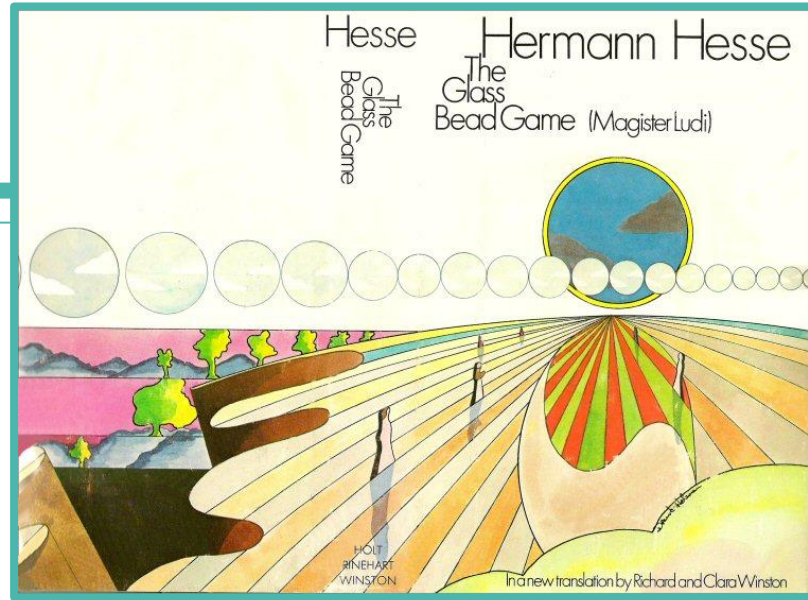
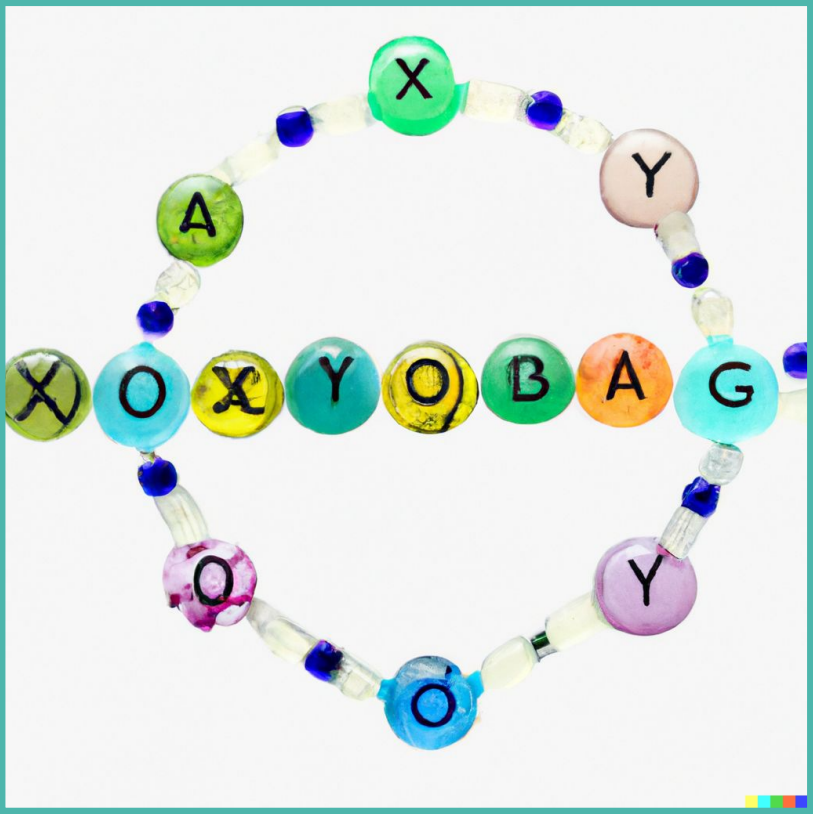


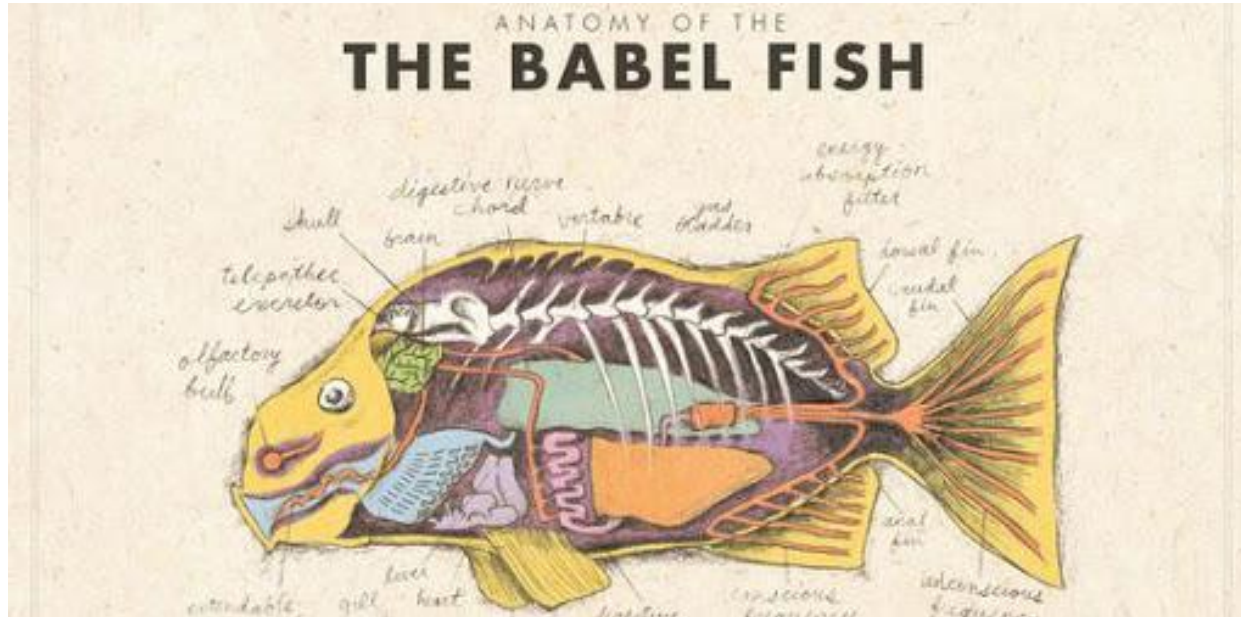
# A Glass Bead Game of

\*  
.  
\*  
\*  
o  
l  
o  
g  
y



Contemporary Computational  
Approaches to **Linguistic Morphology**,  
**Typology**, and  
**Social Psychology**

# Multilinguality



# Approx. 7,000\* languages in the world...



\* Language vs. dialect distinction: “A language is a dialect with its own army and navy” (Max Weinreich)

## Roman Jakobson on differences

## between Languages



“Languages differ essentially in what they must convey and not in what they may convey”

# Chinese (Isolating; strict word order)

wǒmen xué le zhè xiē shēngcí.

I.PL.AN learn .PAST this .PL new word.

``We learned these new words.''

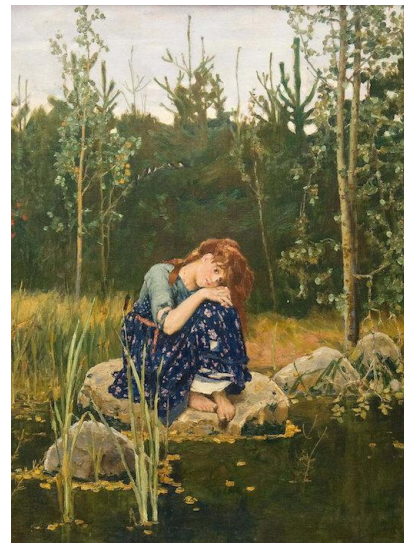


# Russian (Synthetic;flexible word order)

My vyučili eti novyje slova.

We learn.PAST.PL this.ACC.PL new.ACC.PL word.ACC.PL

``We learned these new words.''



# West Greenlandic (Polysynthetic; Fortescue (2017))

Nannu-n-niuti-kkuminar-tu-rujussu-u-vuq.

Polar.bear-catch-instrument.for.achieving-something.good.for-PART-  
big-be-3SG.INDIC

“It (a dog) is good for catching polar bears  
with.”

Speedtalk in “Gulf” by Robert Heinlein: complex syntax, minimal  
vocabulary, and a rich phoneme inventory

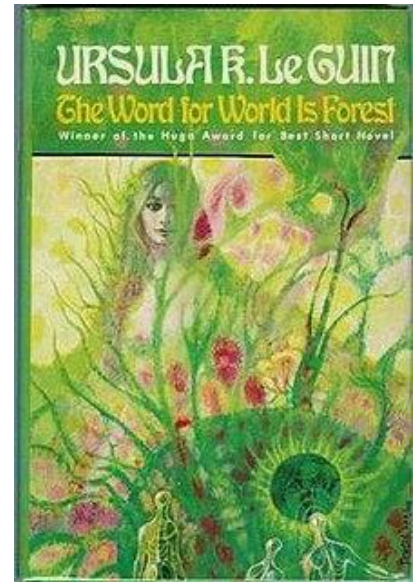


# Kunwinjku (Polysynthetic; Evans (2003))

Aban-yawoith-warrgah-marne-ganj-ginje-ng.

1/3PL-again-wrong-BEN-meat-cook-PP

"I cooked the wrong meat for them again"





# NLP: Universal or Euroversal?

# Approx. 7,000\* languages in the world...

... But the vast majority of NLP technologies only focus on most documented languages such as English, French, German, Russian, Hindi, or Finnish

\* Language vs. dialect distinction: “A language is a dialect with its own army and navy” (Max Weinreich)

# Most NLP is Standard Average European

*SAE was introduced by Whorf, 1939*

... According to Martin Haspelmath (2001), “euroversals” share:

- definite and indefinite articles (e.g. English the vs. a)
- a periphrastic perfect formed with 'have' plus a passive participle (e.g. English I have said);
- the verb is inflected for person and number of the subject, but subject pronouns may not be dropped even when this would be unambiguous
- Some features that are common in European langs but also found elsewhere:
  - lack of distinction between inclusive and exclusive first-person plural pronouns ("we and you" vs. "we and not you")
  - lack of distinction between alienable and inalienable (e.g. body part) possession;

Approx. 7,000\* languages in the world, a variety of...

Descriptive traditions

Sound Systems

Writing Systems (many do not have any)



Semantics

Morphology

Syntax



# Approx. 7,000\* languages in the world, a variety of...

How do we address the differences and unify the linguistic data?

Descriptive traditions

Sound Systems

Writing Systems (many do not have any)



Semantics

Morphology

Syntax

Which features are language specific and which are universal?

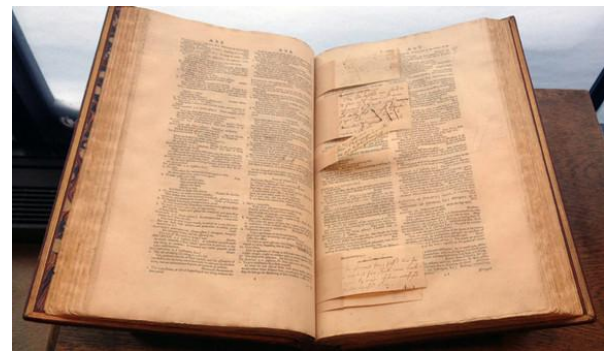
**Linguistic typology** is the discipline that studies variations by means of a systematic comparison of languages (Croft, 2002; Comrie, 1989)



# Standardizing the Linguistic Data

- **WALS** (<https://wals.info/>) for linguistic structures
- **Glottolog** (Hammarström et al., 2021) for language names and codes
- **Concepticon** (<https://concepticon.cld.org>, List et al. 2022c) for unified concept senses
- **CLICS** (<https://clics.cld.org/>; Rzymiski, Tresoldi et al. 2019): cross-linguistic co-lexifications (polysemies)
- **B(road)IPA system** of the Cross-Linguistic Transcription Systems (<https://clts.cld.org>, Anderson et al. 2018)
- **PARSEME**: multi-word expressions
- **UD** (<https://universaldependencies.org/>): cross-linguistic syntactic annotations
- **UniMorph** (<https://unimorph.github.io>): Universal Morphosyntactic Annotation Schema



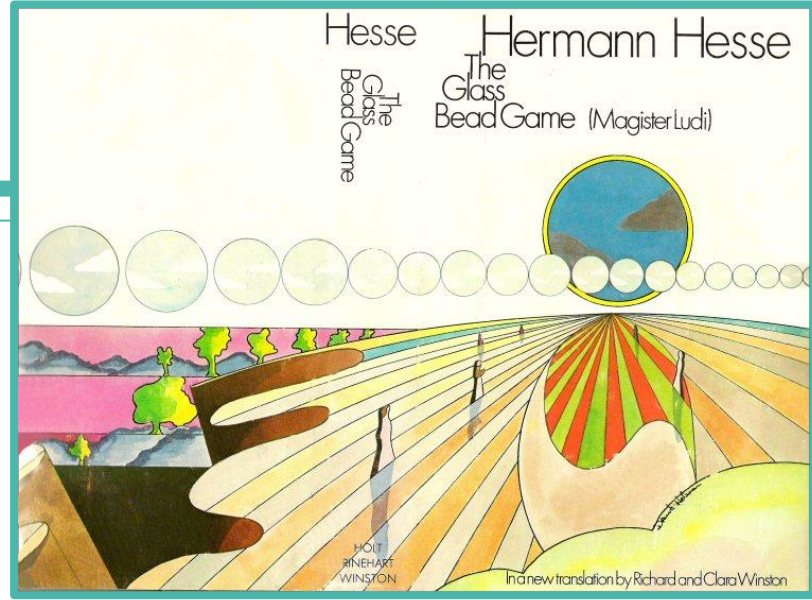
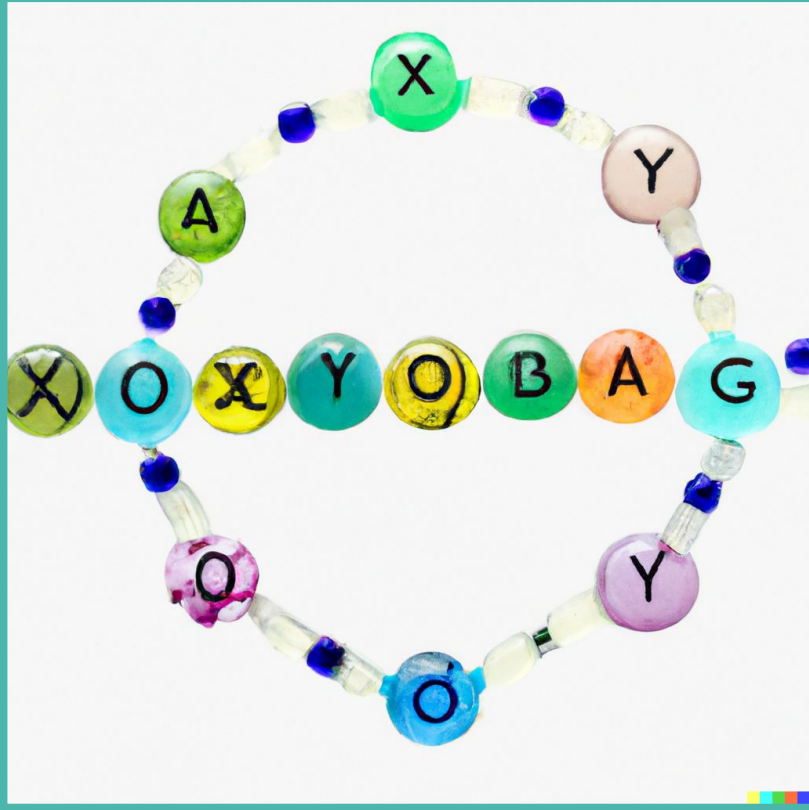


# Standardizing the Linguistic Data

- **WALS** (<https://wals.info/>) for linguistic structures
  - **Glottolog** (Hammarström et al., 2021) for language names and codes
  - **Concepticon** (<https://concepticon.clld.org>, List et al. 2022c) for unified concept senses
  - **CLICS** (<https://clics.clld.org/>; Rzymiski, Tresoldi et al. 2019): cross-linguistic co-lexifications (polysemies)
  - **B(road)IPA system** of the Cross-Linguistic Transcription Systems (<https://clts.clld.org>, Anderson et al. 2018)
  - **PARSEME**: multi-word expressions
  - **UD** (<https://universaldependencies.org/>): cross-linguistic syntactic annotations
  - **UniMorph** (<https://unimorph.github.io>): Universal Morphosyntactic Annotation Schema
- We will focus on this project!

# A Glass Bead Game of

\*.  
\*  
\*  
\*  
o  
l  
o  
g  
y

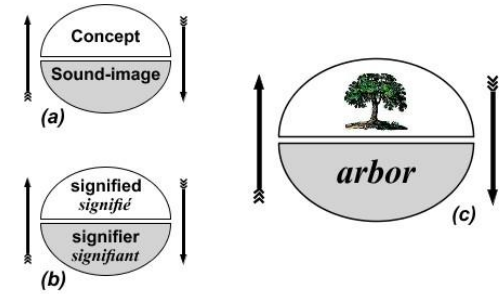


Contemporary Computational  
Approaches to **Linguistic Morphology**,  
**Typology**, and  
**Social Psychology**

**Linguistic morphology** is the study of words, how they are formed, and their relationship to other words in the same language

# Language

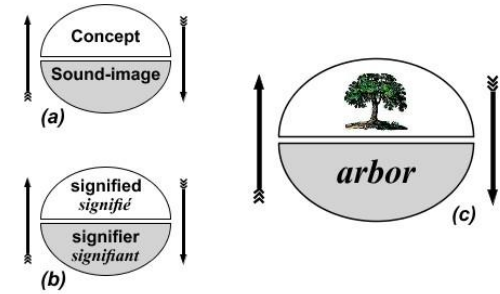
**Arbitrariness of the sign (Saussure; 1916):** the form, **signifier**, has little relationship to the meaning, **signified** (excp. onomatopoeia, e.g. “meow”, and phonesthemes, e.g “gl-” for light-related concepts)



**Efficiency of Communication:** 1) the speaker is focusing on ensuring the listener fully understanding the message he is trying to deliver; 2) deliver the message in a shorter time period (e.g., Zipf's law of abbreviation)

# Language

**Arbitrariness of the sign (Saussure; 1916):** the form, **signifier**, has little relationship to the meaning, **signified** (excp. onomatopoeia, e.g. “meow”, and phonesthemes, e.g. “gl-” for light-related concepts)



**Efficiency of Communication:** 1) the speaker is focusing on ensuring the listener fully understanding the message he is trying to deliver; 2) deliver the message in a shorter time period (e.g., Zipf's law of abbreviation)

→ **Regularity and Systematicity in Language**

# Language: Jabberwocky sentences

From Lewis Carroll:

*“Twas bryllyg, and y<sup>e</sup> slythy toves  
Did gyre and gymble in y<sup>e</sup> wabe:  
All mimsy were y<sup>e</sup> borogoves;  
And y<sup>e</sup> mome raths outgrabe. {...}“\**

Alice, after finishing the poem, *“Somehow it seems to fill my head with ideas—only I don’t exactly know what they are! However, somebody killed something: that’s clear, at any rate.”*

Where does the meaning come from here?



The Jabberwock by John Tenniel, 1871

\* is considered one of the greatest nonsense poems written in English

# Language: Jabberwocky sentences

“The glocky kouzdress steckly brutted the bock and is cuddering the bocklet”

# Language: JabberWocky sentences

“The glocky kouzdress steckly brutted the bock and is cuddering the bocklet”

DALL-E:





# Language: JabberWocky sentences

“The glocky kouzdress steckly brutted the bock and is cuddering the bocklet”

ChatGPT (4.0):



Explain the meaning of the following sentence: "The glocky kouzdress steckly brutted the bock and is cuddering the bocklet"



I'm sorry, but the sentence you provided does not make sense and appears to be a string of nonsensical words. It does not convey any meaning or logical structure in English language or any other language that I am aware of. Please provide a valid sentence or question that I can help you with.

DALL-E:



# Language: JabberWocky sentences

“The glocky kouzdress steckly bruted the bock and is cuddering the bocklet”

ChatGPT (4.0). Another try:



Guess the meaning of the sentence "The glocky kouzdress steckly bruted the bock and is cuddering the bocklet"



The sentence "The glocky kouzdress steckly bruted the bock and is cuddering the bocklet" does not make sense in standard English, as it contains several nonce words or made-up words that are not found in any dictionary. Therefore, it is difficult to accurately guess the intended meaning of the sentence.

However, if we were to assume that the words have some sort of consistent meaning within a certain context or domain, one possible interpretation of the sentence could be:

"The strange dress quickly smashed the block and is now gently holding the small block."

However, it is important to note that this is merely a speculative interpretation based on a loose understanding of the individual words and their potential meanings, and it is not clear whether this interpretation is accurate or even close to the intended meaning.

DALL-E:



# Language: JabberWocky sentences

“The glocky kouzdress steckly bruted the bock and is cuddering the bocklet”

ChatGPT (4.0). POS Tagging:

can you do POS tagging of the sentence?

Yes, I can certainly perform POS tagging on the sentence you provided, even though it does not make any sense. However, since the sentence does not follow any syntactical rules or contain any recognizable English words, the POS tagging would simply be a string of meaningless tags. Here is an example of what the POS tagging of the sentence might look like:

"The/DT glocky/NN kouzdress/NN steckly/RB bruted/VBD the/DT bock/NN and/CC is/VBZ cuddering/VBG the/DT bocklet/NN"

As you can see, each word in the sentence has been assigned a part of speech tag based on its function within the sentence, but these tags do not have any significance or meaning in the absence of a coherent sentence structure and recognizable words.

Well Done!

what about dependency parsing?

Similarly to POS tagging, dependency parsing involves analyzing the grammatical structure of a sentence to identify the relationships between its words. However, since the sentence you provided does not make sense and contains a series of nonsensical words, it would not be possible to perform a meaningful dependency parsing on it.

If you provide me with a valid sentence or a piece of text, I would be happy to demonstrate how dependency parsing works and provide you with a sample dependency parse tree.


DALL-E:





# Language: JabberWocky sentences

“The glocky kouzdress steckly bruted the bock and is cuddering the bocklet”

ChatGPT (4.0). Translation:

 Translate this sentence into Russian: ""The glocky kouzdress steckly bruted the bock and is cuddering the bocklet""

 "Глокий куздресс ударил блок и ласково держит маленький блоклет" 

This is a direct translation of the sentence ""The glocky kouzdress steckly bruted the bock and is cuddering the bocklet"" into Russian. However, as mentioned before, the sentence contains several nonce words or made-up words that do not exist in the English language, and therefore, the translation is likely to be nonsensical or inaccurate.

DALL-E:



“Glocky kouzdress.MASC struck the block.INAN and is gently holding a little blocklet.INAN.”

Resembles its own explanation in English

# Language: JabberWocky sentences

The original utterance in RU (by Lev Shcherba):

Morphology and word ordering provide a significant amount of information!

Glokaja                      kouzdra                      šteko    budlanula  
Glocky.F.SG.NOM    koudress.F.SG.NOM    steckly    brut.PAST.3SG  
bokra                      i    kurdjačit                      bokrënka.  
bock.M.SG.ACC    and    cudder.PRES.3SG    bockling.M.SG.ACC

“The glocky koudress steckly brutted the bock and is cuddering the bocklet”

In Arabic: بضلت قزدرة اغلوكة البكر بإشتيك وتتقرض البكر الصغير

Now translated into 90 Languages: <https://github.com/ivri/kuzdra/> !

# Language: JabberWocky sentences

The original utterance in RU (by Lev Shcherba):

Morphology and word ordering provide a significant amount of information!

Glokaja                      kouzdra                      šteko      budlanula  
Glocky.F.SG.NOM    koudress.F.SG.NOM    steckly    brut.PAST.3SG  
bokra                      i      kurdjačit                      bokrënka.  
bock.M.SG.ACC    and    cudder.PRES.3SG    bockling.M.SG.ACC

“The glocky koudress steckly brutted the bock and is cuddering the bocket!”

In Arabic: بضلت قزدرة اغلوكة البكر بإشتيك وتتقرض البكر الصغير

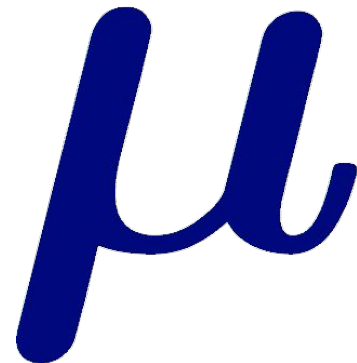
Now translated into 90 Languages: <https://github.com/ivri/kuzdra/>!

Languages differ in the information speakers express in words  
Annotation depends on descriptive traditions

# UniMorph: Universal Morphosyntactic Annotation

# UniMorph (Sylak-Glassman, 2016)

<https://unimorph.github.io/>



- 1) 23 Dimensions of meaning (TAM, case, number) with 212 features
- 2) A-morphous (word-based) morphology (Anderson, 1992)
- 3) Initial paradigms were extracted from the English Edition of Wiktionary (Kirov et al., 2016)
- 4) Follows the Leipzig Glossing Rules (Comrie et al., 2008)

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

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

lemma

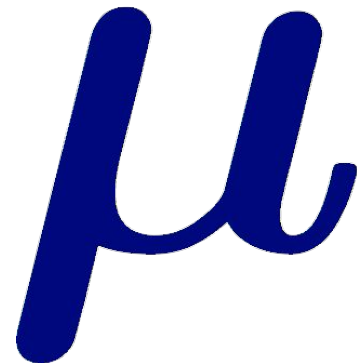
form

tag (features)



# UniMorph (Sylak-Glassman, 2016)

<https://unimorph.github.io/>



- 1) 23 Dimensions of meaning (TAM, case, number) with 212 features
- 2) A-morphous (word-based) morphology (Anderson, 1992)
- 3) Initial paradigms were extracted from the English Edition of Wiktionary (Kirov et al., 2016)
- 4) Follows the Leipzig Glossing Rules (Comrie et al., 2008)

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

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

What's the difference from Wiktionary?

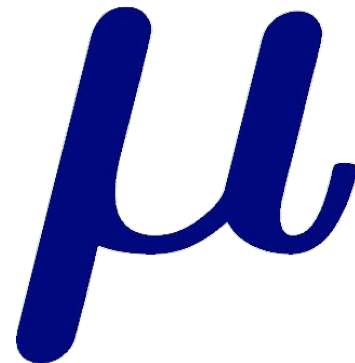
lemma

form

tag (features)

# UniMorph (Sylak-Glassman, 2016)

<https://unimorph.github.io/>



- 1) 23 Dimensions of meaning (TAM, case, number) with 212 features
- 2) A-morphous (word-based) morphology (Anderson, 1992)
- 3) Initial paradigms were extracted from the English Edition of Wiktionary (Kirov et al., 2016)
- 4) Follows the Leipzig Glossing Rules (Comrie et al., 2008)

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

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

What's the difference from Wiktionary?

- + Cross-linguistic annotation
- + More extremely under-resourced languages
- Unattested forms

lemma

form

tag (features)

# **SIGMORPHON Shared Task on Morphological Reinflection**

# SIGMORPHON Shared Task on Morphological (Re-)Inflection

Inflection: RUN + V;PST → ? (predict: “ran”)  
Reinflection: running +V;PST → ? (predict: “ran”)

*Cotterell et al., 2016–2018*

*McCarthy et al., 2019*

*Vylomova et al., 2020*

*Pimentel, Ryskina et al, 2021*

*Kodner et al., 2022*

## Goals:

- Collecting and cleaning the new data for UniMorph
- Evaluation of ML systems

Lemma	Tag	Form
RUN	V;PAST	ran
RUN	V;PRES;1;SG	?
RUN	V;PRES;2;SG	run
RUN	V;PRES;3;SG	?
RUN	V;PRES;PL	run
RUN	V;PART	running

# SIGMORPHON Shared Task on Morphological (Re-)Inflection

Inflection: RUN + V;PST → ? (predict: “ran”)

Reinflection: running +V;PST → ? (predict: “ran”)

*Cotterell et al., 2016–2018*

*McCarthy et al., 2019*

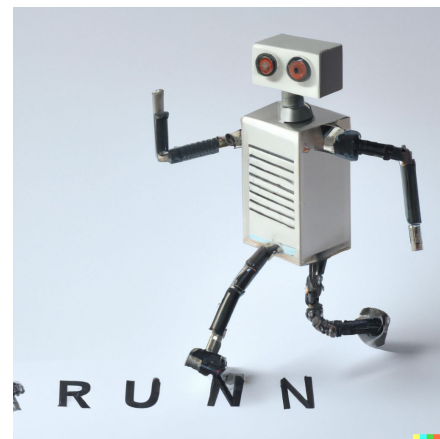
*Vylomova et al., 2020*

*Pimentel, Ryskina et al, 2021*

*Kodner et al., 2022*

A robot writing  
the past tense of  
the verb “run” :-)

© DALL-E



Lemma	Tag	Form
RUN	V;PAST	ran
RUN	V;PRES;1;SG	?
RUN	V;PRES;2;SG	run
RUN	V;PRES;3;SG	?
RUN	V;PRES;PL	run
RUN	V;PART	running

# SIGMORPHON Shared Task on Morphological (Re-)Inflection

Inflection: RUN + V;PST → ? (predict: “ran”)  
Reinflection: running +V;PST → ? (predict: “ran”)

*Cotterell et al., 2016–2018*

*McCarthy et al., 2019*

*Vylomova et al., 2020*

*Pimentel, Ryskina et al, 2021*

*Kodner et al., 2022*

Systems:

2016: FST- and rule-based. The winning system was neural (RNN-based)

2017– now: mainly neural systems

Lemma	Tag	Form
RUN	V;PAST	ran
RUN	V;PRES;1;SG	run
RUN	V;PRES;2;SG	run
RUN	V;PRES;3;SG	runs
RUN	V;PRES;PL	run
RUN	V;PART	running

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

Inflection: RUN + V;PST → ran  
reinflection: running +V;PST → ran

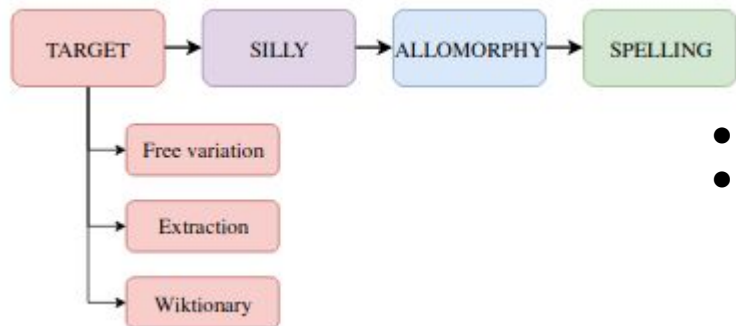
Approx. 96% avg. accuracy on high-resource languages!

Significantly less in under-resourced languages!

Winning systems are neural seq2seq models  
See more details in [my SIGTYP Talk](#)

Lem		
RUN	V;PAST	ran
RUN	V;PRES;1;SG	run
RUN	V;PRES;2;SG	run
RUN	V;PRES;3;SG	runs
RUN	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**

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

## SoTA: char-level Transformer

Best training strategies for language with limited data:

- **data hallucination** (extract common lemma→form transformations and produce more samples with nonce stem)
- **data augmentation** (generate samples from non-lemma form combinations, e.g. use V;PRES;1;SG instead of lemma form)
- **multilingual training**
- **ensembles**
- **replace softmax with sparsemax** (allows 0 probs; reducing the search space)

Lemma	Tag	Form
RUN	V;PAST	ran
RUN	V;PRES;1;SG	run
RUN	V;PRES;2;SG	run
RUN	V;PRES;3;SG	runs
RUN	V;PRES;PL	run
RUN	V;PART	running

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

Still, some languages/data scenarios are more challenging for the systems:

- Tungusic (**large paradigms, very sparse data, not standardized**)
- Oto-Manguean (**complex tonal morphology**)
- Southern Daly (Murrinhpatha: **polysynthetic; 34 classes of verbs**)
- Gunwinyguan (Kunwinjku: **fails in orthography: \*ngurriborlbme** instead of ngurriborle; **looping effects in RNNs** as in \*ngar-rrrrrrmbbbijj instead of karribelbmerrinj)

Lemma	Tag	Form
RUN	V;PAST	ran
RUN	V;PRES;1;SG	run
RUN	V;PRES;2;SG	run
RUN	V;PRES;3;SG	runs
RUN	V;PRES;PL	run
RUN	V;PART	running

# SIGMORPHON Shared Task on Morphological (Re-)Inflection. NNs generalize well!...Or Not?

## SIGMORPHON Shared Task 2021:

– split by lemmas and feature sets (test sets contain either unseen lemmas or feature combinations, or both)

**Prediction:** systems should be able to make correct predictions, especially in agglutinative languages

**Results:** the prediction of inflections with **unseen features proved challenging (even for agglutinative languages)**, with average performance decreased substantially from last year.

## Sample data for Turkish (agglutinative)

Tag	guakamole
N;ACC;SG	?
N;ACC;PL	guakamoleleri
N;DAT;SG	guakamoleye
N;DAT;PL	?
N;ACC;PL;PSS3S	guakamolelerini
N;DAT;PL;PSS3S	guakamolelerine

**SIGMORPHON Shared Task on Morphological  
(Re-)Inflection. Do systems generalize like humans?**

# “The Past Tense” Debate of the 1980s

Rumelhart and McClelland (1986) “On Learning of Past Tenses of English Verbs” (will NNs learn to distinguish English regulars from irregulars?):

Training\*:

run + Past → ran

Guess + Past → guessed

Test:

overrun+Past → ?

sleep+Past → ?

\* the authors provided phonological representations augmented with features

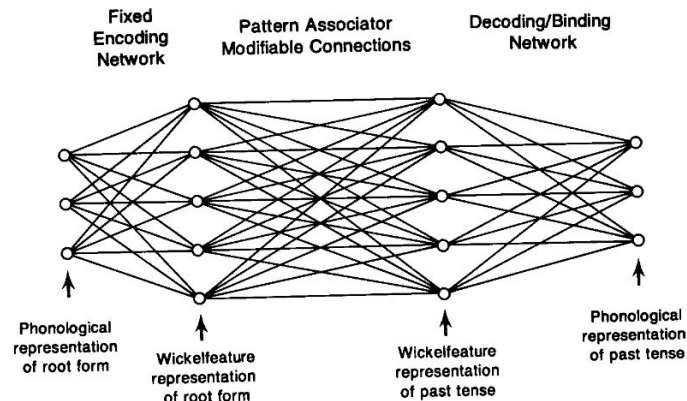


FIGURE 1. The basic structure of the model.

# “The Past Tense” Debate of the 1980s

Pinker and Prince (1988) “On language and connectionism:  
Analysis of a parallel distributed processing model of language  
acquisition”



- (1) it cannot represent certain words
- (2) it cannot learn many rules
- (3) it can learn rules found in no human language
- (4) it cannot explain morphological and phonological regularities
- (5) it cannot explain the differences between irregular and regular forms
- (6) it fails at its assigned task of mastering the past tense of English
- (7) failed at overregularization of irregular forms (e.g., “ated”)

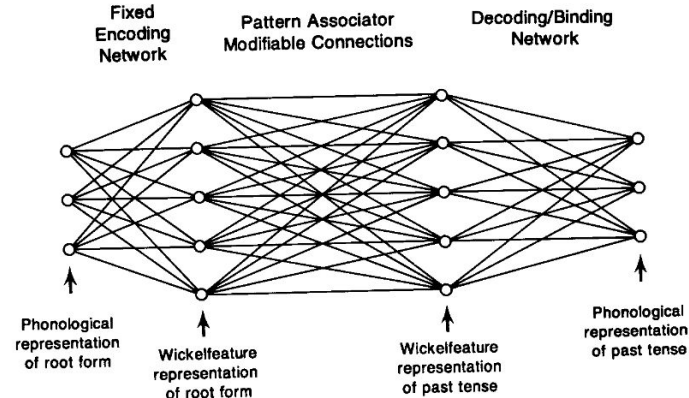


FIGURE 1. The basic structure of the model.

# SIGMORPHON Shared Task on Morphological (Re-)Inflection. Do systems generalize like humans?

## SIGMORPHON Shared Task 2021. Part 2: Model children's developmental trajectories

- English past tense
- German noun plurals
- Arabic noun plurals

Results (all submitted systems were neural such as char-level Transformer):

- American English past tense forms proved the easiest, Modern Standard Arabic noun plural the most challenging
  - systems learned regular patterns from smaller number of samples
  - far fewer nonsense forms of the “mail-membled” type
  - successfully learned over-regularization (e.g., -ed for the English Past Tense)
  - persistent in-human over-irregularization



# SIGMORPHON Shared Task on Morphological (Re-)Inflection 2023!

## 1) Multilingual Inflection

- lemma-split data
- new nesting feature structure
- data in Japanese kanji with loads of unseen chars in the test set
- 18 languages are available for development, surprise languages are coming soon!

## (2) Cognitively Plausible Morphophon. Generalization in Korean

- training data drawn from infant- and adult-directed speech corpora, test data from a linguistic generalization task (wug tests)
- designed to probe how speakers represent existing and novel words

## (3) Models of Acquisition of Inflectional Noun Morphology in Polish, Finnish, Estonian

- training data drawn from child-directed speech corpora
- the aim is NOT building the best model but rather developing a model that shows childlike item-by-item error rates

More details: <https://github.com/sigmorphon/2023InflectionST/>

# References

Kodner J., Khalifa S., Batsuren K., Dolatian H., Cotterell R., Akkus F., Anastasopoulos A., Andrushko T., Arora A., Atanalov N., Bella G., Budianskaya E., Ate Y.G., Goldman O., Guriel D., Guriel S., Guriel-Agiashvili S., Kieras W., Krizhanovsky A., Krizhanovsky N., Marchenko I., Markowska M., Mashkovtseva P., Nepomniashchaya M., Rodionova D., Scheifer K., Sorova A., Yemelina A., Young J., Vylomova E.. 2022. The SIGMORPHON–UniMorph 2022 Shared Task 0: Generalization and Typologically Diverse Morphological Inflection. In Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 176–203

Batsuren K., Goldman O., Khalifa S., Habash N., Kieras W., Bella G., Leonard B., Nicolai G., Gorman K., Ate Y.G., Ryskina M., Mielke S. J., Budianskaya E., El-Khaissi C., Pimentel T., Gasser M., Lane W., Raj M., Coler M., Montoya Samame J.R., Siticonatzí Camaiteri D., Rojas E.S., López Francis D., Oncevay A., López Bautista J., Silva Villegas G.C., Torroba Hennigen L., Ek A., Guriel D., Dirix P., Bernardy J.-P., Scherbakov A., Bayyr-ool A., Anastasopoulos A., Zariquiey R., Scheifer K., Ganieva S., Cruz H., Karahóğa R., Markantonatou S., Pavlidis G., Plugaryov M., Klyachko E., Salehi A., Angulo C., Baxi J., Krizhanovsky A., Krizhanovskaya N., Salesky E., Vania C., Ivanova S., White J., Maudslay R., Valvoda J., Zmigrod R., Czarnowska P., Nikkarinen I., Salchak A., Bhatt B., Straughn C., Liu Z., Washington J., Pinter Y., Ataman D., Wolinski M., Suhardijanto T., Yablonskaya A., Stoehr N., Dolatian H., Nuriah Z., Ratan S., Tyers F.M., Ponti E.M., Aiton G., Arora A., Hatcher R.J., Kumar R., Young J., Rodionova D., Yemelina A., Andrushko T., Marchenko I., Mashkovtseva P., Serova A., Prud'hommeaux E., Nepomniashchaya M., Giunchiglia F., Chodroff E., Hulden M., Silfverberg M., McCarthy A.D., Yarowsky D., Cotterell R., Tsarfaty R., Vylomova E.. 2022. UniMorph 4.0: Universal Morphology. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), Marseille, France

Pimentel T., Ryskina M., Mielke S., Wu S., Chodroff E., Leonard B., Nicolai G., Ate Y., Khalifa S., Habash N., El-Khaissi C. {...}, Vylomova E.. 2021. SIGMORPHON 2021 Shared Task on Morphological Reinflexion: Generalization Across Languages. In Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 229-259.

Shcherbakov A., Muradoglu S., Vylomova E. 2020. Exploring Looping Effects in RNN-based Architectures. In Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association, pp. 115-120.

Vylomova E., White J., Salesky E., Mielke S.J., Wu S., Ponti E.M., Maudslay R.H., Zmigrod R., Valvoda J., Toldova S., Tyers F. et al. 2020. SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection. In Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 1-39.

McCarthy A. D., Kirov C., Grella M., Nidhi A., Xia P., Gorman K., Vylomova E. et al. 2020. UniMorph 3.0: Universal Morphology. In Proceedings of The 12th Language Resources and Evaluation Conference, pp. 3922-3931.

Gorman K., McCarthy A., Cotterell R., Vylomova E., Silfverberg M., Markowska M. 2019. Weird Inflects but OK: Making Sense of Morphological Generation Errors. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL) (pp. 140-151). Hong Kong

McCarthy A., Vylomova E., Wu S., Malaviya C., Wolf-Sonkin L., Nicolai G., Kirov C., Silfverberg M., Mielke S., Heinz J., Cotterell R., Hulden M. 2019. The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, ACL, Florence, Italy

Cotterell R., Kirov C., Sylak-Glassman J., Walther G., Vylomova E., McCarthy A., Kann K., Mielke S., Nicolai G., Silfverberg M., Yarowsky D., Eisner J., Hulden M. 2018. The CoNLL-SIGMORPHON 2018 Shared Task: Universal Morphological Reinflexion., Brussels, Belgium

Kirov C., Cotterell R., Sylak-Glassman J., Walther G., Vylomova E., Xia P., Faruqui M., Mielke S., McCarthy A., Kbler S., Yarowsky D., Eisner J., Hulden M. 2018. UniMorph 2.0: Universal Morphology. In Proceedings of Language Resources and Evaluation (LREC), Miyazaki, Japan

Cotterell R., Kirov C., Sylak-Glassman J., Walther G., Vylomova E., Xia P., Faruqui M., Kubler S., Yarowsky D., Eisner J., Hulden M. 2017. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflexion in 52 Languages. In Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflexion, CoNLL, Vancouver, Canada



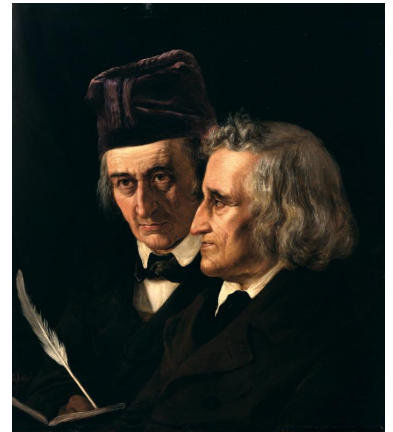
# Language as a Dynamic System: Phonetic Changes

Crucial insights made by historical linguists (Grimm, 1822; Rask, 1818):  
sound change proceeds in a surprisingly **regular, systematic manner**

**Grimm's law, or the First Germanic Sound Shift:**

- $b^h \rightarrow b \rightarrow p \rightarrow f$
- $d^h \rightarrow d \rightarrow t \rightarrow \theta$
- $g^h \rightarrow g \rightarrow k \rightarrow x$
- $g^{wh} \rightarrow g^w \rightarrow k^w \rightarrow x^w$

E.g., Proto-Indo-European  $*h_2ébōl$  → English “apple”, Dutch “appel”



# Language as a Dynamic System: Cognates

**Cognates** sets of words that share a **common origin regardless of their meaning** in historical-comparative linguistics and that should not contain borrowed words.

**Cognate items (reflexes)** typically show **regular sound correspondences**: English `t` typically corresponds to a German `ts` (compare `ten` vs. `zehn`), and English `d` corresponds to German `t` (compare `dove` vs. `Taube`).

The more reflexes a cognate set has in different languages, the easier it is to predict reflexes in individual languages.

# SIGTYP 2022 Shared Task on the Prediction of Cognate Items

Cognate Set	German	English	Dutch
ASH	aʃə	æʃ	ɑs
BITE	baisən	bait	bɛitə
BELLY	baux	?	bœik

*Wordlists were derived from Lexibank (List et al., 2022)*

# SIGTYP 2022 Best Performing Systems (Kirov, Sproat, Gutkin; 2022)

- **Image Inpainting Model**: A CNN that was initially developed to restore 2D images (dimensions correspond to languages and cognate phonemic representations)
- **Neighbor Transformer Model** that was initially developed to find problems in the readings of Japanese place names spelled in kanji. The model was enriched with synthetic n-gram based instances.

Overperforming all non-neural/traditional alignment-based baselines

# SIGTYP 2022 Results

- A good starting point for computer-assisted approaches and can already provide active help in various practical annotation tasks in historical linguistics.
- Scholars working on the reconstruction of certain language families could use predicted proto-forms and later manually correct them

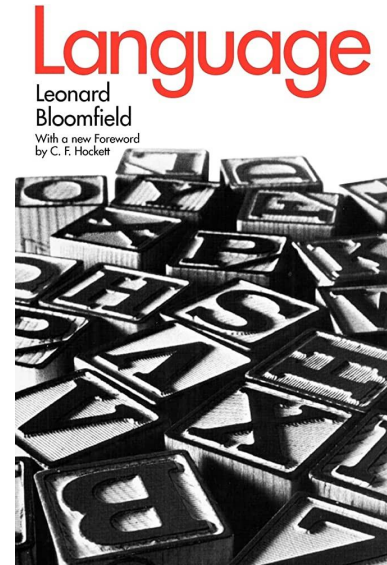
List, J. M., Vylomova, E., Forkel, R., Hill, N., & Cotterell, R. (2022, July). The SIGTYP 2022 Shared Task on the Prediction of Cognate Reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP* (pp. 52-62).

More Information: <https://github.com/sigtyp/ST2022>



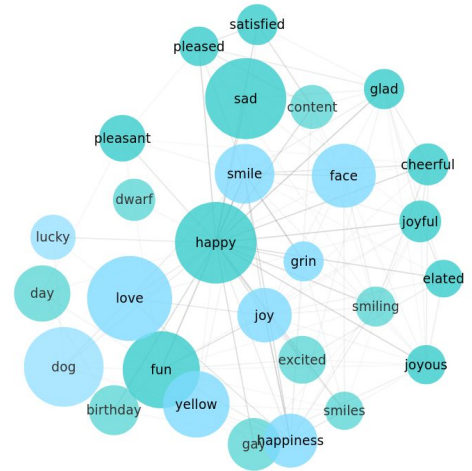
# Language as a Dynamic System: Semantic Changes (Bloomfield, 1933)

- **semantic narrowing-widening**: the Old English mete ‘food’ > meat ‘edible flesh’) vs. the Middle English briddle ‘young birdling’ > bird ‘birds of all ages’
- **extension by analogy**: the Old English bītan ‘to bite’ > the Middle English bitter ‘acid’
- **the litotes–hyperbole axis**: Proto-West Germanic \*kwalljan ‘to make suffer’ > the Old English cwellan ‘to kill’ vs. the Vulgar Latin \*extonare ‘to strike with thunder’ > astonish ‘to surprise greatly’

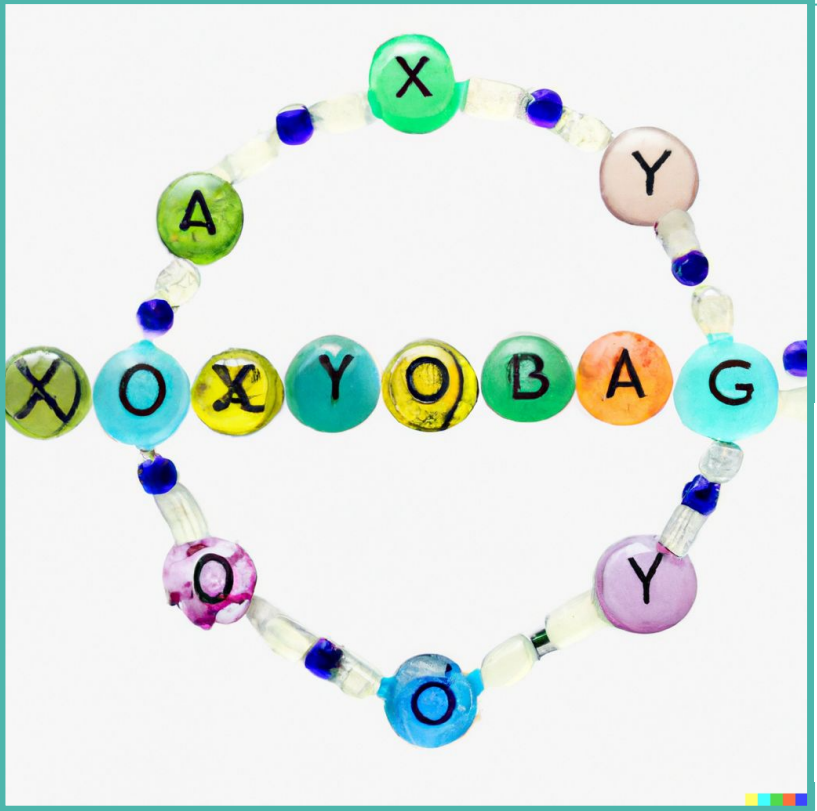


# Language as a Dynamic System: Laws of Lexical Change

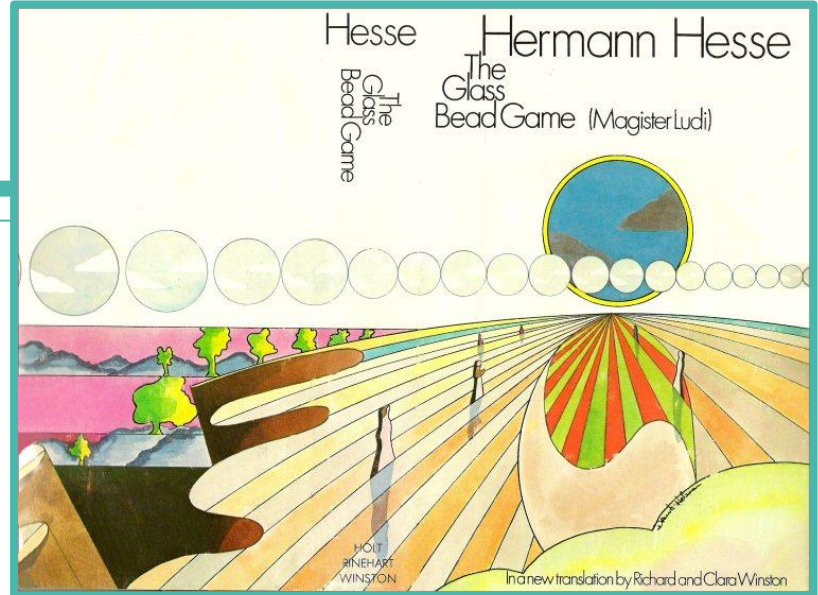
- "law of conformity" (Hamilton, Leskovec, et al. ,2016): frequently used words change slower
- "law of innovation" (Hamilton, Leskovec, et al. ,2016): polysemous words change at faster rates
- "law of prototypicality" (Dubossarsky, 2015): more prototypical words (centroids) change slower
- Pagel et al. (2007): nouns are replaced more easily than verbs
- "law of parallel change" (Xu and Kemp, 2015): words that are semantically linked, like synonyms or antonyms, experience similar change over time
- Winter et al. (2014): words used more diverse contexts are more likely to change
- Jackson et al. (2023): negative concepts changing more rapidly



# A Glass Bead Game of



\*  
.  
\*  
.  
\*  
O  
L  
O  
g  
y

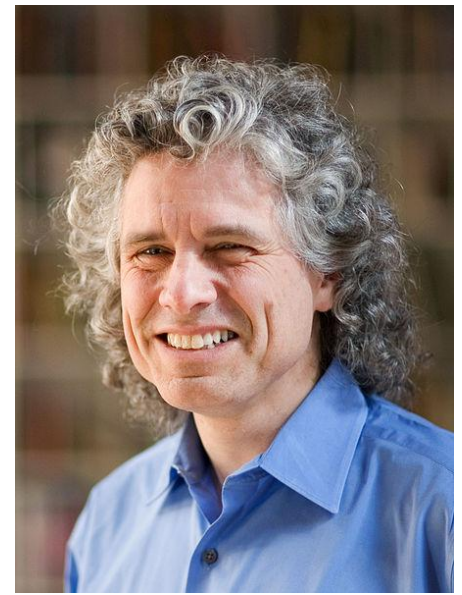


Contemporary Computational  
Approaches to **Linguistic Morphology**,  
**Typology**, and  
**Social Psychology**

# The Better Angels of Our Nature

(Pinker, 2011)

Demonstrates that violence  
(homicide, genocide, torture, criminal justice, wars) has been in  
decline over millennia.



Four “Angels” that reduce violence and increase altruism:  
Empathy, Self-control, Morality and taboo, Reason

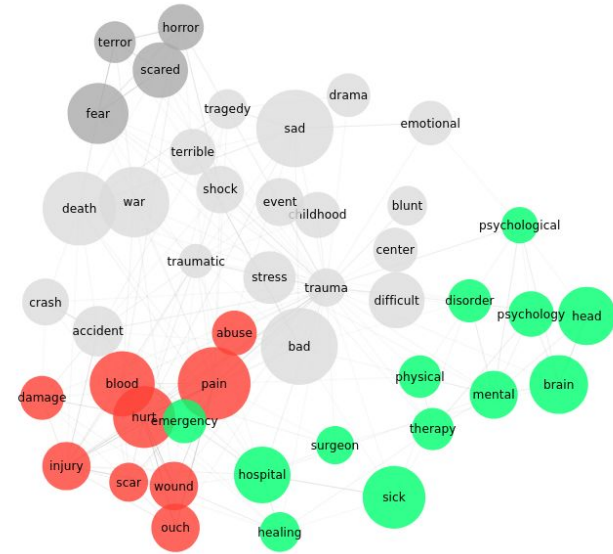
Five “Inner Demons”: Predation, Dominance, Revenge, Sadism,  
Ideology

# “Concept Creep” (Haslam, 2016)

“Concept Creep”: the semantic inflation of harm-related concepts (such as trauma, addiction, bullying, harassment, and prejudice) in the West. This reflects a growing cultural sensitivity to harm.

## Possible Reasons:

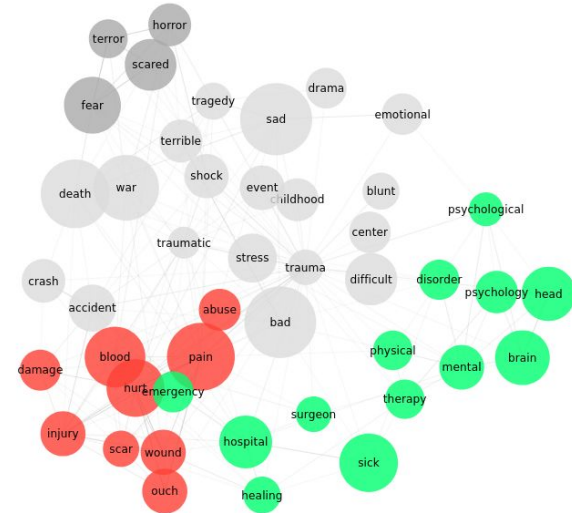
- Decline in violence (Pinker, 2011)
- Rights revolutions of the 1960s
- rise of postmaterialist values in the West (Inglehart, 2008), e.g. an increased concern with quality of life relative to material security, encouraging a reduced tolerance for suffering
- “prevalence-induced concept change” by Levari et al. (2018): *as blue dots, angry faces, and unethical research proposals became scarcer in their studies, participants came to classify stimuli they had previously judged to be purple, neutral, and ethical as examples of these categories.*



# “Vertical and Horizontal Creep” (Haslam, 2016)

In “vertical creep” (Bréal’s “*épaississement de sens*”) concept meanings became less stringent by a relaxation of criteria or a lowering of a threshold, such that **less severe or intense phenomena came to be regarded as examples of the concept** (e.g., vicarious trauma, unrepeated bullying)

In “horizontal creep” (Bréal’s “*élargissement de sens*”) by contrast, concepts broaden by **incorporating qualitatively different and new phenomena** (e.g., addition of new domains of psychopathology in DSM, recognition of new targets of prejudice).



# Concept Creep (Haslam, 2016): Can we verify and measure it computationally? Yes!

## Data:

Psychology domain: A corpus of ~831k abstracts from psychology journals:

- extracted from PubMed and E-Research Databases
- covers the time period of 1930–2017

General Domain: CoCA + CoHA

## Approach:

1. Train a diachronic language model, getting embeddings for each decade/year since the 1970s
2. Measure semantic breadth for each concept (e.g., trauma, bullying) in each decade
3. More fine-grained analysis: explore nearest neighbors for each concept's embedding

# Concept Creep (Haslam, 2016): Can we verify and measure it computationally? Yes!

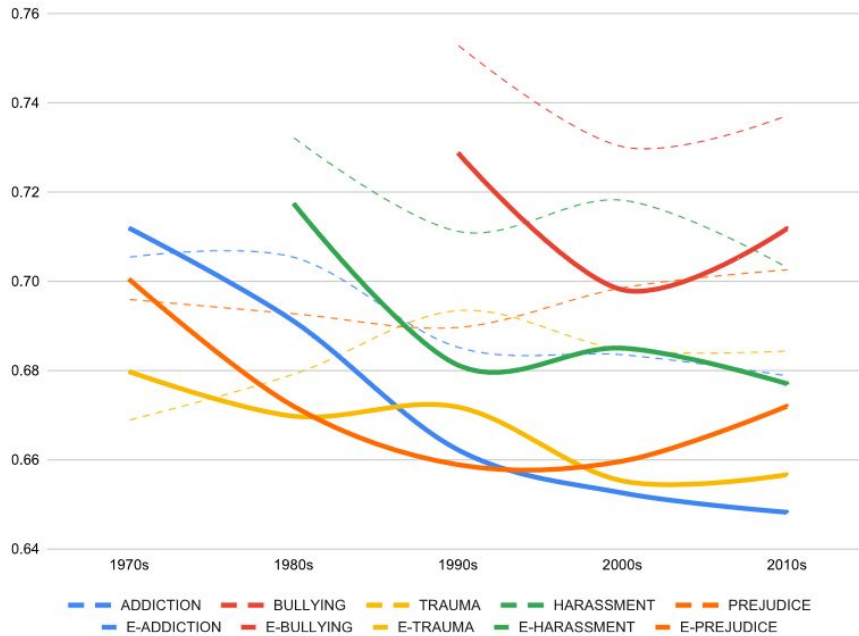
## Approach:

1. Train a diachronic language model, getting embeddings for each decade/year since the 1970s
  - a. Train a LM on the whole corpus, then continue training on each epoch (e.g. decade)
  - b. (Option 2) Train a LM for each epoch (decade) independently, then align vector spaces using orthogonal *Procrustes* (Hamilton et al., 2016)
2. Measure semantic breadth for each concept (e.g., trauma, bullying) in each decade
  - a. Sample 50 sentences containing the target concept from each epoch (year/decade)
  - b. Get contextual representations for each occurrence
  - c. Measure pairwise cosine similarity
  - d. Repeat sampling ~10 times, and get mean similarity values
3. More fine-grained analysis: explore nearest neighbors for each concept's embedding in each epoch/decade (based on cosine similarity)

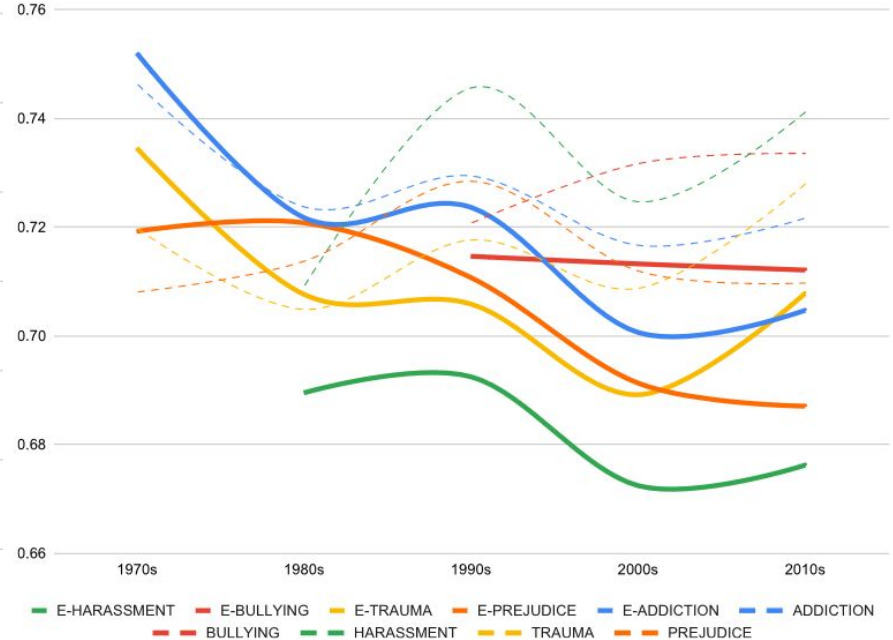
Vylomova, E., Murphy, S., & Haslam, N. (2019, August). Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (pp. 29-34).



# Concept Creep (Haslam, 2016): Semantic Similarity (inv. Breadth)



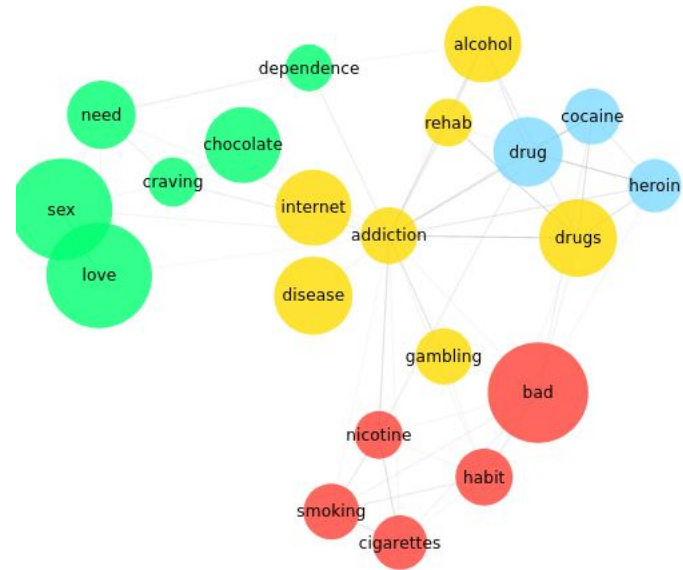
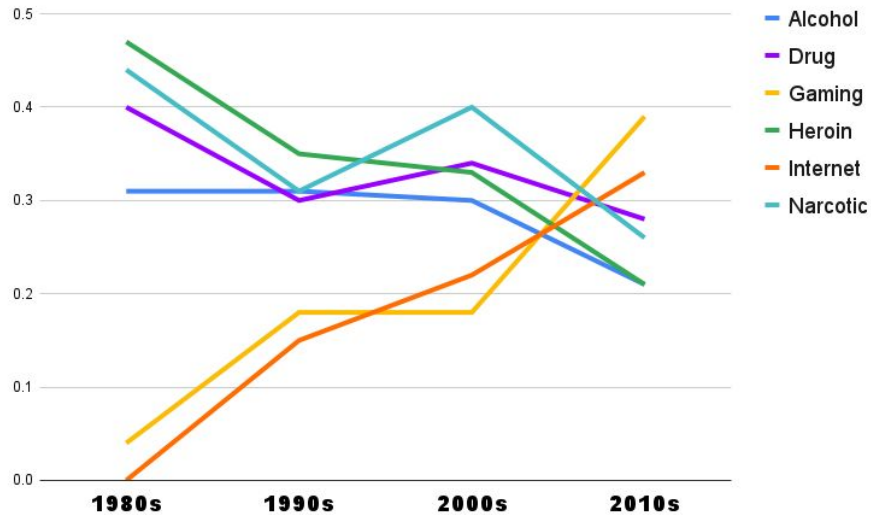
## Psychology



## General Domain

# Concept Creep (Haslam, 2016): Addiction

Substance → Behaviour



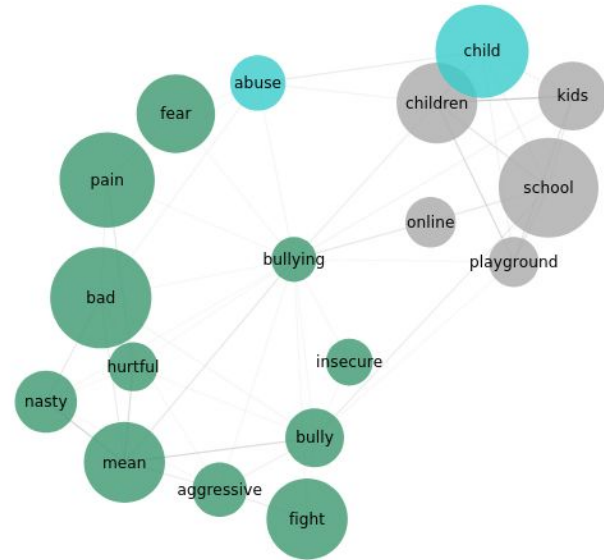
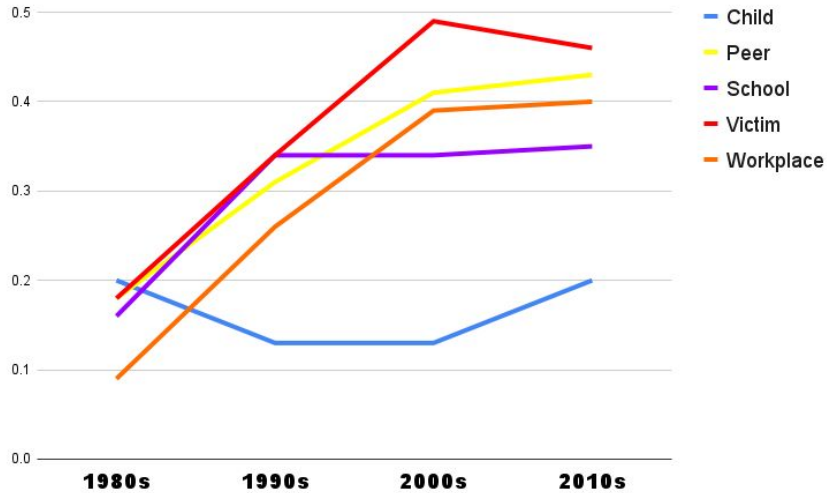
Top Nearest Neighbors (Psychology)

Word Associations

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

# Concept Creep (Haslam, 2016): Bullying

School → Workplace

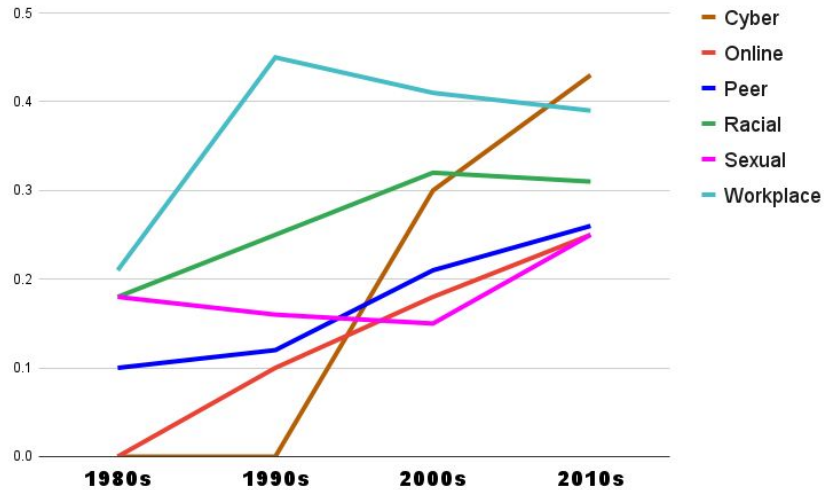


Top Nearest Neighbors (Psychology)

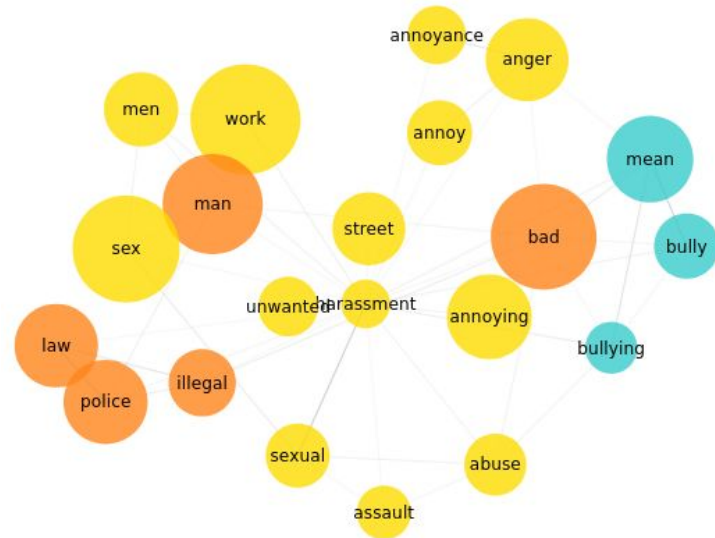
Word Associations

# Concept Creep (Haslam, 2016): Harassment

Workplace → Online



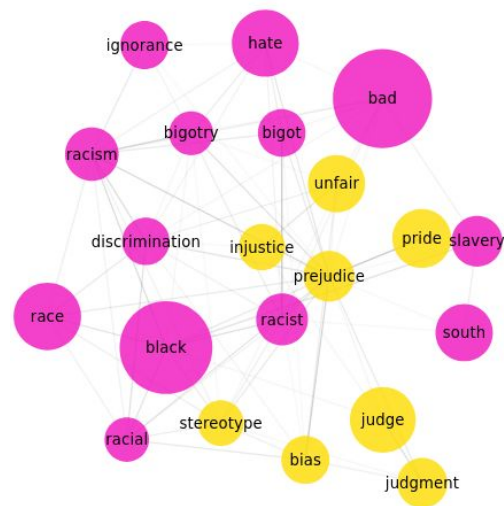
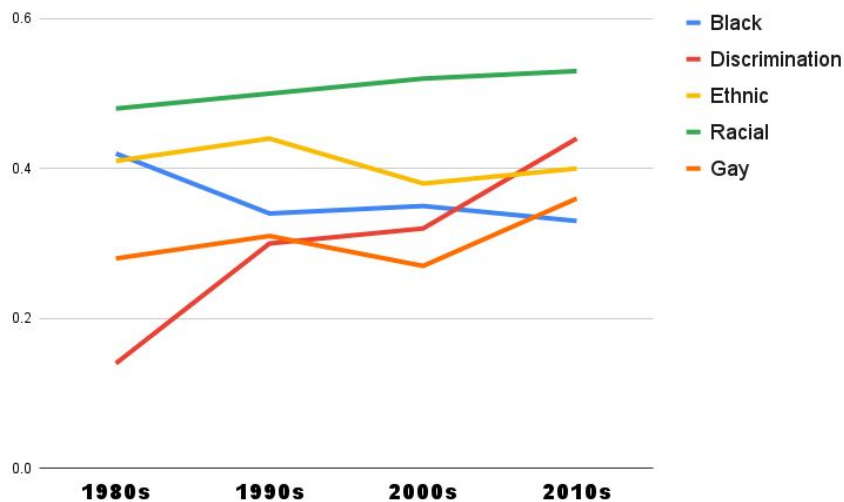
Top Nearest Neighbors (Psychology)



Word Associations

# Concept Creep (Haslam, 2016): Prejudice

## Extends to Include Non-Racial Groups

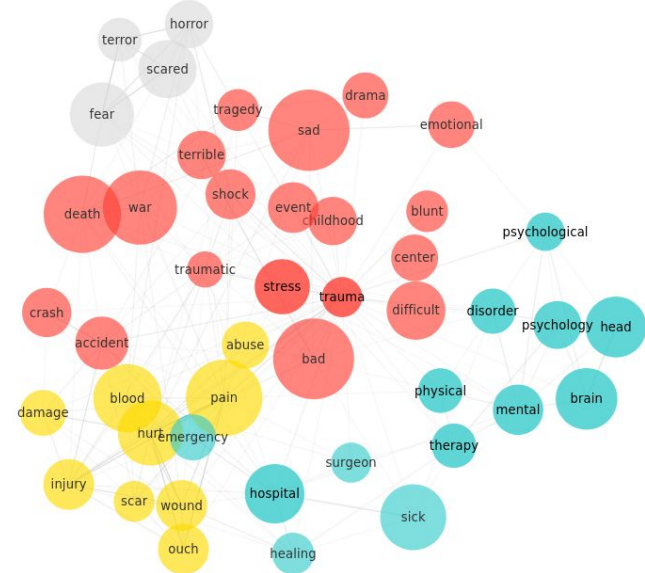
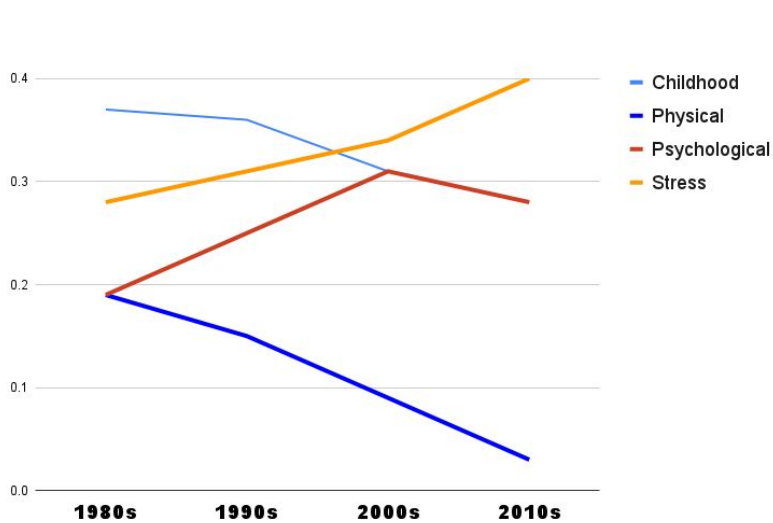


Top Nearest Neighbors (Psychology)

Word Associations

# Concept Creep (Haslam, 2016): Trauma

Childhood/Physical → Emotional/Stress

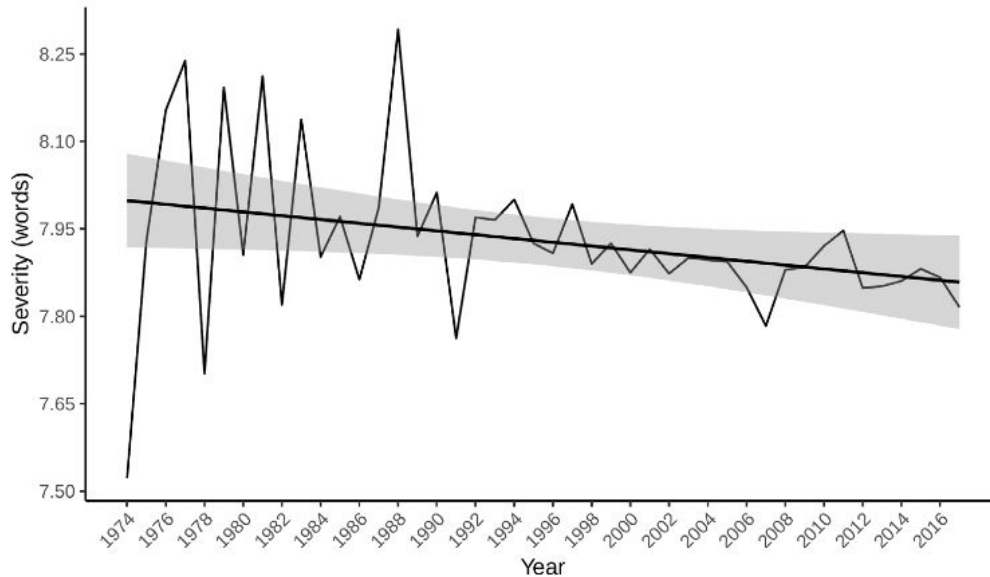


Top Nearest Neighbors (Psychology)

Word Associations

# Concept Creep (Haslam, 2016): Vertical Expansion of Trauma

## Trauma becoming less “severe” in psychology



Investigating valence and arousal (derived from Warriner, 2013) of “trauma” collocates for each year.

Baes, N., Vylomova, E., Zyphur, M., & Haslam, N. (2023). The semantic inflation of “trauma” in psychology. *Psychology of Language and Communication*, 27(1), 23-45.

## Concept Creep (Haslam, 2016): More on the Topic

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, 27(1), 1-17.

Haslam, N., Dakin, B. C., Fabiano, F., McGrath, M. J., Rhee, J., Vylomova, E., Wheeler, M. A. (2020). Harm inflation: Making sense of concept creep. *European Review of Social Psychology*, 31(1), 254-286.

Haslam, N., Vylomova, E., Zyphur, M., & Kashima, Y. (2021). The cultural dynamics of concept creep. *American Psychologist*, 76(6), 1013.



## And **More** Research **Not** mentioned here:

(submitted) Culture in Psychology. Joint w/ Yoshihisa Kashima, Naomi Baes, Takeshi Hamamura, Nick Haslam

(under review) Predicting Human Translation Difficulty Using Automatic Word Alignment. Joint w/Zheng Wei Lim, Trevor Cohn, Charles Kemp

A Computational Approach to Discovering Cultural Keywords across Languages. (2022). Preprint.  
*by Zheng Wei Lim, Harry Stuart, Simon De Deyne, Terry Regier, Ekaterina Vylomova, Trevor Cohn, Charles Kemp*

Wheeler M., Vylomova E., McGrath M., Haslam N. 2021. More Confident, Less Formal: Stylistic Changes in Academic Psychology Writing From 1970 To 2016. *Scientometrics*, 126.12, 9603--9612

Haslam N., Vylomova E., Murphy S.C., Wilson S. 2021. The Neuroscientification of Psychology: The Rising Prevalence of Neuroscientific Concepts in Psychology From 1965 To 2016. *Perspectives on Psychological Science*. 2021 Jul.

Salesky E., Abdullah B., Mielke S., Klyachko E., Serikov O., Ponti E.M., Kumar R., Cotterell R., Vylomova E. 2021. SIGTYP 2021 Shared Task: Robust Spoken Language Identification. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pp. 122-129.

Muradoglu S., Evans N., Vylomova E. 2020. Modelling Verbal Morphology in Nen. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pp. 43-53.

Bjerva J., Salesky E., Mielke S.J., Chaudhary A., Giuseppe C., Ponti E.M., Vylomova E., Cotterell R., Augenstein I. 2020. SIGTYP 2020 Shared Task: Prediction of Typological Features. In *Proceedings of the Second Workshop on Computational Research in Linguistic Typology*, pp. 1-11.

# THANK YOU!

Questions?



*“Wugs” generated by AI*