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

Hessel Haagsma, Malvina Nissim & Johan Bos

university of groningen

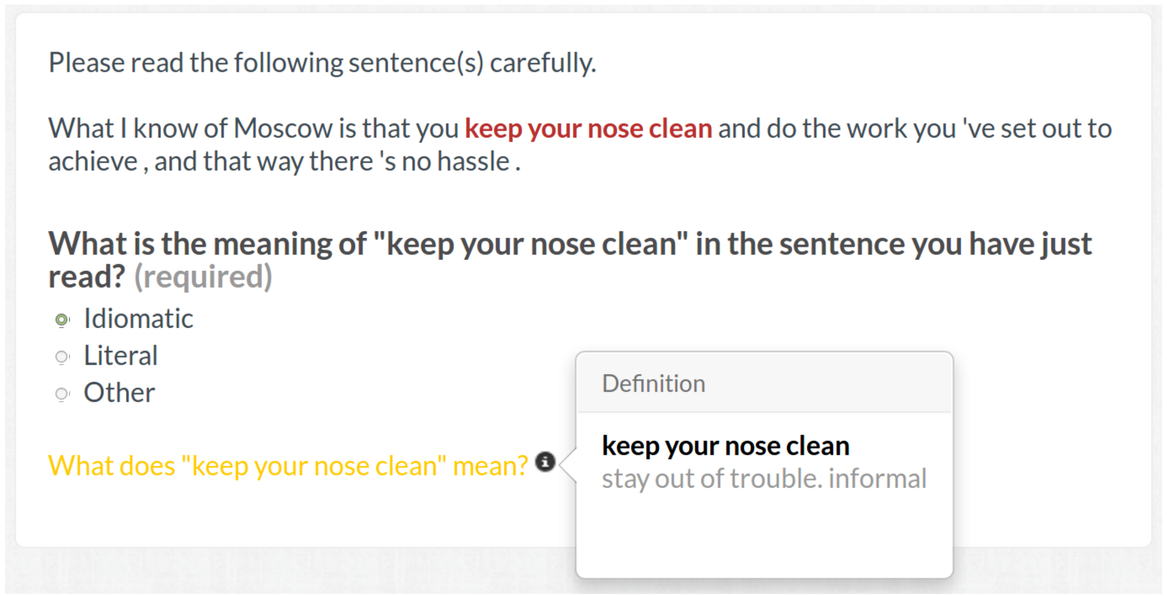
"What I know of Moscow is that if you *keep your nose clean* and do the work, there's no hassle."

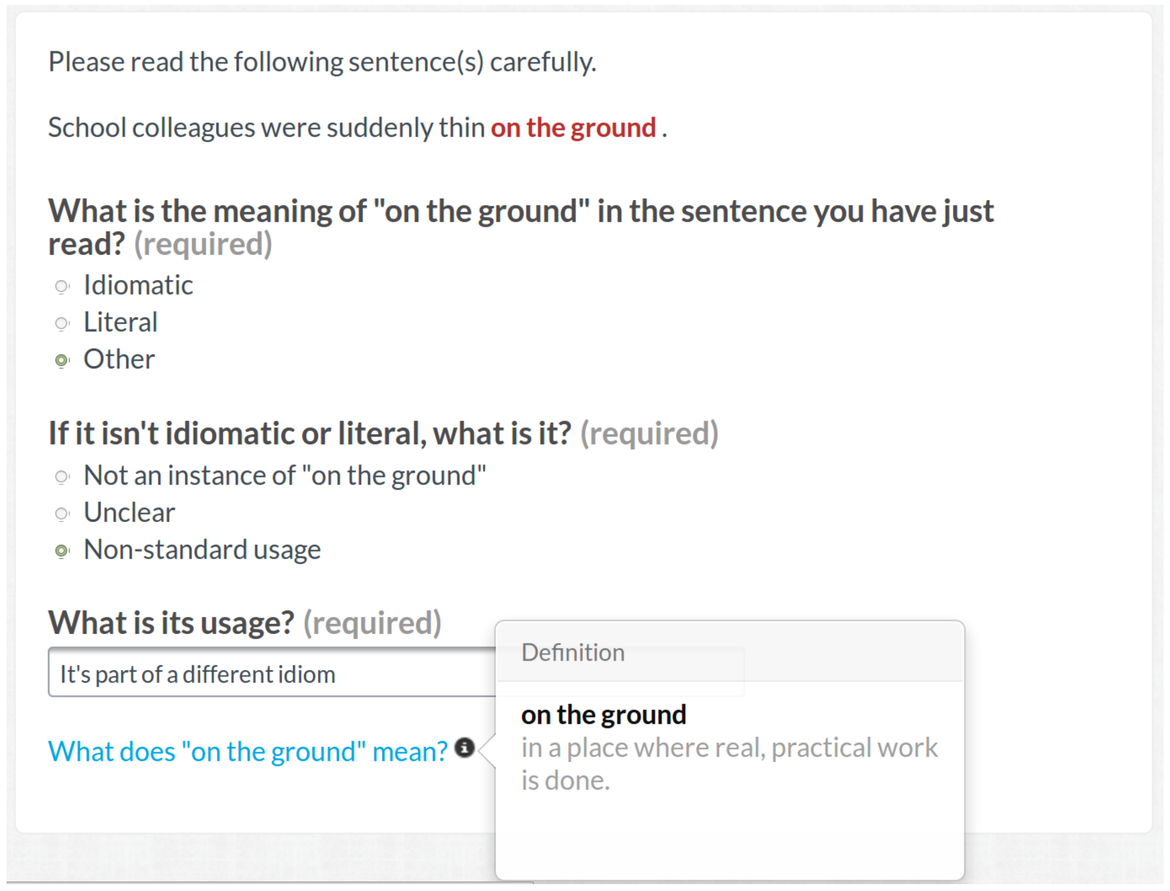
⇓

"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."





Existing Corpora

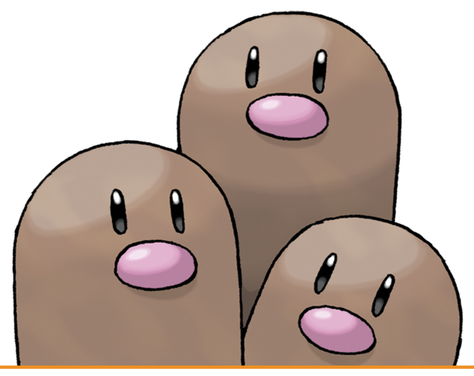
Name	# Idioms	# Instances	Corpus	Patterns
VNC-Tokens	53	2,984	BNC	V+NP
Gigaword	17	3,964	Gigaword	V+NP/PP
IDIX	52	4,022	BNC	V+NP/PP
SemEval-2013	65	4,350	ukWaC	unrestricted

Goals

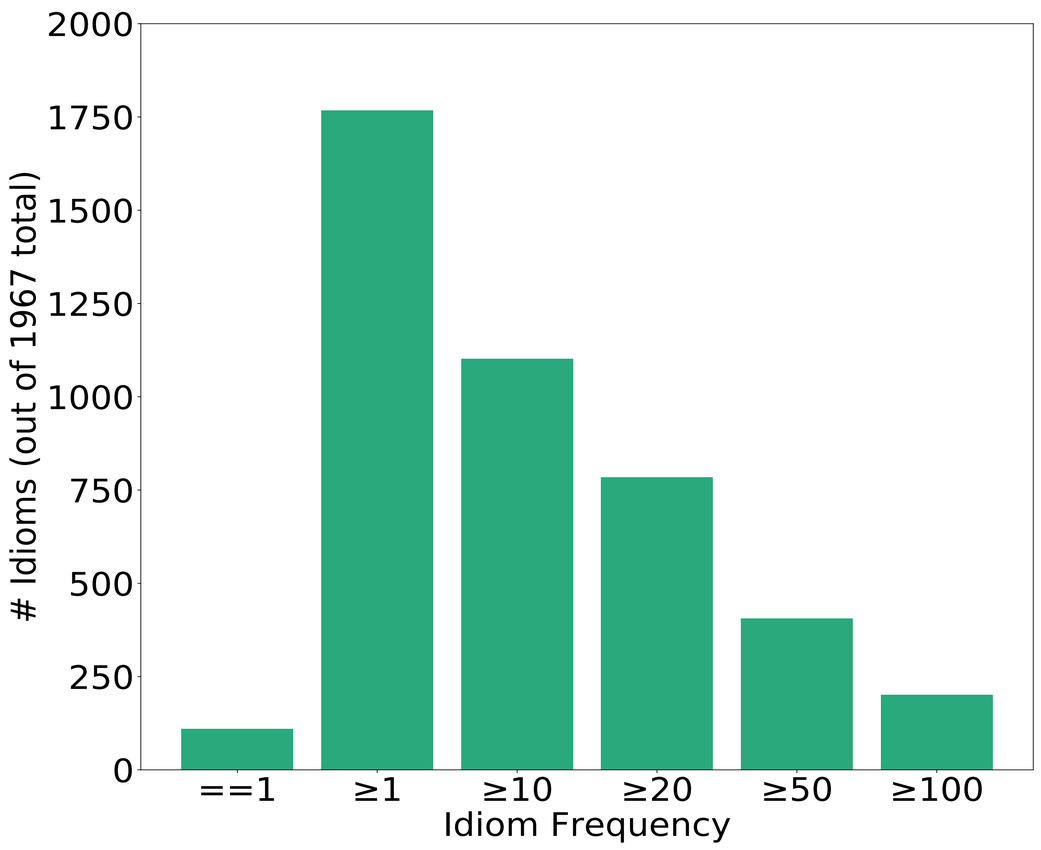
- $>$ 1000 types
- $\leq$ 200 instances per type
- Unrestricted syntactic patterns
- Enable fine-grained evaluation
- Facilitate deep learning approaches
- Focus on cross-idiom disambiguation

Setup

- Automatic pre-extraction of candidates
- Source: British National Corpus
- Dictionaries: Wiktionary, Oxford Dictionary of English Idioms, UsingEnglish.com
- Idioms: (Wiki $\cap$ Oxford) $\cup$ (UE.com $\cap$ 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*.