The Creation of a Corpus of English Metalanguage

Shomir Wilson
Carnegie Mellon (current affiliation)
University of Maryland (research performed)

ACL 2012 – 10 July 2012

Speaking or Writing About Language: Observations

When we write or speak *about language* (to discuss words, phrases, syntax, meaning…):

- We often convey very direct, salient information about language.
- We tend to be instructive, and we (often) try to be easily understood.
- We do this to clarify the meaning of words or phrases we (or our audience) use.

Examples

1) This is sometimes called **tough love**.
2) I wrote "**meet outside**" on the chalkboard.
3) **Has** is a conjugation of the verb **have**.
4) The button labeled **go** was illuminated.
5) That bus, was its name **61C**?
6) **Mississippi** is fun to spell.
7) He said, "**Dinner is served**."

And Yet…

Metalanguage (sometimes described as *self-referential language*, or the "mention" part of the use-mention distinction) should be fertile ground for language technologies. However:

- Metalinguistic constructions have atypical properties.
- Metalanguage defies trends in language (e.g., in syntax, word senses, topicality) that language technologies usually exploit.

What Goes Wrong

Dialog System: Where do you wish to depart from?

User: Arlington.

Dialog System: Departing from Allegheny West. Is this right?

User: No, I said "Arlington".

Dialog System: Please say where you are leaving from.

```
(ROOT
  (S
    (NP
      (NP (DT The) (NN button))
      (VP (VBN labeled)
        (S
          (VP (VB go))))
    (VP (VBD was)
      (VP (VBN illuminated)))
    (. .)))
```

The word "bank" can refer to many things.

bank: n|1| a financial institution that accepts deposits and channels the money into lending activities

Dialog System: Let's Go! (Carnegie Mellon University)

Parser: Stanford Parser (Stanford University)

Word Sense Disambiguation: IMS (National University of Singapore)

Metalanguage and Mentioned Language

The goal of this project was to provide a basis for the study of metalanguage (i.e., *language about language*) in English.

A better understanding of metalanguage will enable us to construct language technologies that (at worst) can cope with it and (at best) exploit the information it conveys.

To make the problem tractable, the focus was on *mentioned language*: instances of metalanguage that can be explicitly delimited within a sentence.

Previous Efforts

- Two proof-of-concept corpora preceded this one:
  - A “pilot corpus” established that Wikipedia was a fertile source of mentioned language [1].
  - A “combined cues corpus” validated the combination of lexical and stylistic cues to gather candidate instances [2].
- Anderson, et al. [3] gathered a metalanguage corpus of human dialog—but it lacked word- or phrase-level annotations and contained substantial noise.
- Many have discussed mentioned language or metalanguage in purely theoretical terms (Saka, Cappelen, Lepore, Maier, Geach, Partee, et al.).

[1] Shomir Wilson. "Distinguishing use and mention in natural language". NAACL HLT Student Research Workshop, 2010
[2] Shomir Wilson "In search of the use-mention distinction and its impact on language processing tasks". CICLING 2011
[3] Anderson, ML, Yoshi A Okamoto, Darsana Josyula, and Donald Perlis. "The Use-Mention Distinction and its Importance to HCI." EDILOLG 2002

Mentioned Language: A Definition

The following definition was used in previous efforts to build pilot corpora of mentioned language:

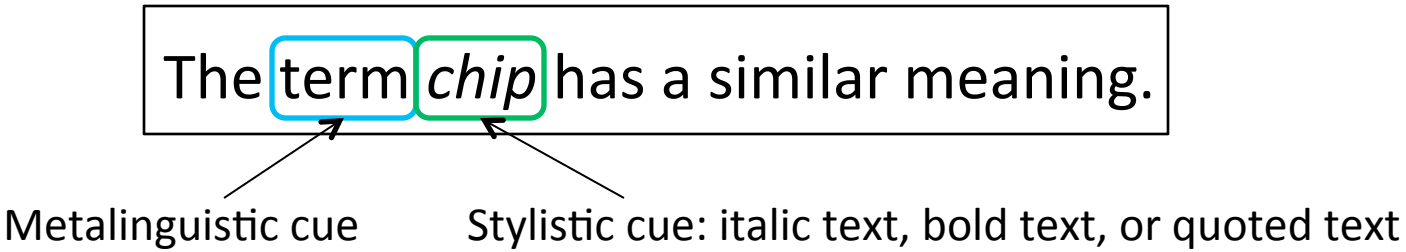
For T a token or a set of tokens in a sentence, if T is produced to draw attention to a property of the token T or the type of T, then T is an instance of mentioned language.

Example: *"The cat is on the mat" is a sentence.*

New in the present effort: an equivalent substitution-based "labeling rubric" was used to produce consistent results. The rubric appears in the paper.

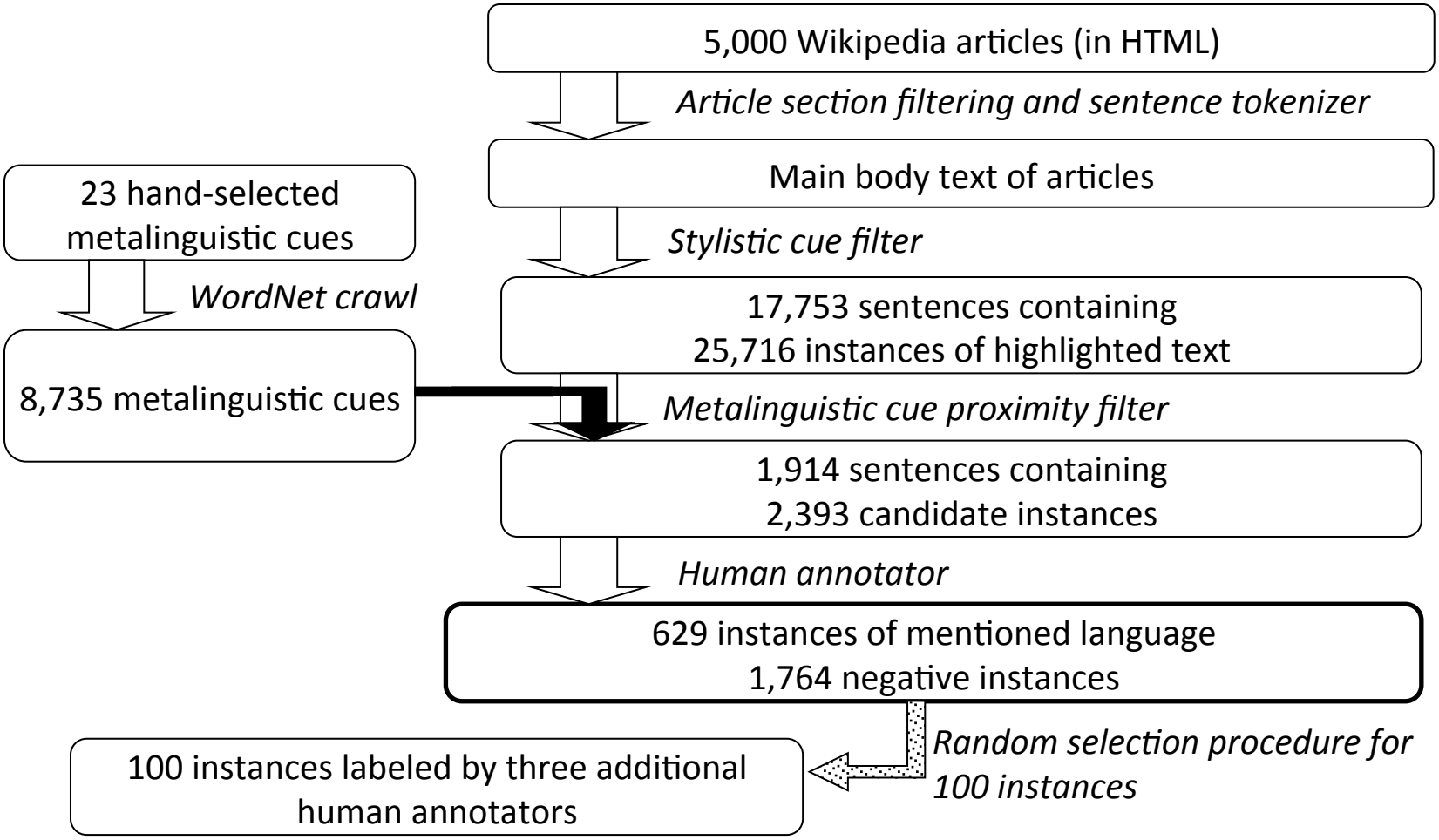
Corpus Creation: Overview

- A randomly subset of English Wikipedia articles was chosen as a text source.
- To make human annotation tractable: sentences were examined only if they fit a combination of cues:



- Mechanical Turk did not work well for labeling.
- Candidate instances were labeled by a human annotator. A subset were labeled by multiple annotators to verify the reliability of the corpus.

Collection and Filtering



Corpus Composition: Frequent Leading and Trailing Words

These were the most common words to appear in the three words before and after instances of mentioned language.

Before Instances

Rank	Word	Freq.	Precision (%)
1	call (v)	92	80
2	word (n)	68	95.8
3	term (n)	60	95.2
4	name (n)	31	67.4
5	use (v)	17	70.8
6	know (v)	15	88.2
7	also (rb)	13	59.1
8	name (v)	11	100
9	sometimes (rb)	9	81.9
10	Latin (n)	9	69.2

After Instances

Rank	Word	Freq.	Precision (%)
1	mean (v)	31	83.4
2	name (n)	24	63.2
3	use (v)	11	55
4	meaning (n)	8	57.1
5	derive (v)	8	80
6	refers (n)	7	87.5
7	describe (v)	6	60
8	refer (v)	6	54.5
9	word (n)	6	50
10	may (md)	5	62.5

Corpus Composition: Categories

Categories were observed through application of the substitution rubric.

Category	Freq.	Example
Words as Words (WW)	438	The IP Multimedia Subsystem architecture uses the term transport plane to describe a function roughly equivalent to the routing control plane.
		The material was a heavy canvas known as duck, and the brothers began making work pants and shirts out of the strong material.
Names as Names (NN)	117	Digeri is the name of a Thracian tribe mentioned by Pliny the Elder, in The Natural History.
		Hazrat Syed Jalaluddin Bukhari's descendants are also called Naqvi al-Bukhari.
Spelling and Pronunciation (SP)	48	The French changed the spelling to bataillon, whereupon it directly entered into German.
		Welles insisted on pronouncing the word apostles with a hard t.
Other Mentioned Language (OM)	26	He kneels over Fil, and seeing that his eyes are open whispers: brother.
		During Christmas 1941, she typed The end on the last page of Laura.
[Not Mentioned Language (XX)]	1,764	NCR was the first U.S. publication to write about the clergy sex abuse scandal.
		Many Croats reacted by expelling all words in the Croatian language that had, in their minds, even distant Serbian origin.

Inter-Annotator Agreement

Three additional expert annotators labeled 100 instances selected randomly with quotas from each category.

Code	Frequency	K
WW	17	0.38
NN	17	0.72
SP	16	0.66
OM	4	0.09
XX	46	0.74

For mention vs. non-mention labeling, the kappa statistic was 0.74. Kappa between the primary annotator and the "majority voter" of the rest was 0.90.

These statistics suggest that mentioned language can be labeled fairly consistently—but the categories are fluid.

Discussion

- A core set of metalinguistic cues appeared frequently, followed by a long, thin tail.
  - A core metalinguistic vocabulary seems to exist.
  - The most popular metalinguistic cues were highly correlated with mentioned language.
- Recurring patterns were observed in how metalinguistic cues related to mentioned language.
  - Noun apposition often occurs between a cue noun and mentioned language:

    "Sometimes the term *scramble crossing* is used."
  - Mentioned language tends to appear in appropriate semantic roles for cue verbs:

    "This precipitation is called *sleet*."

Future Directions

- The corpus is online (URL on next slide) and available for use under a CC BY-SA 3.0 license.
- Next: automatic detection of mentioned language. It appears to be feasible.
- Potential applications to language technologies:
  - Dialog systems
  - Language instruction
  - Dictionary building tools
  - Source attribution
  - Automated typesetting and copyediting

Questions?

Shomir Wilson – shomir@cs.cmu.edu
http://www.cs.cmu.edu/~shomir

The corpus is available at:
http://www.cs.cmu.edu/~shomir/um_corpus.html

Special thanks to:
Don Perlis
Tim Oates

Appendix

The Rubric

Rubric

Given S a sentence and X a copy of a linguistic entity in S:

1) Create X': the phrase "that [item]", where [item] is the appropriate term for linguistic entity X in the context of S.
2) Create S': copy S and replace the occurrence of X with X'.
3) (3) Create W: the set of truth conditions of S.
4) (4) Create W': the set of truth conditions of S', assuming that X' in S' is understood to refer deictically to X.
5) (5) Compare W and W'. If they are equal, X is mentioned language in S. Else, X is not mentioned language in S.

Positive Example

S: *Spain is the name of a European country.*

X: *Spain*.

1) X': *that name*
2) S': *That name is the name of a European country.*
3) W: Stated briefly, *Spain* is the name of a European country.
4) W': Stated briefly, *Spain* is the name of a European country.
5) W and W' are equal. *Spain* is mentioned language in S.

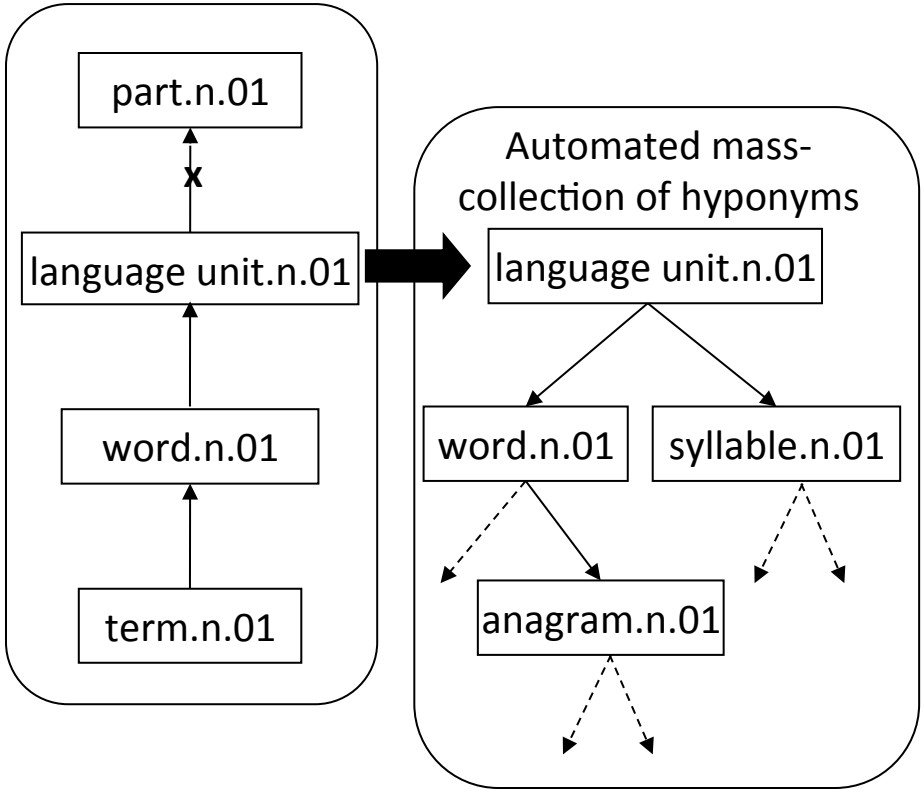
Negative Example

S: *Spain is a European country*.

X: *Spain.*

1) X': *that name*
2) S': *That name is a European country.*
3) W: Stated briefly, Spain is a European country.
4) W': Stated briefly, the name *Spain* is a European country.
5) W and W' are not equal. *Spain* is not mentioned language in S.

"Mention Word" Collection via WordNet



Seed mention word: "term"

For each of 23 mention words from the previous ("combined cues") corpus:

1) A human annotator found its most general linguistically-significant hypernym
2) All descendants of the hypernym were gathered

Corpus Composition: Cumulative Coverage of Top Words