Are idioms surprising?

J. Nathanael Philipp[1,2] Michael Richter[2] Erik Daas[2] Max Kölbl[3]

1: Serbski Institut, August-Bebel-Straße 82, 03046 Cottbus, nathanael@philipp.land

2: Leipzig University, Augustusplatz 10, 04109 Leipzig, mprrichter@gmail.com, erik.daas.uni@outlook.de

3: Osaka University, 1-5 Yamadaoka, Suita, 565-0871 Osaka, max.w.koelbl@gmail.com

Aim

Identification of **Idiomatic Expressions** (**IE**) in English with an information theoretic model.

General assumption: **IE** are semantic outliers compared to literals, thereby standing out more prominently in the text.

IE in focus

Verb-Noun-Constructions (**VNC**) that can be interpreted both literally and idiomatic:

pull plug, ***get the sack***, ***blow whistle***, etc.

Data

British National Corpus:

1,997 sentences, labelled as idioms

535 sentences, labelled as literals

Leipzig Wortschatz:

2 x 1 million sentences of natural language, from News- and Wikipedia-corpora (reference dataset)

Information theory: Concepts and measures

Topic Context Model (**TCM**):

- an extended topic model (LDA),
- calculates contextualised information, i.e., **surprisal**, as feature of words,
- contexts are topics in the environment of words.

➢ The **surprisal** represents the amount of un-/certainty regarding the language processor's expectations.

$$\overline{\boldsymbol{surprisal}}(w_d) = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P(w_d|t_i)$$

$$P(w_d|t_i) = \frac{c_d(w_d)}{|d|} WT_{w_d,t_i} P(t_i|d)$$

Information Density:

- comparison of information flow between **IE** and literals,
- utilizes the concept of **Local Uniform Information Density**

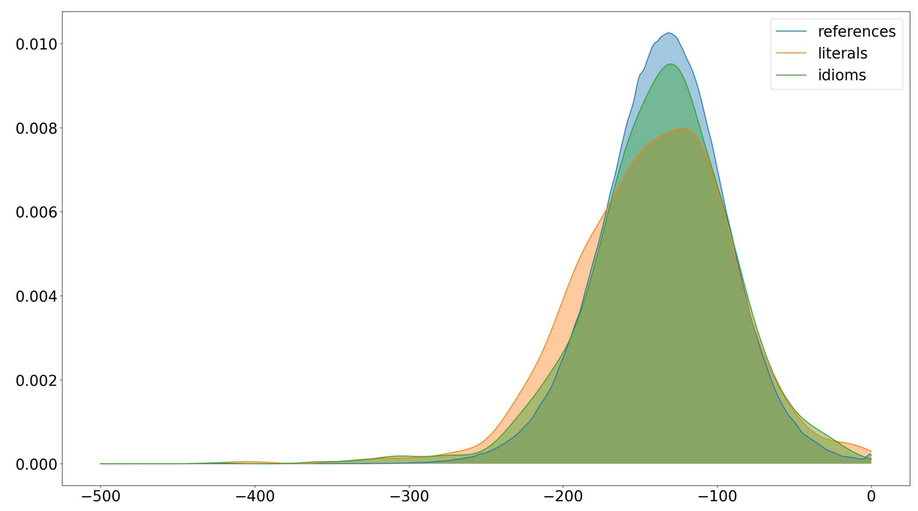
$$\boldsymbol{UID_{LOCAL}} = -\frac{1}{n}\sum_{i=1}^{n} (id_i - id_{i-1})^2$$

Limitations

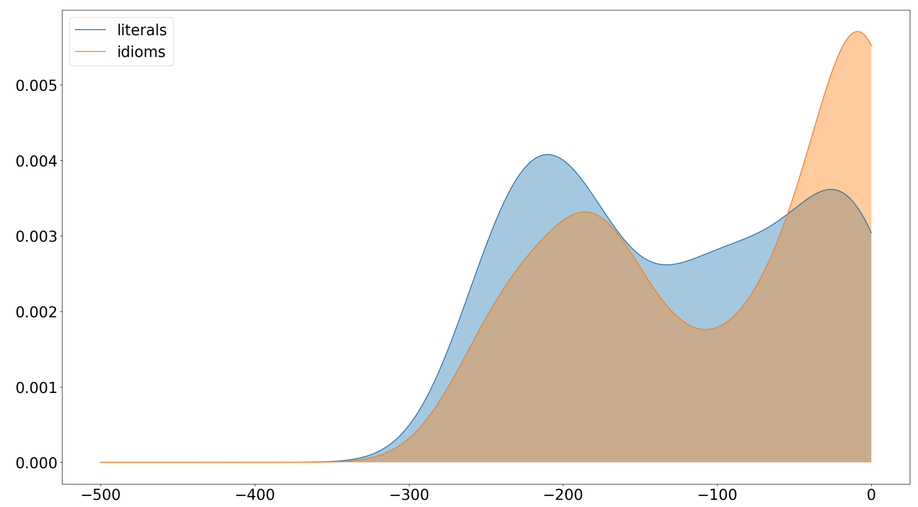
- The News and Wikipedia corpora consist of single sentences, limiting the **TCM**'s effectiveness.
- Future research should use longer texts for more valid results, especially with larger corpora for literals.
- Expanding to other languages requires annotated corpora for training classification models, a future research goal.

Results

- **Welch tests** were conducted to compare mean differences in data for **surprisal**.
- Significant mean differences:
  - ➢ between **IE** and literals ($t = 2.19$, $p = 0.029$),
  - ➢ between News-Wikipedia and literals ($t = 2.23$, $p = 0.025$).
- No significant mean differences:
  - ➢ between **IE** and News-Wikipedia ($t = -0.40$, $p = 0.69$).
- Effect sizes by Cohen's d were consistently small (e.g., 0.022 for **IE** and literals).
- **Figure 1** shows the distribution of $\mathbf{UID_{LOCAL}}$-values:



- Focus to a **VNC**-List of 12 constructions in **IE** and literals:
- significant difference ($t = -1.955$, p-value = 0.05) with a higher Cohen's d of 0.104.
- **Figure 2** displays $\mathbf{UID_{LOCAL}}$ in **VNC**, indicating that information jumps tend to be smaller in **IE** compared to



Conclusion and discussion

- Surprisingly, **IE** and reference dataset exhibit smaller differences in **surprisal** and $\mathbf{UID_{LOCAL}}$ than literals.
- Global measures were used, and local measures (**VNC**) increased the effect size of **surprisal**.
- Reference dataset may have an idiomatic character, challenging the assumption that **IE** are semantic outliers.

➢ Future research should investigate whether language in general tends to be more idiomatic or literal.