English isn't generic for language, despite what NLP papers might lead you to believe

Emily M. Bender - @emilymbender
University of Washington

Symposium on Data Science & Statistics
Bellevue, WA
May 30, 2019

The structure behind ‘unstructured’ data

- Natural language processing allows computers to access unstructured data expressed as speech or text

- Speech or text data does involve linguistic structure

- Linguistic structures vary depending on the language

- … and yet most NLP research looks only at English

Levels of linguistic structure, illustrated with ambiguity

- Phonetics & phonology (sounds): *It's hard to wreck a nice beach*.
- Morphology, the structure of words: *This safe is unlockable*.
- Syntax, the structure of sentences: *I saw the kid with a telescope.*
- Lexical semantics (word meaning): *The book about statistics is on the shelf*.
- Compositional semantics (sentence meaning): *Kim believes a unicorn is in the garden.*
- Speech acts: *Have you emptied the dishwasher?*

See Bender 2013,
Bender & Lascarides forthcoming

Languages of the world

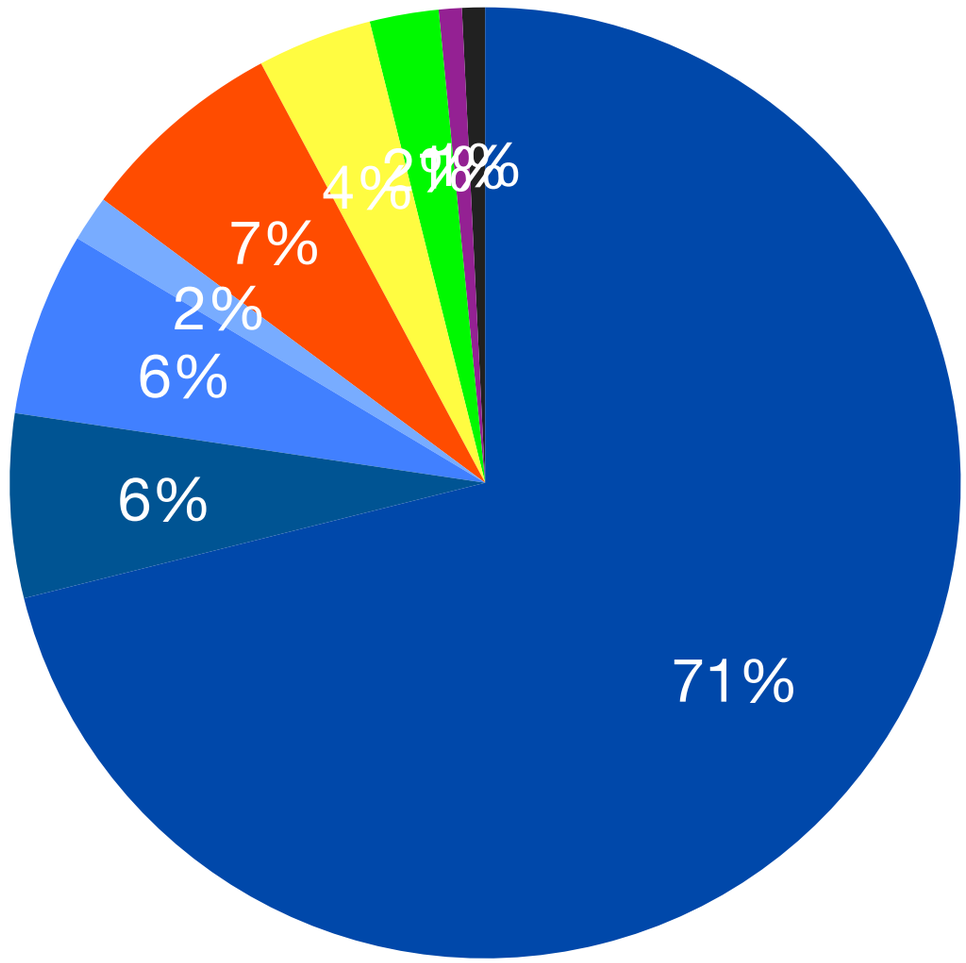
- 240 language families, according to glottolog.org
    - English belongs to Indo-European
- ~7000 languages in the world (ethnologue.com)
    - Most native speakers: Mandarin, Spanish, English, Hindi/Urdu, Arabic
    - Most total speakers: English, Mandarin, Hindi/Urdu, Spanish, French
    - Seattle's most common languages: English, Spanish, Arabic, Cantonese, Korean, Russian, Somali, Tagalog, Vietnamese (onecityproject.org)
    - Language of Seattle's indigenous people: Lushootseed

Languages of the world

- 240 language families, according to glottolog.org
    - English belongs to Indo-European
- ~7000 languages in the world (ethnologue.com)
    - Most native speakers: Mandarin, Spanish, English, Hindi/Urdu, Arabic
    - Most total speakers: English, Mandarin, Hindi/Urdu, Spanish, French
    - Seattle's most common languages: English, Spanish, Arabic, Cantonese, Korean, Russian, Somali, Tagalog, Vietnamese (onecityproject.org)
    - Language of Seattle's indigenous people: Lushootseed

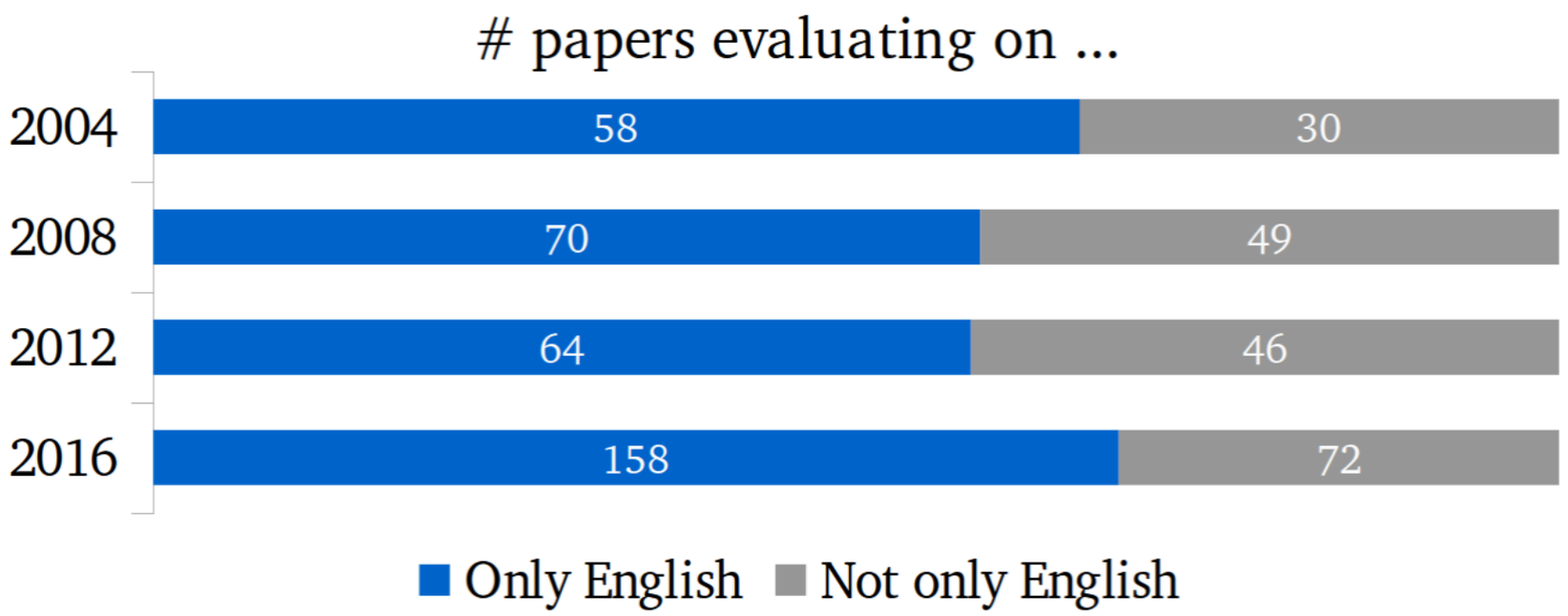
Languages of NLP: ACL 2008 (Bender 2009)

- Germanic
- Romance
- Semitic
- Japanese
- West Barkly
- Slavic
- Indic
- Chinese
- Turkish



English: 63%

Languages of NLP: ACL 2004-2016 (Mielke 2016)



Name that language (Bender 2011, 2018)

- EACL 2009: 33/45 English-only papers don't include the word "English"
- NAACL 2018: 42 tasks reported among 50 papers surveyed don't specify the language

Emily M. Bender
@emilymbender

Dear Computer Scientists,

"Natural Language" is *not* a synonym for "English".

That is all.
-Emily

9:32 AM - 26 Nov 2018

Emily M. Bender
@emilymbender

There’s about a dozen of us prepared to ask which language at #naacl2019 even if it’s obviously English because you didn’t say. Don’t want to get this question? State the language(s) you’re working with up front.

> **Emily M. Bender** @emilymbender
>
> Replying to @WWRob @_roryturnbull @BayesForDays
>
> At #naacl2019 should I ask at every talk/poster that doesn't specify their language, "What language(s) did you work with?"

1:30 PM - 18 May 2019

David

Replying to @emilymbender

Why is that?

10:02 PM - 19 May 2019

1

Why is that?

1

Emily M. Bender
@emilymbender

Replying to

Most work in #NLProc that is on English fails to say so, treating English as either the default language or as a proxy for all languages. Both are problematic:

6:51 AM - 20 May 2019

> Treating English as the unmarked, default, normative language plays into all sorts of unfair narratives about who counts as 'intelligent' etc, especially in Anglophone countries.

> Treating English as a proxy for all languages masks the specificity of English with its own linguistic idiosyncrasies. This is part of a larger pattern of NLP work ignoring the very linguistic systems it tries to handle.

Why does this matter?

- English isn’t a good representative language
- Mistaking form for meaning
- Exacerbating the digital divide

How is English non-representative?

- It's a spoken language, not a signed language
- It has a well-established, long-used roughly phone-based orthographic system
- … with white space between words
- … using (mostly) only lower-ascii characters
- It has relatively little morphology and thus fewer forms of each word
- It has relatively fixed word order
- English forms might 'accidentally' match database field names, ontology entries, etc.
- It has massive amounts of training data available (like the 3.3B tokens used to train BERT (Devlin et al 2019))

Mistaking form for meaning

- As fluent speakers, we often assume that the meaning of an utterance or text is right the words.
- Furthermore, using text-as-data often conflates the text with the world.
- But in fact there are at least four separate things (Bender & Lascarides forthcoming; Lascarides & Asher 2009, Hobbs 1979):

1. The form of the utterance
2. Its conventional meaning
3. The utterer's communicative intent
4. The relationship between that intent and the world

Mistaking form for meaning

- #1 and #2 here are language-specific:

    1.The form of the utterance

    2. Its conventional meaning

    3. The utterer’s communicative intent

    4. The relationship between that intent and the world

- By making at least the language we are working on visible, can we do better at keeping an eye on this articulated structure?

The digital divide

- Access to language technology is important for speakers:
    - Autocorrect, predictive text, voice prosthetics, internet search
- Access to language technology is important for languages:
    - Minority languages are already vulnerable to language shift (Fishman 1991)
    - Languages lacking technological support are used in fewer domains and perceived as less valuable
- Valuation of minority languages through technology is important for speakers (e.g. Lewis and Yang 2012)
- Even in Anglophone regions, not all stakeholders are English speakers/prefer speaking English

The digital divide

- If we don’t even acknowledge that we’re working (mostly) only on English, other languages get left in the dust

- If English gets to go unnamed, then work on other languages looks “language-specific” while work on English is “NLP”

- If we only value results on English, work on other languages isn’t incentivized

So what can we do?
(NLP researchers, reviewers, consumers)

- Value work on non-English languages
    - What languages has this been tested on? How well does it work for them?
- Always state the language being worked on up front
    - … and not just the name of the language either (=> Data Statements)
- Expect this information to be available, and demand it when it isn’t

Data Statements (Bender & Friedman 2018)

- A. Curation Rationale
- B. Language Variety
- C. Speaker Demographic
- D. Annotator Demographic
- E. Speech Situation
- F. Text Characteristics
- G. Recording Quality
- H. Other
- I. Provenance Appendix

(See also: Gebru et al 2018, AI Now Institute 2018, Yang et al 2018, Mitchell et al 2019)

Data statements

As consumers of datasets or products trained with them, NLP researchers, developers, and the general public would be well advised to use systems only if there is access to the information we propose should be included in data statements. *(Bender & Friedman 2018: 600)*

Conclusion

- Natural language isn’t just English, and NLP work should stop pretending that it is.

- If you’re a consumer of NLP tech (e.g. for text as data research), demand better

- This is a special (both senses!) case of: Always know your training data

References

AI Now Institute. (2018). *Algorithmic impact assessments: Toward accountable automation in public agencies.* Available from `https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde` (Medium.com)

Bender, E. M. (2011). On achieving and evaluating language independence in NLP. *Linguistic Issues in Language Technology*, *6*, 1–26.

Bender, E. M. (2013). *Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax*. Morgan & Claypool.

Bender, E. M. (2018). *How to make ends meet: Why general purpose NLU needs linguistics.* (Talk presented at the Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP (RELNLP) at ACL 2018, Melbourne, Australia, July 19, 2018)

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. Available from `https://doi.org/10.1162/tacl_a_00041`

Bender, E. M., & Lascarides, A. (forthcoming). *Linguistic fundamentals for natural language processing: 100 essentials from semantics and pragmatics*. Morgan & Claypool.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Naacl 2019.*

Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages* (Vol. 76). Multilingual matters.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., et al. (2018). *Datasheets for datasets.* (arXiv:1803.09010v1)

Hobbs, J. (1979). Coherence and coreference. *Cognitive Science*, *3*(1), 67–90.

Lascarides, A., & Asher, N. (2009). Agreement, disputes and commitment in dialogue. *Journal of Semantics*, *26*(2), 109–158.

Lee, J. S. (2006). Exploring the relationship between electronic literacy and heritage language maintenance. *Language Learning & Technology*, *10*(2), 93–113.

Lewis, W. D., & Yang, P. (2012). Building mt for a severely under-resourced language: White hmong. In *Association for machine translation in the americas.*

Mielke, S. J. (2016). *Language diversity in ACL 2004 - 2016.* (Blog post, `https://sjmielke.com/acl-language-diversity.htm`)

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220–229).

Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H. V., & Miklau, G. (2018). A nutritional label for rankings. In *Proceedings of the 2018 international conference on management of data* (pp. 1773–1776). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/3183713.3193568`