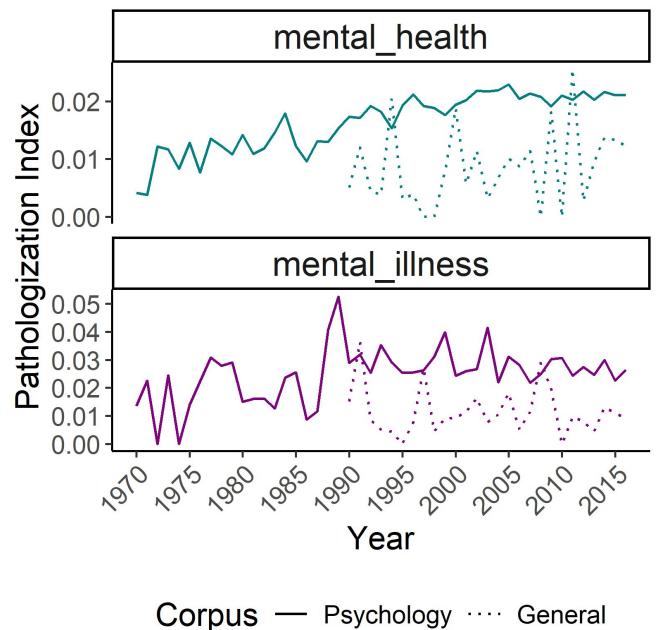
Theme of interest: Pathology, given concerns about pathologization of target concepts (Brinkmann, 2016).

Dictionary: **Pathologization dictionary** (developed by Baes et al., 2023) to compute pathologization index

- Generate disease-related words with restricted range in meaning: “clinical”, “disorder”, “symptom”, “illness”, “pathology”, “disease”
- List their forward word associations from the English Small World of Words project (De Deyne et al., 2019) and remove duplicates
- Filter the list for lemmas reflecting pathologization (i.e., to view or characterize as medically or psychologically abnormal)
“ailment”, “clinical”, “clinic”, “cure”, “diagnosis”, “disease”, “disorder”, “ill”, “illness”
“medical”, “medicine”, “pathology”, “prognosis”, “sick”, “sickness”, “symptom”,
“treatment”
- Compute the pathologization index: relative frequency of disease-related collocates of the target term

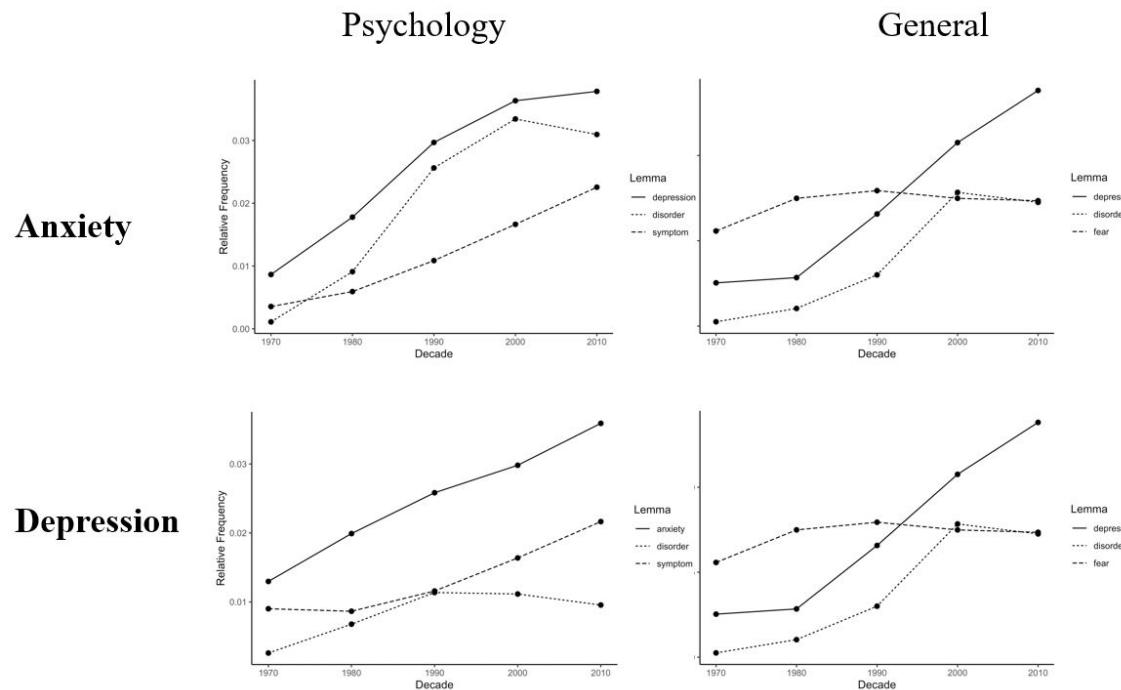
Rising pathologization of “Mental health” and “Mental illness”

Mental health and *mental illness* become significantly more associated with pathology-related terms in the psychology corpus, but not in the general corpus



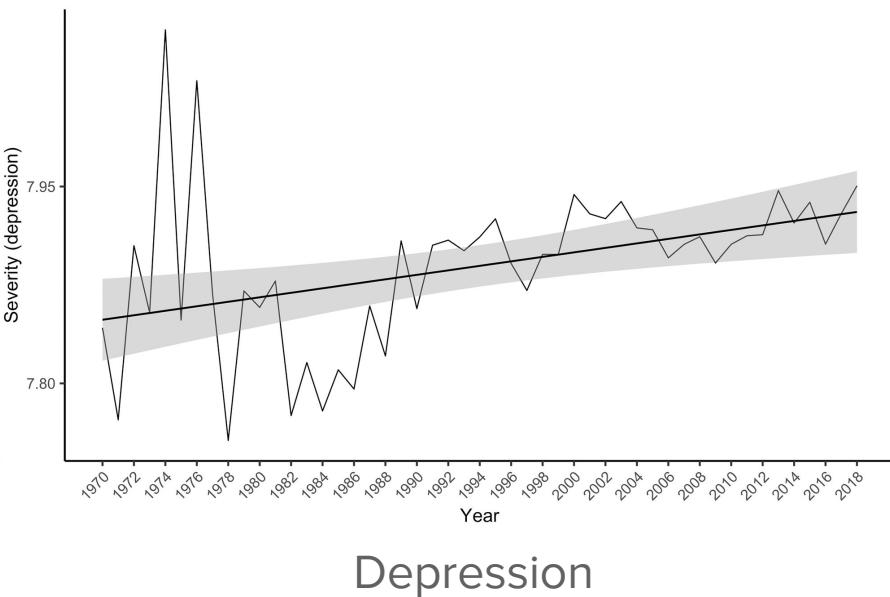
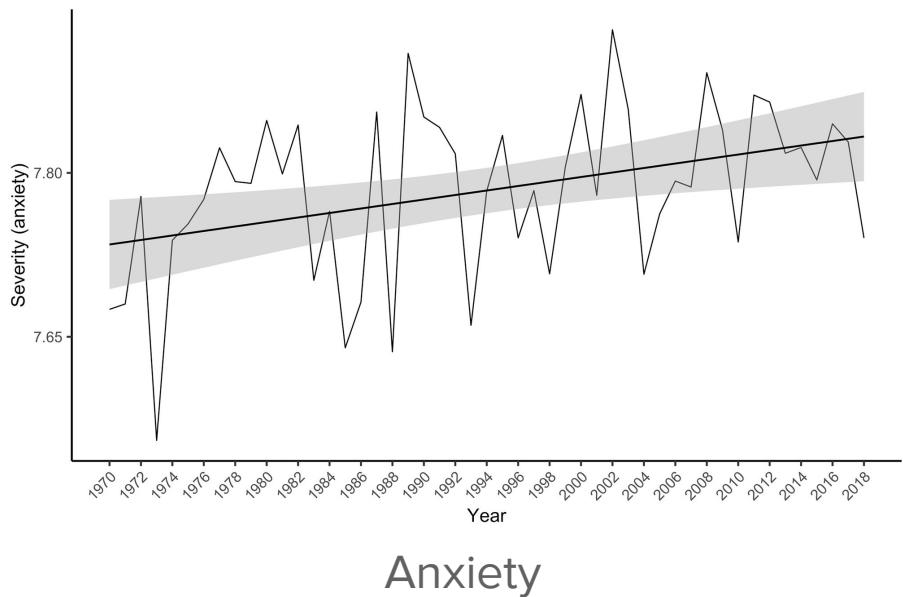
Rising pathologization of anxiety and depression

Clinical terminology (e.g., “disorder”, “symptom”) appeared frequently in the context of target terms



Rising intensity of anxiety and depression

The meanings of anxiety and depression showed significant increases in their emotional intensity from 1970 onwards in a corpus representing psychology and general US English text



NLP for Field Linguistics and Linguistic Typology

Proposal #2: incorporate NLP models to make linguistic data collection, processing and interpretation more efficient



Lexical Elaboration

How to find words that are not already recorded in the dictionary? How to get more nuanced meanings?

- Story telling
- Going through domain-specific terms (kins terms, place names)



Here, instead we propose a method for **creating possible words that can be confirmed by speakers** (or not) as being words in the language.

Lexical Elaboration

Task: Given short wordlists (no translations/texts), produce plausible words (lemmas), increasing the likelihood of hits (i.e. generated forms being real words).

<https://paradisec.org.au/wordgen/wg.php>

Word generator(v1.2) Info & updates

Abstraction factors wildcard+ 70 **Required confidence, %** 100% Reset master

N-Gram size Generated Word Count (max. 10 syllables)

3 4 5 200 500 1000 2000 5000

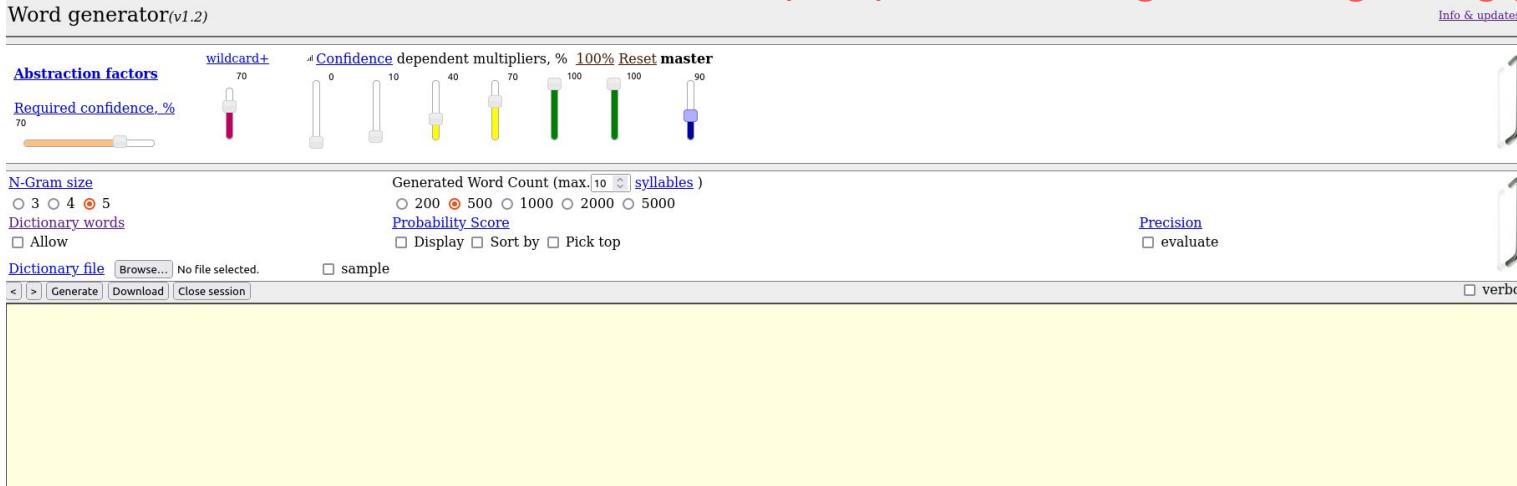
Dictionary words Probability Score

Allow Display Sort by Pick top

Dictionary file Precision evaluate

No file selected. sample verbose

"/>



Lexical Elaboration

Task: Given short wordlists (no translations/texts), produce plausible words (lemmas), increasing the likelihood of hits (i.e. generated forms being real words).

<https://paradisec.org.au/wordgen/wg.php>

E.g., based on English Swadesh list (woman, man person, child, wife, husband, mother, father), around 200 items

Word generator(v1.2)

Abstraction factors

wildcard+ Confidence dependent multipliers, % 100% Reset master

Required confidence, % 70

N-Gram size

Generated Word Count (max. 10 syllables)

3 4 5 200 500 1000 2000 5000

Dictionary words

Probability Score

Allow

Display Sort by Pick top

Dictionary file Browse... en-words sample

< > Generate Download Close session

Direct link to this variant settings

an	bar	be
becaus	bell	bellow
bellower	bir	bit
blo	blowe	blower
bon	breas	breat
breath	breather	breathere
bur	clo	cor
cou	coun	cound
countain	dir	dirt
drim	eart	fath
fathe	fathe	fe
fea	feart	fearth
feat	feath	feathe
feathere	fig	figh
fin	fing	fingern
fir	fiver	flo

Full list:

https://en.wiktionary.org/wiki/Appendix:English_Swadesh_list

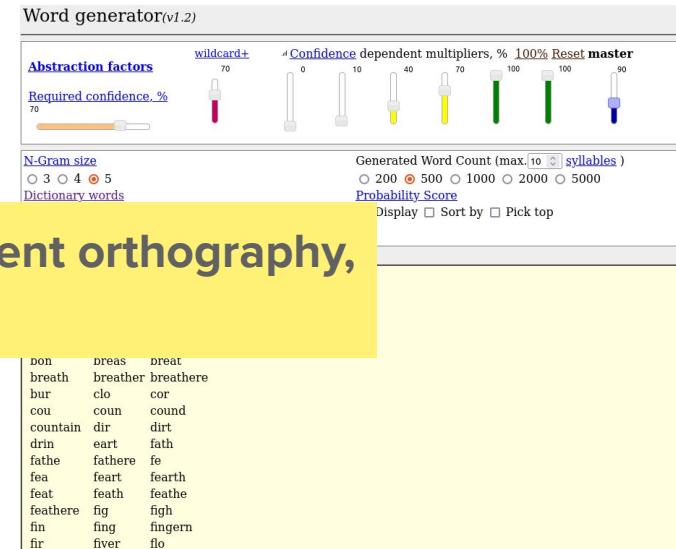
Lexical Elaboration

Task: Given short wordlists (no translations/texts), produce plausible words (lemmas), increasing the likelihood of hits (i.e. generated forms being real words).

<https://paradisec.org.au/wordgen/wg.php>

E.g., based on English Swadesh list (woman, man person, child, wife, husband, mother, father), around 200 items

Note: English has less transparent orthography, with many conventions!

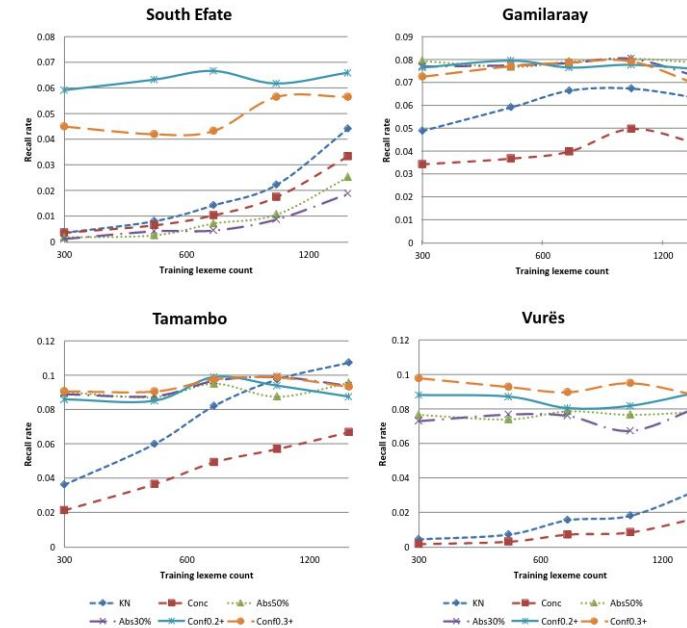
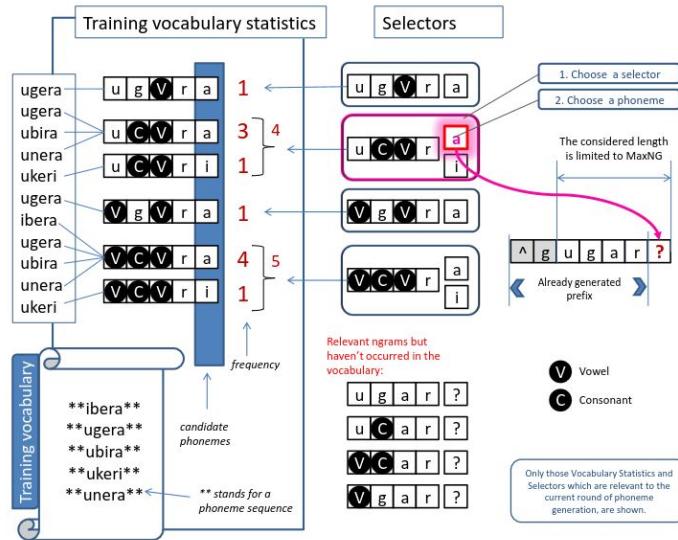


Full list:

https://en.wiktionary.org/wiki/Appendix:English_Swadesh_list

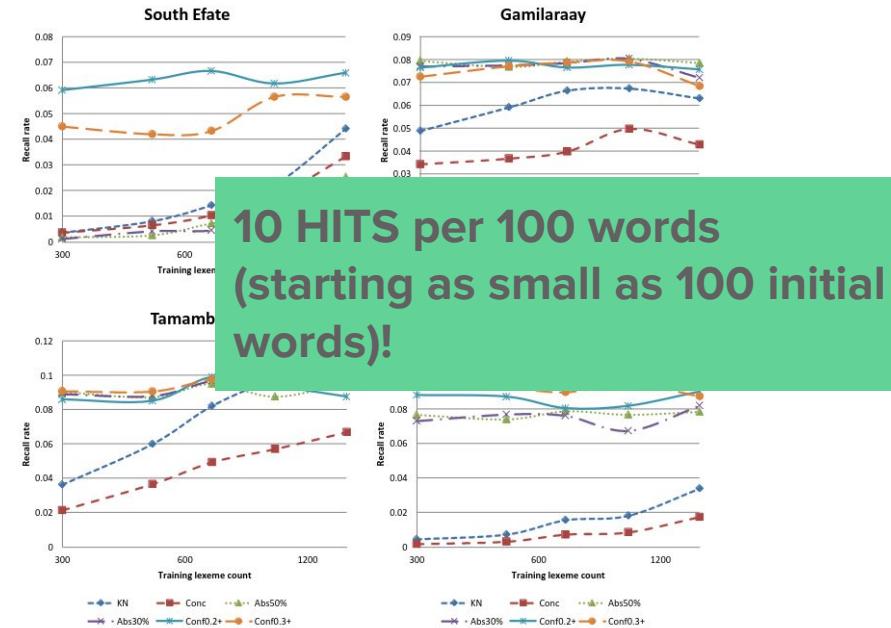
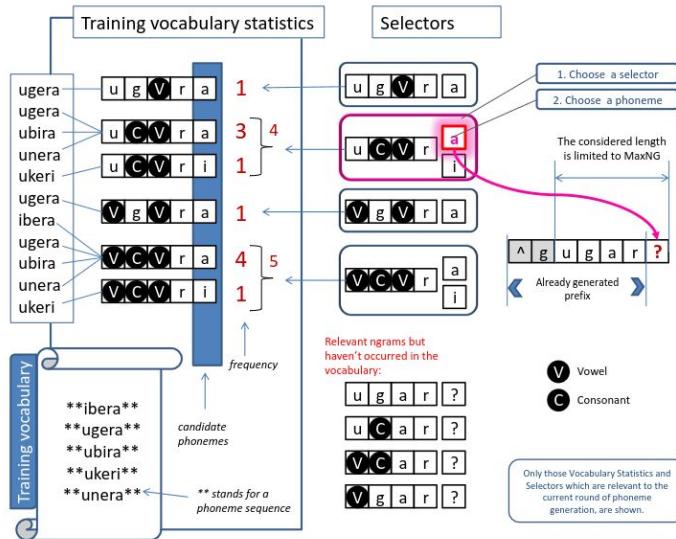
Lexical Elaboration

Task: Given short wordlists (no translations/texts), produce plausible words (lemmas), increasing the likelihood of hits (i.e. generated forms being real words).



Lexical Elaboration

Task: Given short wordlists (no translations/texts), produce plausible words (lemmas), increasing the likelihood of hits (i.e. generated forms being real words).



Lexical Typology

Bininj Gun-Wok:

Worrowkme -- the hop of a person

Djarrwahme -- the hop of the female antilopine kangaroo

Mawudme -- the hop of the male antilopine kangaroo

Muribardme -- the hop of the female euro

Nqedjme -- the hop of the male walaroo



burrmarlarla

“The name for a male antelope kangaroo that is resting, lying on its side in part shade during the heat of the day and the movement of the dappled light on its fur gives it the appearance of having covered itself with white clay, just as hunters do when they hunt these large animals”

(an example from Rachel Nordlinger)



Lexical Elaboration with Dictionaries

Current work: Exploring lexical elaboration based on bilingual dictionaries!

Do Eskimo-Aleut languages have more words for “snow”?

Do languages of peoples living in desert areas have less words for “rain”?

Do Australian languages have rich vocabularies for “kangaroo”?

Do languages spoken in the rainforest environments (humid, hot) have more “smell” terms? How often the speakers are hunter-gatherers?

Which language has the richest “dance” vocabulary?

Lexical Elaboration with Dictionaries

Current work: Exploring lexical elaboration based on bilingual dictionaries!

Do Eskimo-Aleut languages have more words for “snow”?

Do languages of peoples living in desert areas have less words for “rain”?

Do Australian languages have rich vocabularies for “kangaroo”?

Do languages spoken in the rainforest environments (humid, hot) have more “smell” terms? How often the speakers are hunter-gatherers?

Which language has the richest “dance” vocabulary?

Ongoing work with Temuulen Khishigsuren, Charles Kemp, Terry Regier

Morphological Paradigm Completion

Collecting paradigm data efficiently

Support we'd like to elicit all form
for “fing” now.

Word generator(v1.2)

Abstraction factors
wildcard+ 70
Required confidence, % 70

N-Gram size
3 4 5 (selected)
Dictionary words
 Allow
Dictionary file Browse... en-words
 sample
< > Generate Download Close session

Generated Word Count (max. 10 syllables)
200 500 1000 2000 5000
Probability Score
 Display Sort by Pick top

Direct link to this variant settings

an	bar	be
because	bell	bellow
bellower	bir	bit
blo	blowe	blower
bon	breas	breat
breath	breather	breathere
bur	clo	cor
cou	coun	cound
countain	dir	dirt
drin	eart	fath
fathe	father	fe
fea	feart	fearth
feat	feath	feathe
feathere	fig	figh
fin	fig	finger
fir	fiver	flo

What are the options?

- 1) “fing” is a noun → (likely) regular + “-s” → “**fings**”
- 2) “fing” is a verb, i.e.
 - a) “I/We/You/They” Present → “**fing**”
 - b) “He/She” Present → “**fings**”
 - c) Continuous → “**finging**”
 - d) Past → “**finged**” or “**fang**”??

Morphological Paradigm Completion

Collecting paradigm data efficiently

Support we'd like to elicit all form
for “fing” now.

Word generator(v1.2)

Abstraction factors
wildcard+ 70
Required confidence, % 70

N-Gram size
3 4 5 (selected)
Dictionary words
 Allow
Dictionary file Browse... en-words
 sample
< > Generate Download Close session

wildcard+ Confidence dependent multipliers, % 100% Reset master
Generated Word Count (max. 10 syllables)
200 500 1000 2000 5000
Probability Score
 Display Sort by Pick top

Direct link to this variant settings
an bar be
because bell bellow
bellower bir bit
blo blowe blower
bon breas breat
breath breather breathere
bur clo cor
cou coun cound
countain dir dirt
drin eart fath
fathe fathere fe
fea feart fearth
feat feath feathe
feathere fig figh
fin fing fingern
fir fiver flo

Need to ask
the
speaker!

- 1) “fing” is a noun → (likely) regular + “-s” → “fings”
- 2) “fing” is a verb, i.e.
 - a) “I/We/You/They” Present → “fing”
 - b) “He/She” Present → “fings”
 - c) Continuous → “finging”
 - d) Past → “finged” or “fang”??

Morphological Paradigm Complexity

Two components of complexity (Ackerman and Malouf, 2013):

- **E-complexity:** the number of morphosyntactic distinctions a word overtly marks (approx. the paradigm size)
- **I-complexity** (form predictability/regularity)

Potentially, a tradeoff between the two:

- English verbs: small paradigms, irregulars in the past tense
- Archi: up to 1000 forms in a single paradigm, but high regularity

Morphological Paradigm Completion

A sample Paradigm for the Polish (fusional) lemma "książka" (book)

Syncretic forms

książka	książką	N;INS;SG
książka	książka	N;NOM;SG
książka	książkę	N;ACC;SG
książka	książki	N;ACC;PL
książka	?	N;GEN;SG
książka	książki	N;NOM;PL
książka	?	N;VOC;PL
książka	książkom	N;DAT;PL
książka	książko	N;VOC;SG

lemma

form

tag (features)

A sample Paradigm for the Turkish (agglutinative) lemma "guacamole"

Tag	guacamole
N;ACC;SG	?
N;ACC;PL	guacamoleleri
N;DAT;SG	guacamoleyeye
N;DAT;PL	?
N;ACC;PL;PSS3S	guacamolelerini
N;DAT;PL;PSS3S	guacamolelerine

Morphological Paradigm Completion

A sample Paradigm for the Polish (fusional) lemma "książka" (book)

Syncretic forms

książka	książką	N;INS;SG
książka	książka	N;NOM;SG
książka	książkę	N;ACC;SG
książka	książki	N;ACC;PL
książka	?	N;GEN;SG
książka	książki	N;NOM;PL
książka	?	N;VOC;PL
książka	książkom	N;DAT;PL
książka	książko	N;VOC;SG

lemma

form

tag (features)

A sample Paradigm for the Turkish (agglutinative) lemma "guacamole"

Tag	guacamole
N;ACC;SG	guakamole(y)i
N;ACC;PL	guakamoleleri
N;DAT;SG	guakamoleye
N;DAT;PL	guakamolelere
N;ACC;PL;PSS3S	guakamolelerini
N;DAT;PL;PSS3S	guakamolelerine

Morphological Paradigm Completion

A sample Paradigm for the Polish (fusional) lemma "książka" (book)

książka książką N;INS;SG

less regularity

higher syncretism

need to estimate classes

książka ? N;VOC;PL
książka książkom N;DAT;PL
książka książko N;VOC;SG

lemma

form

tag (features)

A sample Paradigm for the Turkish (agglutinative) lemma "guacamole"

Tag	guacamole
N;ACC;SG	guakamole(y)i
N;ACC;P	higher regularity
N;DAT;SG	need to learn morpheme order and their alignment with features
N;DAT;PL	
N;ACC;PL;PSS3S	guakamolelerini
N;DAT;PL;PSS3S	guakamolelerine

SIGMORPHON Shared Task on Morphological (Re-)Inflection. NNs generalise well!

Inflection: RUN + V;PST → ran

reinflection: running +V;PST → ran

Approx. 96% avg. accuracy on high-resource languages!

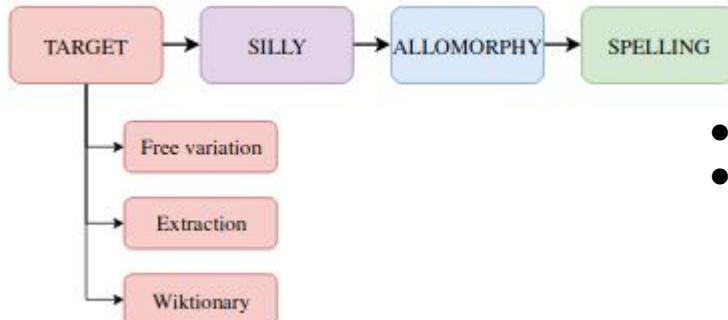
Significantly less in under-resourced languages!

Winning systems are neural seq2seq models

See more details in [my SIGTYP Talk](#)

Lemm		
RUN	V;PAST	ran
RUN	V;PRES;1;SG	run
RUN	V;PRES;2;SG	run
JN	V;PRES;3;SG	runs
JN	V;PRES;PL	run
RUN	V;PART	running

Error Taxonomy (Gorman et al., 2019)



- **Free variation error:** more than one acceptable form exists
- **Silly errors:** “bizarre” errors which defy any purely linguistic characterization (“*membled” instead of “mailed” or enters a loop such as “ynawemaylmyylmyylmyylmyylmyylmyym...” instead of “ysnewem”)
- **Allomorphy errors:** misapplication of existing allomorphic patterns
- **Spelling errors:** forms that do not follow language-specific orthographic conventions

Error Taxonomy (Gorman et al., 2019)

Language	Target	Silly		Allomorphy		Spelling	
		UE-LMU-I	CLUZH-7	UE-LMU-I	CLUZH-7	UE-LMU-I	CLUZH-7
Dutch	8	1	1	19	16	5	7
English	3	0	0	18	18	7	11
Finnish	11	7	7	33	48	0	0
German	3	4	10	54	67	9	9
Hungarian	83	21	9	37	44	1	0
Italian	5	5	1	11	16	0	2
Latin	119	2	0	76	93	0	0
Polish	5	6	3	60	67	2	4
Portuguese	1	1	0	6	7	1	2
Romanian	54	3	5	61	69	1	2
Russian	7	7	0	48	45	23	28
Spanish	7	2	1	12	12	6	6
Total						55	71

Majority of errors are due to allomorphy

Morphological Paradigm Completion



A simulation of linguistic fieldwork:

A linguist: a pre-existing dictionary (~lemmas are known) and a morphological paradigm structure

A native speaker: an oracle system provides access to complete paradigms for all lemmas: as an input it receives (1) a lemma, (2) target tags/features, (3) system id (linguist).

The oracle system has access to all forms, but it comes at a certain cost. Participants can send requests to retrieve a form or to check whether their prediction is correct (-1 for retrieval and incorrect forms). Their task is to reduce the penalty and get the highest accuracy at all forms prediction.

Morphological Paradigm Completion



Research Questions:

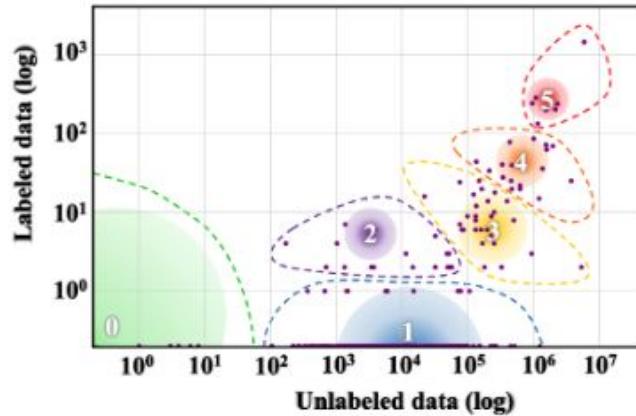
- What is the minimum number of samples to get the highest accuracy?
- What is the best strategy to sample selection?
- What are the most essential paradigm parts (principal parts of the paradigm)? How do we learn them automatically?
- How well do the systems learn syncretic forms?

NLP for Better Technology in Small Languages

Proposal #3: incorporate NLP models to digitise and translate pre-existing linguistic resources



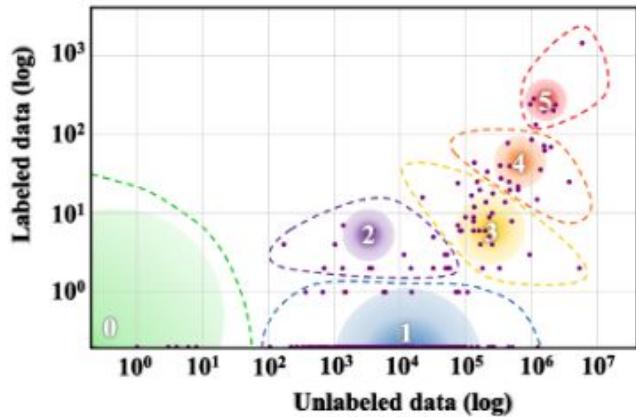
The State and Fate of Linguistic Diversity and Inclusion in the NLP World



- 0 - The Left-Behinds
- 1 - The Scraping-Bys
- 2 - The Hopefuls
- 3 - The Rising Stars
- 4 - The Underdogs
- 5 - The Winners

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

The Case of Mambai (Timor-Leste)

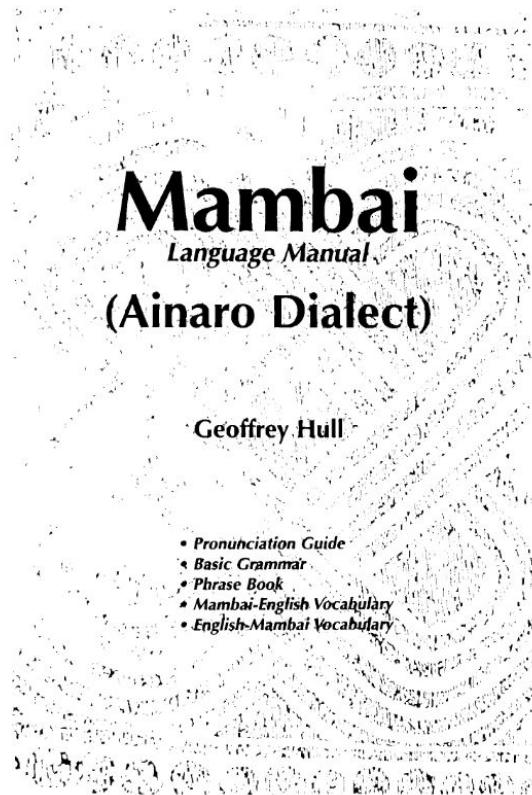


- Mambai (also spelled Mambae) is second most common mother tongue in Timor-Leste after Tetun, with around 200,000 native speakers
- An Austronesian language, does not have a standardised orthography
- **Translating to Mambai can bring valuable material closer to Mambai-speaking communities**

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Mambai belongs here!

The Case of Mambai (Timor-Leste)



Can we use large language models (LLMs) for translating English into Mambai?

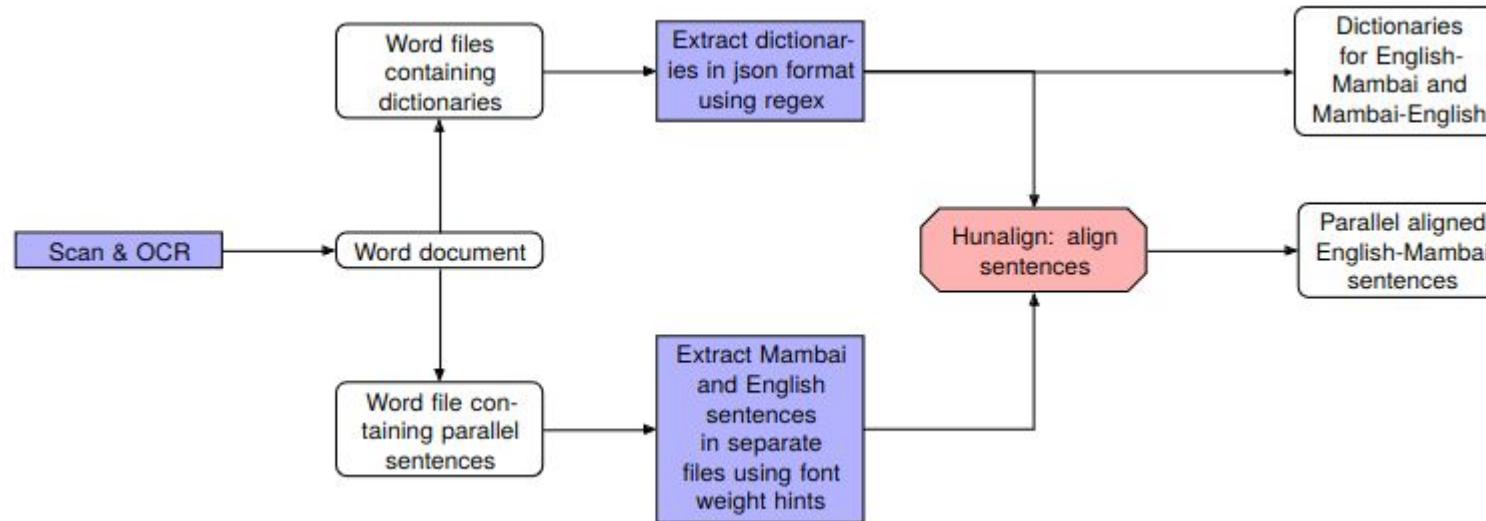
Initially provided:

A 109-pages book that includes a pronunciation guide, a grammar, a phrase book (1275 bitexts), and English-Mambai (1790 entries) and Mambai-English(1592)

LLMs: GPT-4-Turbo, Mixtral, Llama

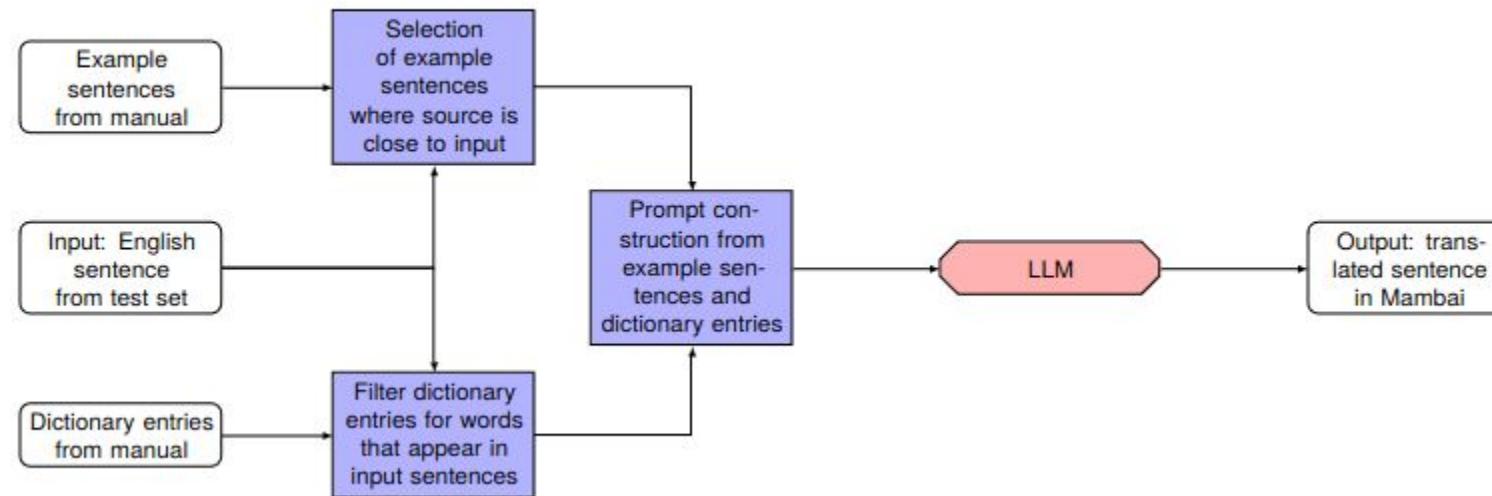
The Case of Mambai (Timor-Leste)

Preprocessing:



The Case of Mambai (Timor-Leste)

Translation:



The Case of Mambai (Timor-Leste)

Results:

Model	N_{TFIDF}	N_{embed}	UseDict	BLEU	ChrF	ChrF++
gpt-4-turbo	0	0	FALSE	3.7	22.4	19.9
gpt-4-turbo	0	0	TRUE	6.9	25.3	24.7
gpt-4-turbo	10	0	FALSE	16.1	40.3	39.7
gpt-4-turbo	10	0	TRUE	20.9	41.8	41.6
gpt-4-turbo	0	10	FALSE	16.8	38.2	37.4
gpt-4-turbo	0	10	TRUE	18.3	39.6	39.5
gpt-4-turbo	5	5	FALSE	17.7	40.4	39.6
gpt-4-turbo	5	5	TRUE	21.2	41.8	41.6
Mixtral 8x7B	5	5	TRUE	9.0	30.9	30.4
LlaMa 70b	5	5	TRUE	12.3	32.3	31.8

The Case of Mambai (Timor-Leste)

Results:

Model	N_{TFIDF}	N_{embed}	UseDict	BLEU	ChrF	ChrF++
gpt-4-turbo	0	0	FALSE	3.7	22.4	19.9
gpt-4-turbo	0	0	TRUE	6.9	25.3	24.7
gpt-4-turbo	10	0	FALSE	16.1	40.3	39.7
gpt-4-turbo	10	0	TRUE	20.9	41.8	41.6
gpt-4-turbo	0	10	FALSE	16.8	38.2	37.4
gpt-4-turbo	0	10	TRUE	18.3	39.6	39.5
gpt-4-turbo					39.6	
gpt-4-turbo					41.6	
Mixtral 8x7B					30.4	
LlaMa 70b					31.8	

BLEU over 20! Looks promising! But
does it generalise beyond the textbook?

The Case of Mambai (Timor-Leste)

Results:

Model	A	B	C	D	E	F
gpt-4-turbo	10	0	FALSE	16.1	40.3	39.7
gpt-4-turbo	10	0	TRUE	20.9	41.8	41.6
gpt-4-turbo	0	10	FALSE	16.8	38.2	37.4
gpt-4-turbo	0	10	TRUE	18.3	39.6	39.5
gpt-4-turbo						39.6
gpt-4-turbo						41.6
Mixtral 8x7B						30.4
LlaMa 70b						31.8

Let's ask a native speaker to provide us with a few more translations and then evaluate our system on them!

BLEU over 20! Looks promising! But does it generalise beyond the textbook?

The Case of Mambai (Timor-Leste)

Results:

Model	N _{TFIDF}	N _{embed}	UseDict	BLEU	ChrF	ChrF++
gpt-4-turbo	0	0	TRUE	3	30.7	27.9
gpt-4-turbo	0	0	FALSE	0	30.8	26.9
gpt-4-turbo	10	0	TRUE	4	36.9	33.8
gpt-4-turbo	10	0	FALSE	0	33.4	29.9
gpt-4-turbo	0	10	TRUE	3.4	34.5	31.6
gpt-4-turbo	0	10	FALSE	0	31.4	27.8
gpt-4-turbo	5	5	TRUE	4.4	35.9	33
gpt-4-turbo	5	5	FALSE	0	33.7	29.9
Mixtral 8x7B	5	5	TRUE	3.5	26.8	24.6
LlaMa 70b	5	5	TRUE	0	27.7	24.7

Such a drop of BLEU!

The Case of Mambai (Timor-Leste)

Results:

Model	N _{TFI}	N _{BLEU}	BLEU	TRUE	N _{BLEU}	N _{BLEU}	N _{BLEU}
gpt-4-turbo	0						
gpt-4-turbo	0						
gpt-4-turbo	10						
gpt-4-turbo	10						
gpt-4-turbo	0						
gpt-4-turbo	0						
gpt-4-turbo	5						
gpt-4-turbo	5						
Mixtral 8x7B	5	5	TRUE	3.5	26.8	24.6	
LlaMa 70b	5	5	TRUE	0	27.7	24.7	

Such a drop of BLEU!

Reasons? the native speaker had...

- 1) More idiosyncratic usage
- 2) Longer sentences
- 3) Language variation (orthography conventions, more Portuguese and Timor Dili)

The Case of Mambai (Timor-Leste)

Results:

Model	N_{TF}
gpt-4-turbo	0
gpt-4-turbo	0
gpt-4-turbo	10
gpt-4-turbo	10
gpt-4-turbo	0
gpt-4-turbo	0
gpt-4-turbo	5
gpt-4-turbo	5
Mixtral 8x7B	5
LlaMa 70b	7

Such a drop of BLEU!

Reasons? the native speaker had...

- 1) More idiosyncratic usage
- 2) Longer sentences
- 3) Language variation (orthography conventions, more Portuguese and Timor Dili)

Out-of-distribution generalisation is one of major topics in ML! Larger and more diverse data improve it!

TRUE 0.5 0.8 24.6
7 24.7

A Few more Experiments
with LLMs...

Multimodal Models for Digitisation

ábat₂ +V, —A D. +N. D:i--an.
To perform a ceremony for someone who has had a spirit encounter; such a ceremony. (1) lyabátam nan nasakit. Perform the ábat ceremony for the sick person.

aabátan₁ A place where the ábat ceremony may be performed. This is usually at one of the village entrances. Cf. aapoyan, papatáyan, wawalitan.

omaábat +N. A person able to perform the ábat ceremony.

paábat₁ +N. A person for whom the ábat ceremony is performed.

abbawil —A CON. A:in-, CON:i-. To support oneself by leaning on one arm against an object; to hang by one arm from an object. (1) Nan komáwat sinan báttang, in-ab-abbawil sinan panga. The person who climbed the pinetree hooked his arm around a branch. See: baw-ídang.

abbingi +N. +V, —P. P:na--an. A sixth finger, usually an extra thumb, occurring in the children of certain families. (1) Si Aggingi, naabbingiyan nan límana. Aggingi has six fingers on one hand. Cf. gammingi.

Claude-3 OPUS: “This is a part of a bilingual dictionary containing translations into English, the data is provided in two columns. Parse the dictionary, extract the words in the source language and provide their translations into English.”

RESPONSE

[Markdown](#) [Text](#)

The image shows a bilingual dictionary with translations from a language into English. Some of the translated words and phrases include:

ábat - To perform a ceremony for someone who has had a spirit encounter; such a ceremony.

lyabátam nan nasakit, Perform the ábat ceremony for the sick person.

aabátan - A place where the ábat ceremony may be performed. This is usually at one of the village entrances.

omaábat - A person able to perform the ábat ceremony.

pábat - A person for whom the ábat ceremony is performed.

abbawil - A CON. To support oneself by leaning on one arm against an object; to hang by one arm from an object.

abbingi - A sixth finger, usually an extra thumb, occurring in the children of certain families.

Sí Agálngi, naabbingiyan nan Ifmana. Aggingi has six fingers on one hand.

The dictionary does not identify or name any individuals. It focuses on defining and translating words and phrases from one language to English.

Multimodal Models for Digitisation

ábat₂ +V, —A D. +N. D:i--an.
To perform a ceremony for someone who has had a spirit encounter; such a ceremony. (1) lyabátam nan nasakit. Perform the ábat ceremony for the sick person.

aabátan₁ A place where the ábat ceremony may be performed. This is usually at one of the village entrances. Cf. aapoyan, papatáyan, wawalitan.

omaábat +N. A person able to perform the ábat ceremony.

paábát_L +N. A person for whom the ábat ceremony is performed.

abbawil —A CON. A:in-, CON:i-. To support oneself by leaning on one arm against an object; to hang by one arm from an object. (1) Nan komáwat sinan báttang, in-ab-abbawil sinan panga. The person who climbed the pinetree hooked his arm around a branch. See: baw-ídang.

abbingi +N. +V, —P. P:na--an. A sixth finger, usually an extra thumb, occurring in the children of certain families. (1) Si Aggingi, naabbingiyan nan límana. Aggingi has six fingers on one hand. Cf. gammingi.

Claude-3 OPUS: “This is a part of a bilingual dictionary containing translations into English, the data is provided in two columns. Parse the dictionary, extract the words in the source language and provide their translations into English.”

RESPONSE

[Markdown](#) [Text](#)

The image shows a bilingual dictionary with translations from a language into English. Some of the translated words and phrases include:

ábat - To perform a ceremony for someone who has had a spirit encounter; such a ceremony.

lyabátam nan nasakit, Perform the ábat ceremony for the sick person.

aabátan - A place where the ábat ceremony may be performed. This is usually at one of the village entrances.

omaábat - A person able to perform the ábat ceremony.

pábat - A person for whom the ábat ceremony is performed.

abbawil - A CON. To support oneself by leaning on one arm against an object; to hang by one arm from an object.

abbingi - A sixth finger, usually an extra thumb, occurring in the children of certain families.

Sí Agálngi, naabbingiyan nan Ifmana. Aggingi has six fingers on one hand.

The dictionary does not identify or name any individuals. It focuses on defining and translating words and phrases from one language to English.

Is able to perform OCR (a few errors) and parse the dictionary structure!

Multimodal Models for Digitisation

ábat₂ +V, —A D. +N. D:i--an.
To perform a ceremony for someone who has had a spirit encounter; such a ceremony. (1) lyabátam nan nasakit. Perform the ábat ceremony for the sick person.

aabátan₁ A place where the ábat ceremony may be performed. This is usually at one of the village entrances. Cf. aapoyan, papatáyan, wawalitan.

omaábat +N. A person able to perform the ábat ceremony.

paábat₁ +N. A person for whom the ábat ceremony is performed.

abbawil —A CON. A:in-, CON:i-. To support oneself by leaning on one arm against an object; to hang by one arm from an object. (1) Nan komáwat sinan báttang, in-ab-abbawil sinan panga. The person who climbed the pinetree hooked his arm around a branch. See: baw-ídang.

abbingi +N. +V, —P. P:na--an. A sixth finger, usually an extra thumb, occurring in the children of certain families. (1) Si Aggingi, naabbingiyan nan límana. Aggingi has six fingers on one hand. Cf. gammingi.

Claude-3 OPUS: “This is a part of a bilingual dictionary containing translations into English, the data is provided in two columns. Parse the dictionary, extract the words in the source language and provide their translations into English.”

RESPONSE

[Markdown](#) [Text](#)

The image shows a bilingual dictionary with translations from a language into English. Some of the translated words and phrases include:

ábat - To perform a ceremony for someone who has had a spirit encounter; such a ceremony.

lyabátam nan nasakit, Perform the ábat ceremony for the sick person.

aabátan - A place where the ábat ceremony may be performed. This is usually at one of the village entrances.

omaábat - A person able to perform the ábat ceremony.

pábat - A person for whom the ábat ceremony is performed.

abbawil - A CON. To support oneself by leaning on one arm against an object; to hang by one arm from an object.

abbingi - A sixth finger, usually an extra thumb, occurring in the children of certain families.

Sí Agálngi, naabbingiyan nan Ifmana. Aggingi has six fingers on one hand.

The dictionary does not identify or name any individuals. It focuses on defining and translating words and phrases from one language to English.

Is able to perform OCR (a few errors) and parse the dictionary structure!

Can we do Translation?

Multimodal Models with Digitisation

аңј- A широкий, широко <V-A181>

аңјү нареч широко <V-A181>

аңјлах прил 1) плоский <VK15> 2) широкий (по отношению к плоским предметам: стена, доска, скатерть, покрывало, стол и т.п.) <V-A181>

аңјү нареч широко <V-A181>

аңјлах прил 1) плоский <VK15> 2) широкий (по отношению к плоским предметам: стена, доска, скатерть, покрывало, стол и т.п.) <V-A181>;
Тюш антълях широкий

аңойк'а (мн. аңойк'а?н) сущ бесхвостый (всякое животное) <V-A182>

аофү сущ блюдо из запеченной толченой свежей лососевой икры с клубнями сараны или картофеля <СТРК>

апјк'алах var. апәдјлах прил плоский, мелкий (напр. блюдце) <V-A185>;
Тюш апльках;
см. (б) даплкало

Itelmen – Russian →? English

Claude-3 OPUS: “This is a part of

Itelmen-Russian dictionary containing translations into Russian. Parse the dictionary, extract the word in Itelmen language and provide their translations into English.”

RESPONSE

Markdown Text

This image contains a portion of an Itelmen-Russian dictionary with Itelmen words on the left and their Russian translations on the right. Here are the Itelmen words and their English translations:

аңј-л - wide, broad

аңїн - to widen, expand

аңїнах - 1) flat, 2) wide (in relation to flat objects: wall, board, cloth, cover, etc.)

аңїн - wide

аңјлах - 2) wide (in relation to flat objects: wall, board, cloth, cover, etc.)

аңїлях - flat, wide

аңёнк'а - all, whole (about a beastly living creature)

аоф - to block the way, to block the fresh air with a club or cartilage

аңїк'алах - wide, spacious

аңїк'алах - flat, wide

аңїках - to widen

Many Itelmen diacritics/special characters are wrong.
Some translations are very misleading (hallucinations).

Multimodal Models with Digitisation

ајј- A широкий, широко <V-A181>

ајјю нареч широко <V-A181>

ајјлах прил 1) плоский <VK15> 2) широкий (по отношению к плоским предметам: стена, доска, скатерть, покрывало, стол и т.п.) <V-A181>

ајјю нареч широко <V-A181>

ајјлах прил 1) плоский <VK15> 2) широкий (по отношению к плоским предметам: стена, доска, скатерть, стол и т.п.)

Тюш антълях широкий

аңойк'а (мн. аңойк'а?н) сущ бесхвостый

аофк сущ блюдо из запеченной толченой

клубнями сараны или картофеля <ст>

аплк'алах var. апәлјлах прил плоский, мелкий (напр. блюдце) <V-A185>;

Тюш аплъках;

см. (б) даплкало

Itelmen – Russian →? English

Gemini: “This is a part of Itelmen-Russian dictionary containing translations into Russian. Parse the dictionary, extract the word in Itelmen

The translation part looks promising but needs to be checked by a native (bilingual) speaker, or at least other systems for consistency!

into English, based on the image:

- аул - wide, widely
- анлк - widely
- андлах - flat, wide (in relation to flat objects: wall, board, tablecloth, bedspread, table, etc.)
- аңойк'а (тп. аңойк'а?н) - tailless (any animal)
- аофк - a dish of baked mashed fresh salmon roe with sarana tubers or potatoes
- аплк'алах or апәлјлах - flat, shallow (e.g., saucer)

Many Itelmen diacritics/special characters are wrong.
But translations are better.

Issues:

- 1) privacy/copyright: unlike OCR systems that can run 100% offline (client-side), these models require data uploading and processing on their servers
- 2) No clear option of PDF processing (single images)
- 3) Hallucinations
- 4) The performance across diverse scripts is not well-studied

Overall, looks promising, and I see many opportunities to improve digitisation of linguistic resources, especially with linguists and community members involved!

THANK YOU!

Questions?

