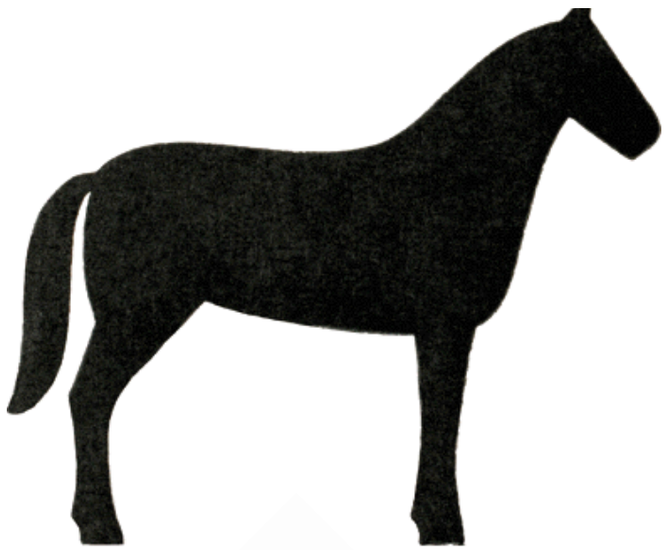
An empirically grounded expansion of the supersense inventory

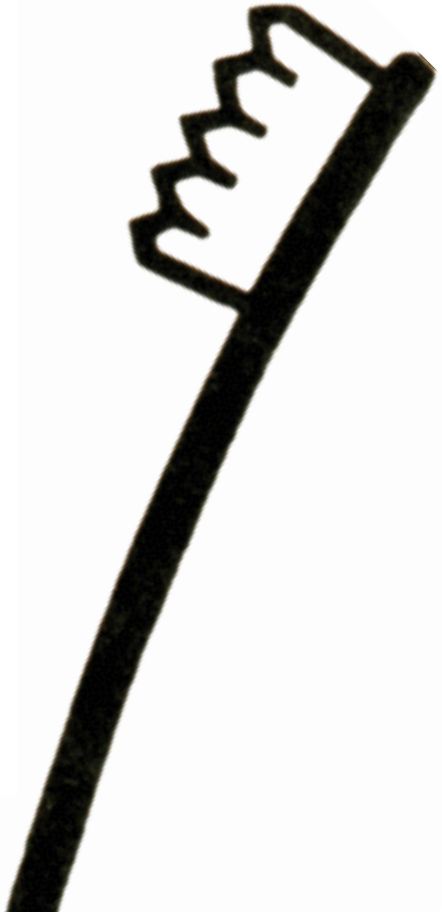
Héctor Martínez Alonso, Anders Johannsen,Sanni Nimb†,
Sussi Olsen, Bolette Sandford Pedersen
University of Copenhagen (Denmark)
Danish Society of Language and Literature (Denmark) †

alonso@hum.ku.dk, bspedersen@hum.ku.dk

Many words; many senses.

Lexicographer files





Supersense inventory (SSI)

The set of coarse target senses that supersense tagging aims to predict.

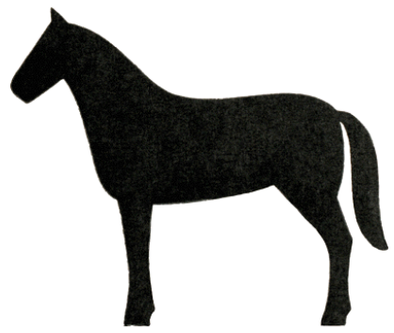
These senses are normally taken from WordNet lexicographer file names.

Supersense tagging

"Beatrix Potter had always loved drawing those sweet bunnies"

Supersense tagging (SST)





B	I			B	B			B
Beatrix	Potter	had	always	loved	drawing	those	sweet	bunnies

But…

Lexicographer files were devised to organize synsets.

What are their shortcomings regarding SST?

And…

What about adjectives?

What this work is not

1. A proposal for a new canonical supersense inventory (SSI).
2. A reorganization of lexicographer files.

What this work *is*

A methodological outline on how to expand the SSI

1. to improve its usefulness for a certain corpus choice
2. while keeping the SSI backward-compatible

A bit of background

1. We wanted to deploy Danish SST based on DanNet.

A bit of background

1. We wanted to deploy Danish SST based on DanNet.
2. DanNet has no immediate synset-supersense links.

A bit of background

1. We wanted to deploy Danish SST based on DanNet.
2. DanNet has no immediate synset-supersense links.
3. However, it has EuroWordNet ontological types.

A bit of background

1. We wanted to deploy Danish SST based on DanNet.
2. DanNet has no immediate synset-supersense links.
3. However, it has EuroWordNet ontological types.
4. We matched the ontological the existing SSI.

A bit of background

1. We wanted to deploy Danish SST based on DanNet.
2. DanNet has no immediate synset-supersense links.
3. However, it has EuroWordNet ontological types.
4. We matched the ontological the existing SSI.
5. When some mismatches appeared with enough frequency, we considered suggesting a new supersense.

A bit of background

1. We wanted to deploy Danish SST based on DanNet.
2. DanNet has no immediate synset-supersense links.
3. However, it has EuroWordNet ontological types.
4. We matched the ontological the existing SSI.
5. When some mismatches appeared with enough frequency, we considered suggesting a new supersense.
6. This work describes how we evaluate whether a new supersense is worth incorporating into the SSI.

Mapping ontological types

Ontological type	Supersense
Property+Physical+Colour	ADJ.PHYSICAL
Liquid+Natural	NOUN.SUBSTANCE
Dynamic+Agentive+Mental	VERB.COGNITION

Corpus

We use the Danish Clarin corpus:
newswire, blogs, chatrooms, magazines,
parliamentary speeches

plus the test section of the Danish Dependency treebank:
more newswire, some literature

3 1/2 inclusion criteria

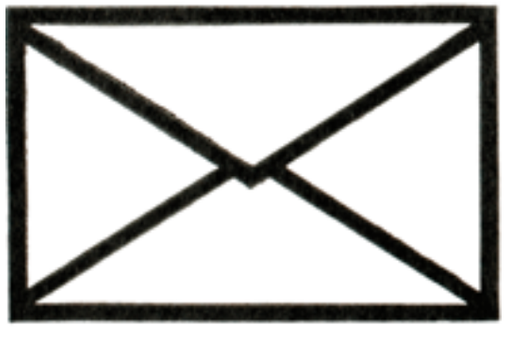
1. Agreement
2. Frequency
3. Association
4. Entity

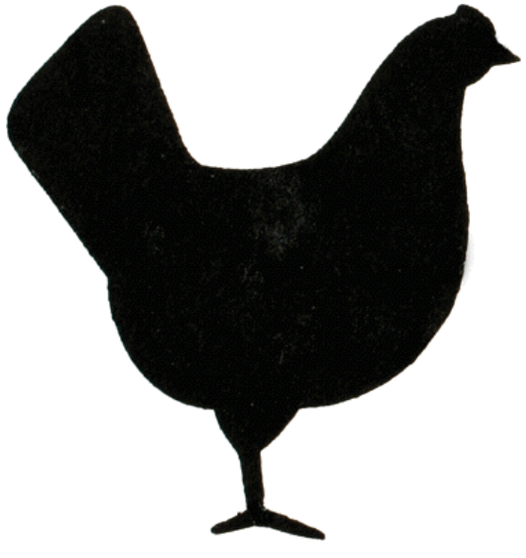
And a second step

1. Agreement
2. Frequency
3. Association
4. Entity
5. Post-annotation analysis

Nouns

Canonical SSI: 27 classes
Extension: 7 classes (out of 9 suggested)

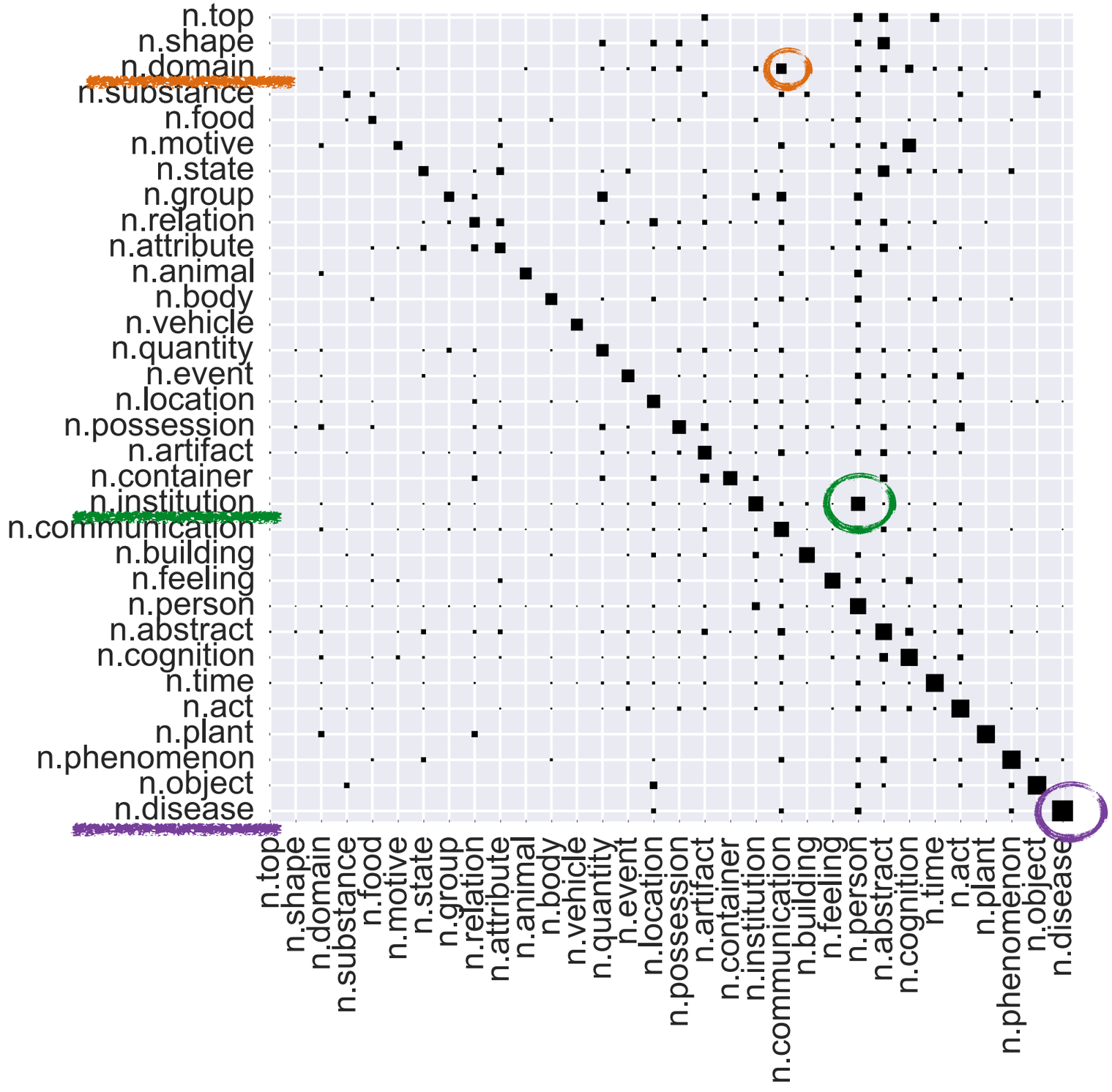




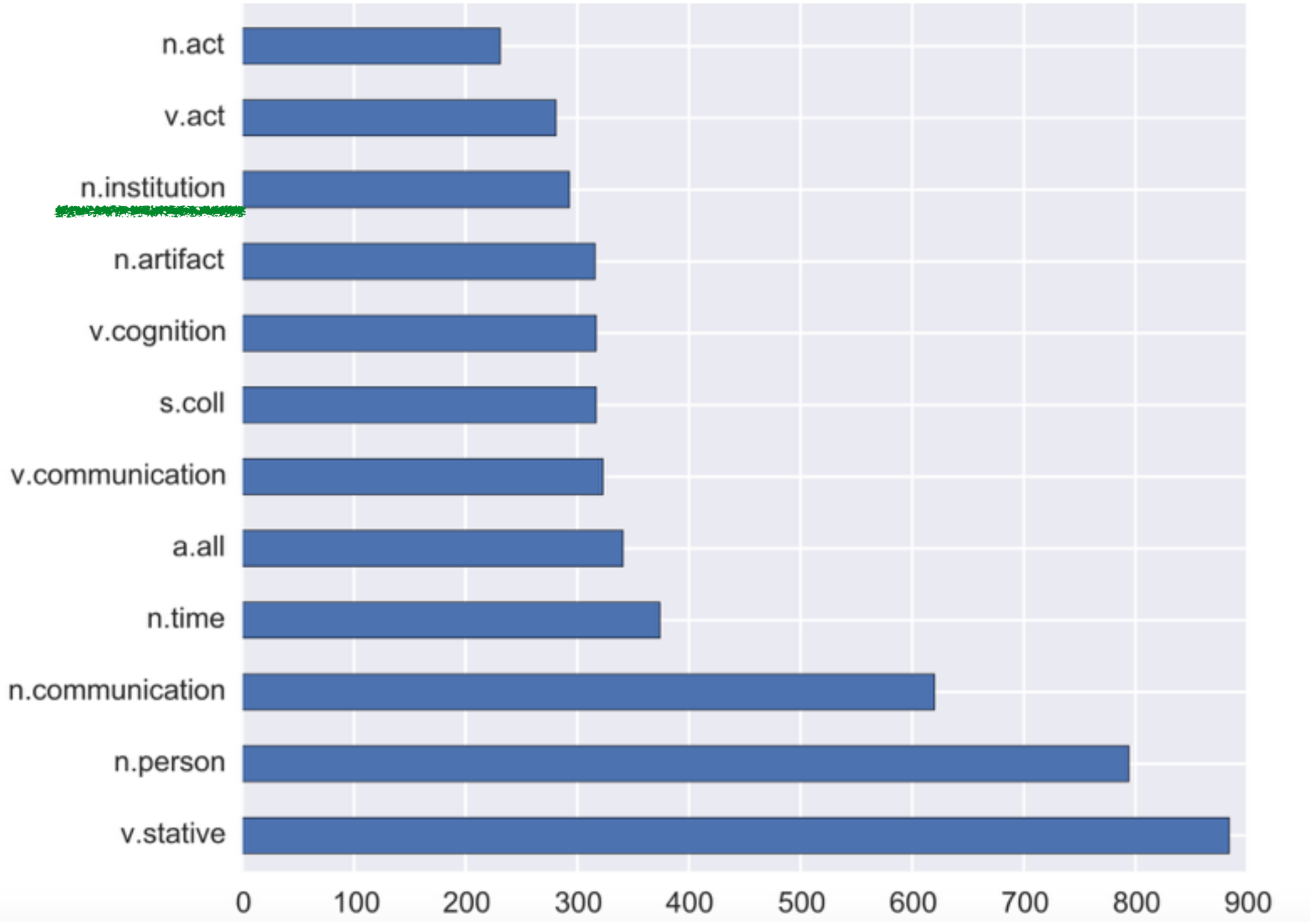
Nouns

Noun	
VEHICLE	ARTIFACT
BUILDING	ARTIFACT
CONTAINER	ARTIFACT
DOMAIN	COGNITION
ABSTRACT	COGNITION
INSTITUTION	GROUP
DISEASE	STATE
LANGUAGE	COMMUNICATION
DOCUMENT	COMMUNICATION

Criterion I: Agreement



Criterion II: Frequency



Dictionary ≠ Encyclopedia

DanNet has no coverage for named entities,
which are pervasive when annotating some NLP-
typical text types like newswire.

Dictionary ≠ Encyclopedia

DanNet has no coverage for named entities, which are pervasive when annotating some NLP-typical text types like newswire.

The a priori sense distribution given by DanNet underestimates the frequency of e.g. noun.person or noun.location

Criterion III: Association

Danish (extended)	
v.consumption	n.food
v.contact	n.body
n.food	**n.container†**
v.body	n.body
n.disease†	n.body
v.competition	n.event
v.motion	v.contact
v.contact	n.artifact
n.substance	n.object
n.shape	n.body
n.vehicle†	n.substance

Criterion IV: Entity

noun.group

noun.institution

Organization

Criterion IV: Entity

noun.artifact

noun.building

Location

Nouns: Evaluation

New supersense	Agr.	Freq.	Assc.	NER
ABSTRACT	x	x		
BUILDING	x			x
CONTAINER	x		x	
DISEASE	x			x
DOMAIN				
INSTITUTION	x	x		x
VEHICLE	x		x	
LANGUAGE	–			
DOCUMENT	–	x		x

Verbs

Canonical SSI: 15 classes
Extension: 2 classes (plus verb satellite tags)




Verbs

Verb

ASPECTUAL } STATIVE

PHENOMENON } CHANGE

Verbs

Verb

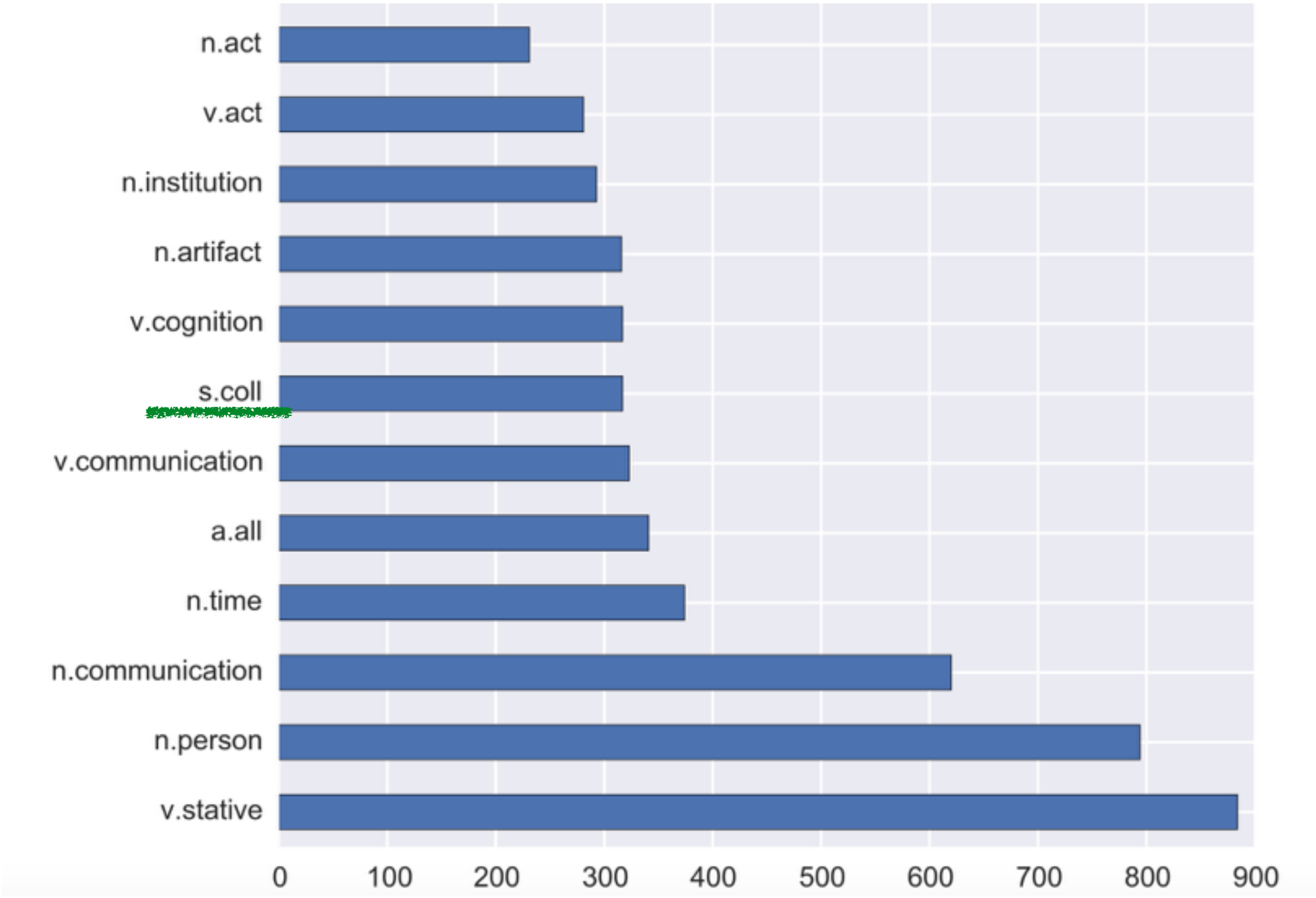
ASPECTUAL } STATIVE

PHENOMENON } CHANGE

Satellite

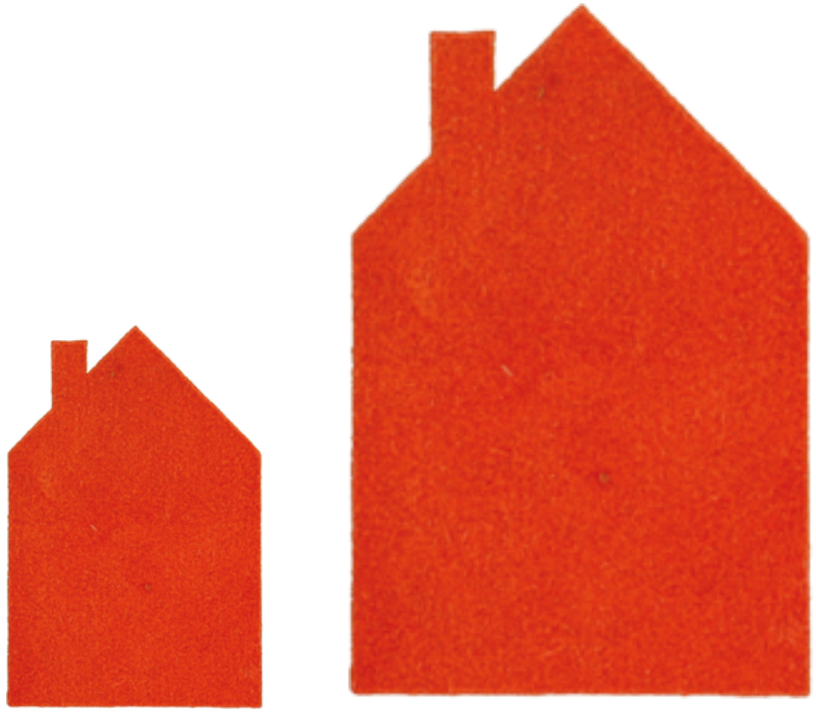
COLLOCATION
PARTICLE
REFLPRON
} *none*

Criterion II: Frequency



Adjectives

Canonical SSI: 0-3 classes
Extension: 5 classes



Adjectives

Adjective

MENTAL

PHYSICAL

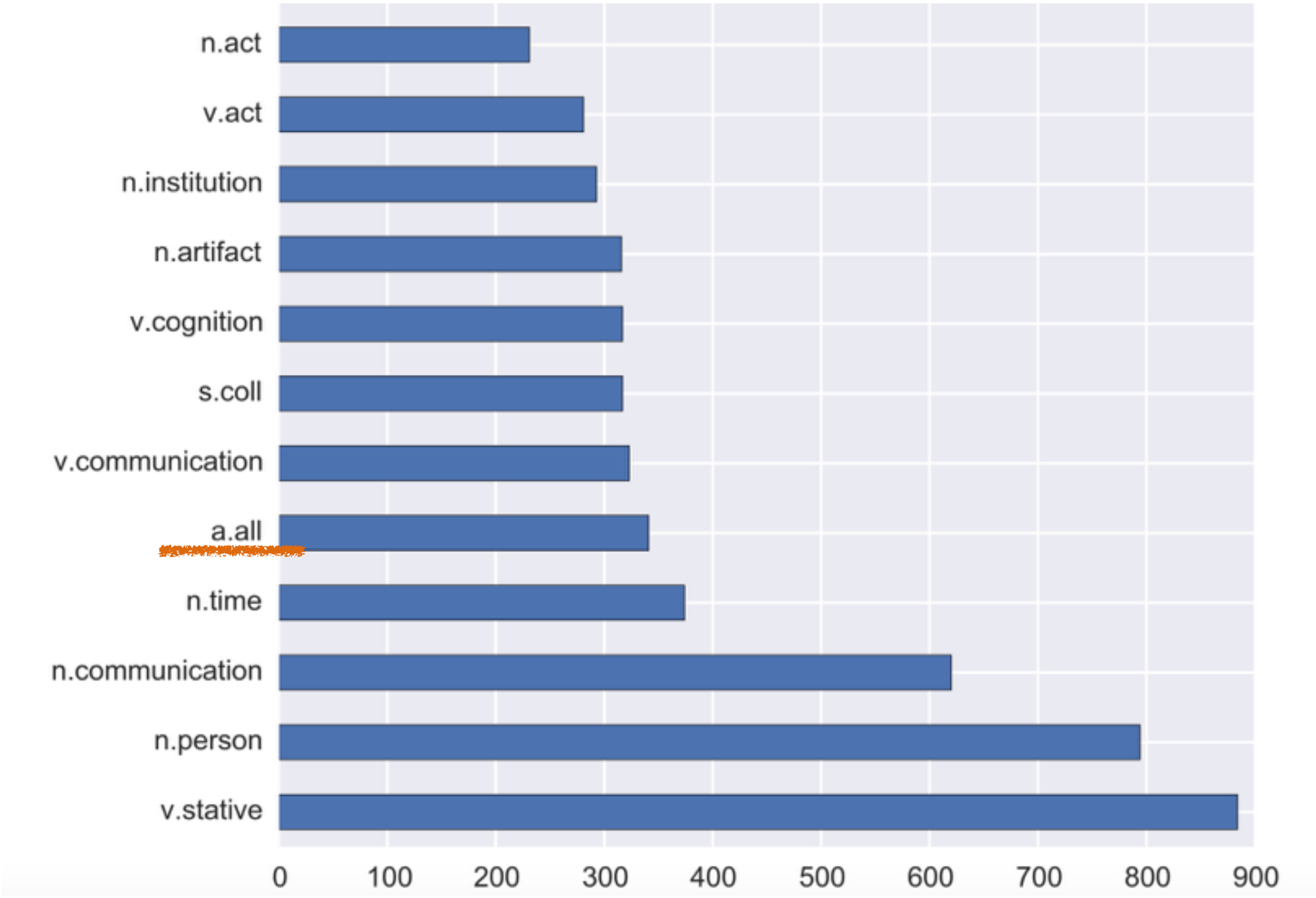
SOCIAL

TIME

FUNCTION

} ALL

Criterion II: Frequency



Summary

1. EWN ontological type + POS → Supersense
2. 3 1/2 Criteria: Agreement, Freq. , Association, Entity
3. This method is corpus dependent, but yields
    1. n.institution and n.document as robust candidates
    2. v.aspect and v.phenomenon
    3. four adjective classes plus a.function which greatly reduces the size of a.all

Thanks! Questions?

All the annotated data, and supersense conversions are available under

`https://github.com/coastalcph/semdax`

Cf. also Olsen (2015) for annotation task, Martínez Alonso et al (2015a) and (2015b) for SST, and Pedersen (2016, to appear) for corpora.

The research leading to these results has been funded by Danish Free Research Council under the project *Semantic Parsing Across Domains*.