

Gotta catch 'em all: Crowdsourcing a wide-coverage corpus of idiomatic expressions in English



university of
 groningen

Hessel Haagsma, Malvina Nissim & Johan Bos

“What I know of Moscow is that if you *keep your nose clean* and do the work, there’s no hassle.”



“What I know of Moscow is that if you *wash your nose* and do the work, there’s no hassle.”

or

“What I know of Moscow is that if you *behave properly* and do the work, there’s no hassle.”

Please read the following sentence(s) carefully.

What I know of Moscow is that you **keep your nose clean** and do the work you've set out to achieve, and that way there's no hassle.

What is the meaning of "keep your nose clean" in the sentence you have just read? (required)

- Idiomatic
- Literal
- Other

What does "keep your nose clean" mean?

Definition

keep your nose clean
stay out of trouble. informal

Please read the following sentence(s) carefully.

School colleagues were suddenly thin **on the ground**.

What is the meaning of "on the ground" in the sentence you have just read? (required)

- Idiomatic
- Literal
- Other

If it isn't idiomatic or literal, what is it? (required)

- Not an instance of "on the ground"
- Unclear
- Non-standard usage

What is its usage? (required)

It's part of a different idiom

What does "on the ground" mean?

Definition

on the ground
in a place where real, practical work is done.

Existing Corpora

Name	# Idioms	# Instances	Corpus	Patterns
VNC-Tokens	53	2,984	BNC	V+NP
Gigaword	17	3,964	Gigaword	V+NP/PP
IDIX	52	4,022	BNC	V+NP/PP
SemEval-2013	65	4,350	ukWaC	unrestricted

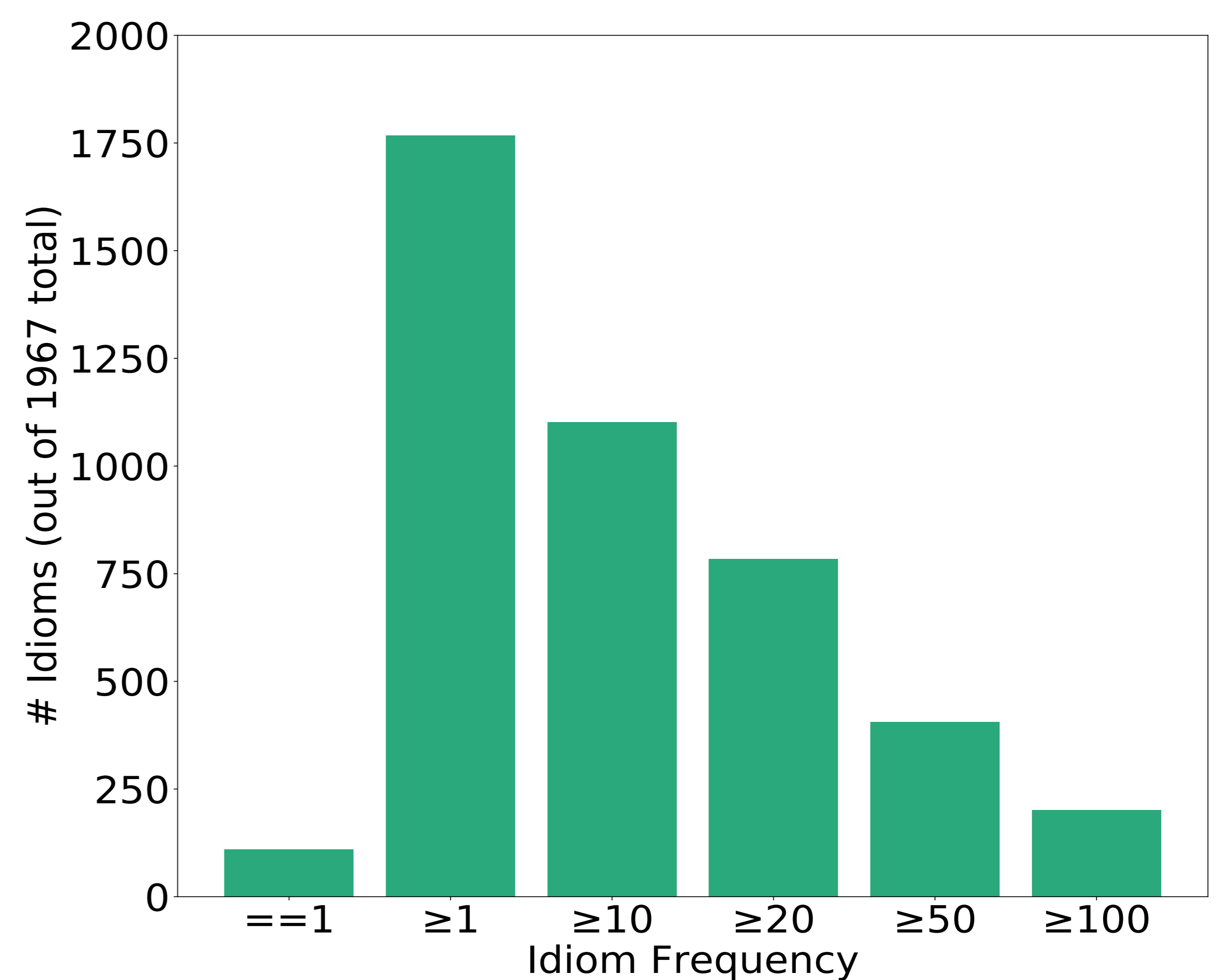
Goals

- ▶ > 1000 types
- ▶ ≤ 200 instances per type
- ▶ Unrestricted syntactic patterns
- ▶ Enable fine-grained evaluation
- ▶ Facilitate deep learning approaches
- ▶ Focus on cross-idiom disambiguation

Setup

- ▶ Automatic pre-extraction of candidates
- ▶ Source: British National Corpus
- ▶ Dictionaries: Wiktionary, Oxford Dictionary of English Idioms, UsingEnglish.com
- ▶ Idioms: (Wiki ∩ Oxford) ∪ (UE.com ∩ Oxford)
- ▶ Platform: FigureEight (f.k.a CrowdFlower)
- ▶ Test questions from existing PIE Corpus

Idiom Frequencies



Open Questions

- ▶ Agreement threshold
- ▶ Number of annotations
- ▶ Idioms within other idioms
- ▶ Single-sense idioms
[.] and the prince *thinks the world* of her.
- ▶ Strong ambiguity
Commandos were digging in *all over the place*.
- ▶ Limits on modification
[.] *jumping on the Tetris/Columns bandwagon*.

