September 2025

# A Survey of Idiom Datasets for Psycholinguistic and Computational Research

Michael Flor
Educational Testing Service
Princeton, New Jersey
USA

Xinyi Liu
Montclair State University
Montclair, New Jersey
USA

Anna Feldman
Montclair State University
Montclair, New Jersey
USA

presented at

KONVENS 2025

Hildesheim, Germany

# Introduction

Datasets are an important driving force in many areas of research

Research on idiomatic expressions
has accumulated over 50 datasets in the past 30 years

Two domains are particularly prolific for idioms:
psycholinguistics and computational linguistics

We present a survey of idioms datasets,
describing what is in them and what research they support

# Idioms

Figurative idiomatic expressions,

where the meaning of an expression
is not a composition of the meanings of the components

*kick the bucket, spill the beans, best thing since sliced bread, shed light on…*

*steal one’s heart, on the warpath, game changer, thinly veiled*

September 2025

# Psycholinguistic Research

….

# Psycholinguistics

Psycholinguistic research on idiomatic expressions is focused on issues like:

- How are idioms represented (in the mind) ?
- How are idioms processed during comprehension ?
- What are the connotations of idioms ?
- Difficulties learners experience with acquisition and production of idioms in a foreign language (L2)
- Differences between L1 and L2 speakers
- Experiments with human participants !

Norming studies are always about **types**, not tokens (instances).

Researchers are interested in **norming studies** on variables like:

- Knowledge (of idiom)
- Meaningfulness
- Familiarity
- Frequency
- Literality
- Age of Acquisition
- Predictability
- Decomposability / Compositionality
- Syntactic Flexibility
- etc…

# Psycholinguistics

Variables recorded with participants in psycholinguistic norming studies 1

### **Knowledge** of idiom

- Knowledge of the figurative meaning
- Sometimes just binary yes/no
- Sometimes rated on a scale 1-5 (**meaningfulness**)
- Sometimes actually provide definition or description (**objective knowledge**)

### **Familiarity** of idiom

- Sometimes called **subjective frequency**
- Subjective estimate how familiar one is with the expression
- Sometimes on a scale 1-5 or 1-7
- Sometimes estimation of how other people are familiar with the expression

### **Frequency** of idiom

- Sometimes called **objective frequency**
- Not easy to measure!
- Sometimes measure frequency of constituent words in corpus with ‘corrections’
- Sometimes measure frequency of expression in corpus (but:
  In what sense?
  In which forms? )

### **Literality** of idiom

- Does the idiom have a plausible literal meaning?

  *Break the ice*
  *Shoot the breeze*
- Sometimes rate on a scale
- Also called **literal plausibility,** sometimes called **ambiguity**

# Psycholinguistics

Variables recorded with participants in psycholinguistic norming studies 2

### **Predictability** of idiom

- Probability of completing an incomplete expression idiomatically
- Often estimated with a cloze task

*Ben is in the seventh* __ (heaven)

### **Age of Acquisition**

- Sometimes called **AoA**
- Subjective estimate of age (range) when expression was first learned
- Typically for native speakers

### **Length** of idiom

- Number of words
- Number of characters

### **Syntactic form**

- Many verb-noun idioms
- But less common: noun phrases (NPs), adjectival and adverbial phrases

### **Syntactic Flexibility**

- Does the idiom allow syntactic operations, and still carries its meaning?
- It depends on POS of the expression, etc.!!!!

  *Break the ice:*
  *At the party, the ice was broken.*
  *Shoot the breeze:*
  **The breeze was shot.*
- Not an easy measure, might need to show modifications to participants

*ets MONTCLAIR STATE UNIVERSITY

# Psycholinguistics

Variables recorded with participants in psycholinguistic norming studies 3

## Compositionality

- Also called **Decomposability**
- To what extent the component words of the idiom contribute individually to the meaning of expression

*Spill the beans = reveal secrets*

*Kick the bucket = die*

- Often estimated on a scale rating (1-5 or 1-7)

## Transparency

- How easy it is to infer the figurative meaning from the literal meaning
- Also called **semantic transparency** and opaqueness

*Keep in touch*
*Kick the bucket*

- Rated on a scale
- Problematic measure, since it is based on participant intuitions that can be wrong

## Emotional Valence

- The degree of emotional/sentiment values (positive or negative)
- Rated on a scale

## Arousal

- Excitation-potential of the stimulus (idiom) (arousing or calm)
- Rated on a scale

## Concreteness

- Does the idiom refer to a state or event or attribute that can be experienced via sensory modalities?
- Rated on a scale

## Imageability

- Does the idiom (figurative sense) refer to something that can be visualized?
- Rated on a scale

# Psycholinguistics

- Some psycholinguistic idioms norming studies: how many idioms are covered

| Authors | Idioms | Language |
|---|---|---|
| Cronk et al. 1993 | 245 | English |
| Libben & Titone, 2008 | 210 | English |
| Callies 2009 | 300 | French |
| Tabossi et al., 2011 | 245 | Italian |
| Bonin et al., 2013 | 305 | French |
| Citron et al., 2016 | 619 | German |
| Li et al., 2016 | 350 | Chinese |
| Beck & Weber, 2016 | 300+300 | English & German |
| Bulkes & Tanner, 2017 | 870 | English |
| Nordmann & Jambazova, 2016 | 90+100 | Bulgarian & English |
| Bonin et al., 2018 | 160 | French |
| Hubers et al., 2019 | 384 | Dutch |
| Gavilan et al., 2021 | 1252 | Spanish |
| Pagliai 2023 | 150+150 | English & Italian |
| Lada et al., 2024 | 400 | Greek |
| Morid & Sabourin, 2024 | 210 | English |
| Gridneva et al., 2025 | 376 | Russian |

September 2025

# Computational Linguistic Research

…..

# Computational Linguistics

In Computational Linguistics, research on idiomatic expressions is focused a variety of topics:

- Finding idiomatic expressions (types)
- Detecting idiomatic expressions (tokens/instances) in context.
- Connecting idiomatic expressions across languages
- Translation of idioms (+ in context)
- Computing sentiment and emotions in texts, based on idioms
- Detection/identification of compositionality

Researchers in computational linguistics develop datasets with:

- Annotations of idioms in context, especially figurative vs. literal instances
- Paraphrases of idioms
- Cross-lingual handling of idioms (translations),
  types and instances-in-context
- Sentiment annotations of idioms
- Ratings of compositionality

# Computational Linguistics

- Datasets for idiom detection in context

| Dataset | Idioms | Language |
|---|---|---|
| VNCTokens 2008 | 53 types, 3K tokens in context | English |
| OpenMWE 2008 | 146 types, 102K Sentences | Japanese |
| Semeval2013, task 5b | 85 types, 4350 instances, 5-sent. context | English |
| Idioment 2015 | 580 types, 2421 in sentences + sentiment ratings | English |
| RU idioms 2018 | 100 types, 2.4K in paragraphs + 3K literal | Russian |
| CCT 2018 | 7395 types, 100K sentences | Chinese |
| ChID 2019 | 3848 types, 518K paragraphs | Chinese |
| MAGPIE 2020 | 1756 types, 56K instances in context | English |
| EPIE 2021 | 359 types, 22K sentences | English |
| PIE 2021 | 823 types, 5170 sentences + with paraphrases | English |
| AStitchInLanguageModels | 223+113 types, 4.5/1.8K sentences | Eng+Portugese |
| Semeval2022, task2 | 5352/2555/776 sentences | E+P+Galician |
| PIE-English 2022 | **20K instances** | English |
| IDEM 2024 | 9685 sentences, + annotated emotions | English |

# Computational Linguistics

- Datasets for paraphrase, sentiment, and compositionality

| Dataset | Idioms | Language |
|---|---|---|
| Idiom Paraphrases 2015 | 2432 idioms with paraphrases, 1400 annotated as mutual paraphrases | English |
| Idiom substitution 2016 | 176 idioms with definitions + substitutions | English |
| SLIDE 2018 | 5000 types with sentiment annotation | English |
| FLUTE 2022 | 1000 idiomatic sent. + contradict/entail | English |
| | | |
| CIKN 2010 | 38K idioms with linguistic annotations | Chinese |
| Reddy et al., 2011 | 90 nominal compounds with compositionality | English |
| Nominal Compounds 2019 | 190/180/180 rated for compositionality | Eng., French, Portuguese |
| Swedish MWEs 2020 | 96 MWEs rated for compositionality | Swedish |
| NCS 2021 | 280/180 types, 5620/3600 sentences | Eng., Portug. |

# Computational Linguistics

- Datasets for translation, multilingual, and working with LLMs

| Dataset | Idioms | Language |
|---|---|---|
| IMIL 2018 | 2208 English idioms + translations and sentiment, in 7 Indian languages, 250K sentences | English + Hindi, Urdu, Bengali,Tamil,Telugu, Gujarati, Malayalam |
| Idiom Translation DS | 1500+1500 translations in sentences | Eng. + German |
| LIDIOMS 2018 | 815 idioms (total types*), linked + definitions + translations | Eng, German, Italian, Portuguese, Russian |
| ID10M 2022 | Auto generated: 10K idioms (types), 262K sentences,<br>Gold 200 idioms: Eng,German,Ital.,Span. | 10 lang.: Eng., Chin., Dutch, French, German, Italian, Jap., Polish, Portug., Span. |
| PETCI 2022 | 4310 Chinese idioms, 30K English translations | Chinese, English |
| IDIOMEM 2023 | 814 idiom types – for LLM probing | English |
| IdiomsInCtx-MT 2024 | Idiom instances in sentences, 1000 professional translations per lang.pair | Eng.-German, Eng.-Russian |
| MAPS 2024 | 360-420 idioms per lang., with (non/)-entailing sentences | Eng + 4 lang. |
| Multiling. Idioms and Similes in LLM 2024 | 316 instances, LLM-generated sentences | Eng + 11 lang. |
| IdiomKB 2024 | A merger of several large datasets | Eng.,Chin,,Jap. |

*ets MONTCLAIR STATE UNIVERSITY

# Idiom datasets

## The inventory

- We maintain a continuously expanding list of idiom datasets at [https://github.com/maafiah/IdiomsResearch](https://github.com/maafiah/IdiomsResearch)
- With links to papers and datasets
- Contributions and collaboration are welcome!

## NOW

- There is little integration between psycholinguistic and computational directions
- But **cross-disciplinary** connections are appearing, for example:
  Oh et al. (arXiv:2506.01723v3) analyzed LLM interpretation of idioms using Cronk et al. (1993) dataset.

## NEXT

- There is place for much additional work on datasets.
- Expanding attributes of idioms, like linguistic and psycholinguistic features, sentiment, etc.,
  - for **larger** sets of idioms
- Expanding to more languages

# Thank you

Project website:

https://github.com/maafiah/IdiomsResearch

Contact:

mflor@ets.org