CLUSTER OF EXCELLENCE

Universität des Saarlandes

Institut für
Maschinelle
Sprachverarbeitung

Universität Stuttgart

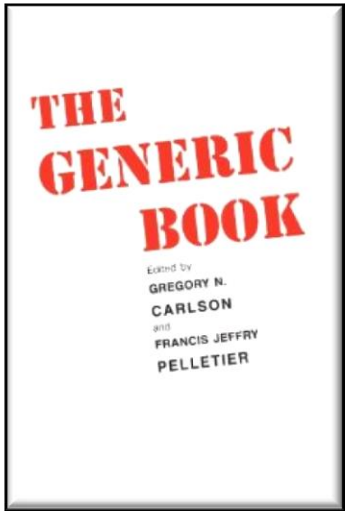
Annotating genericity: a survey, a scheme, and a corpus

Annemarie Friedrich[1], Alexis Palmer[2],
Melissa Peate Sørensen[1] and Manfred Pinkal[1]

[1]Computational Linguistics, Universität des Saarlandes
[2]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

June 2015. 9th Linguistic Annotation Workshop (LAW IX). Denver, Colorado.

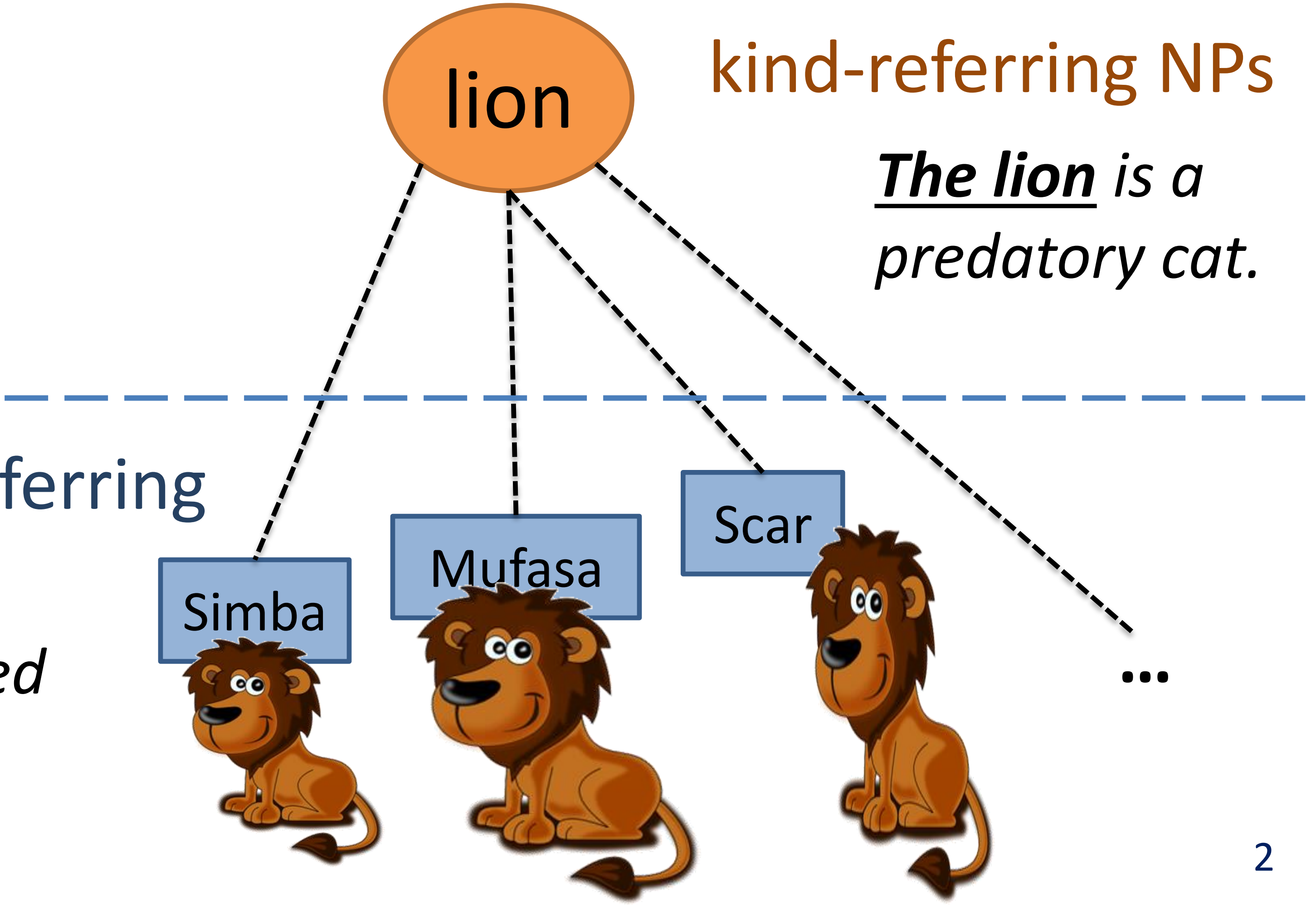
Reference to kinds



THE GENERIC BOOK

Edited by GREGORY N. CARLSON and FRANCIS JEFFRY PELLETIER

Krifka et al. (1995): Genericity: An Introduction.



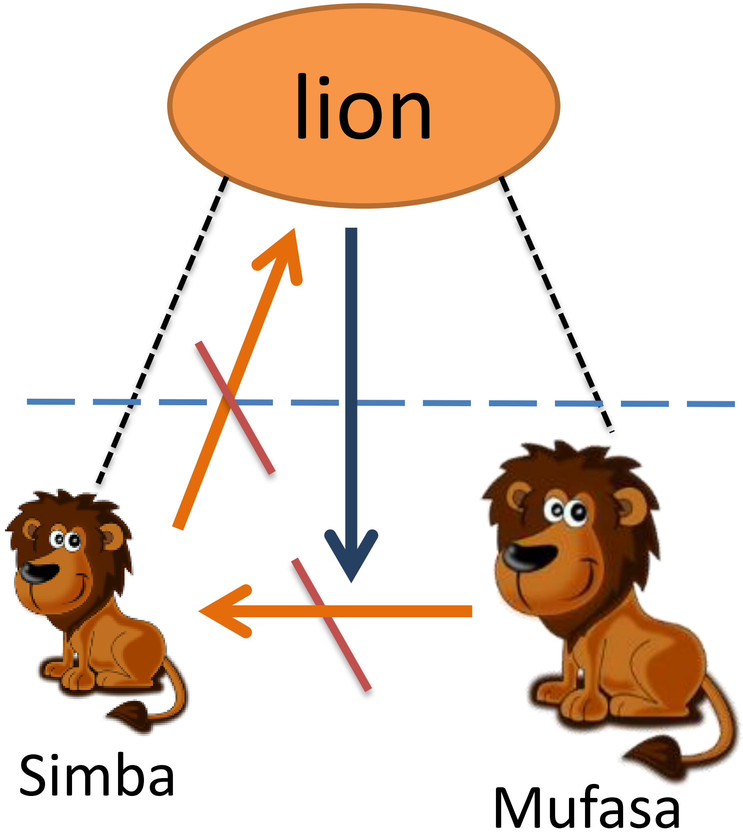
kind-referring NPs

<u>The lion</u> *is a predatory cat.*

non-kind-referring NPs

<u>Simba</u> *escaped from the zoo.*

Generic vs. non-generic expressions

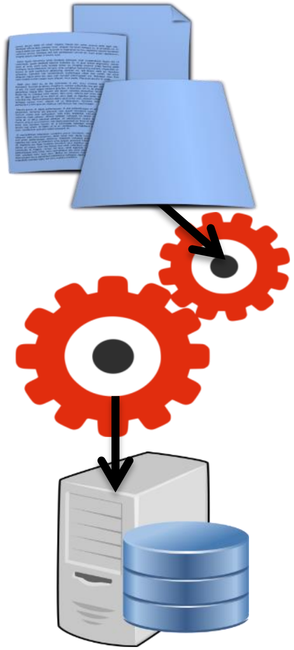


different
entailment properties

Lions are dangerous.

Mufasa is dangerous.
Simba is dangerous.

Identifying generic expressions: why?



knowledge extraction from text

natural language understanding

Motivation

Survey

existing approaches
semantic theory literature

Previously existing corpora

problematic points in annotation guidelines or small data sets

Aim:

computational models for identifying **generic expressions**

Our corpus / annotation scheme

guidelines motivated by semantic theory
large data set
→necessary for successful machine learning approaches to genericity identification

NP-level: reference to kinds

form of NP not sufficient

	kind-referring	**non-kind-referring**
definite NPs	*The lion is a predatory cat.*	*The cat chased the mouse.*
indefinite NPs	*Lions eat meat.*	*Dogs were barking outside.*
quantified NPs	*Some (type of) dinosaur is extinct.*	*Some dogs were barking outside.*
proper names	*Panthera leo persica was first described by the Austrian zoologist Meyer.*	*John likes ice cream.*

clause / context matters

Terminology: clause-level genericity

characterizing sentences

	lexically characterizing sentences	**habitual sentences**
kind-referring subject	*Lions have manes.*	*Lions eat meat.*
non-kind-referring subject	*John is tall.*	*John drives to work.*

Survey: annotating genericity

Level	Corpus	Scheme	Size
NP	**ACE-2**	generic, specific	40K entity mentions
	ACE-2005	GEN, SPC, USP, NEG	40K entity mentions
	ECB+	GEN, non-GEN	12.5K entity mentions
	GNOME	generic-yes, generic-no	900 clauses
	Herbelot & Copestake	ONE, SOME, MOST, ALL, QUANT	300 subject mentions
	CFD (Bhatia et al.)	GENERIC_KIND, GENERIC_INDIVIDUAL	3422 NPs (131 generic)
clause	**Mathew & Katz**	habitual, episodic	1052 sentences
	Louis & Nenkova	general, specific	894 sentences
NP, clause	**MASC** **WikiGenerics**	GEN_gen, NON-GEN_gen, NON-GEN_non-gen	20k clauses 10k clauses

Survey: clause-level annotations

[Mathew & Katz 2009]

episodic *John has finished the cake*.

vs. **habitual** *John drives to work. (regularity)*

[Louis & Nenkova 2011]

general sentences vs. **specific** sentences

≠ genericity as treated in literature

“broad statements about a topic”

A handful of serious attempts have been made to eliminate diseases.

vs. “detailed information”

Solid silicon compounds are already familiar – as rocks, glass, …

Survey: NP-level annotations

[Nedoluzhko 2013]

coreference resolution research

no consistent definition

ignore generic entity mentions? avoid mixed chains?

GNOME corpus: generic-yes, generic-no

[Herbelot & Copestake 2009/2011]

Cats are mammals.	ALL cats
Cats have four legs.	MOST cats
Cats are black.	SOME cats
A cat chased the mouse.	ONE cat

[Bhatia et al. 2014]

GENERIC_KIND_LEVEL	*Dinosaurs are extinct.*
GENERIC_INDIVIDUAL_LEVEL	*Cats have fur.*

ACE entity class annotations

Automatic Content Extraction (2002-2008)

- largest corpora annotated with NP-level genericity to date
- basis for computational modeling [Reiter & Frank 2010]

ACE-2005:

GEN kind-referring

SPC non-kind-referring

NEG negatively quantified NPs

There are no confirmed suspects yet.

USP underspecified

ambiguous cases

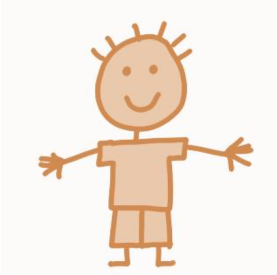
There are new opportunities for women in New Delhi.

and mentions of entities “whose identity would be difficult to locate”

Officials reported …

ACE-2005: agreement study

533 documents



adjudication

final corpus

news, broadcast news,
broadcast conversation,
forum and weblog texts

annotations available from LDC

agreement study:

exactly-matching entity mention spans (~90%)

		annotator 2			
		SPC	**USP**	**GEN**	**NEG**
annotator 1	**SPC**	28168	1575	684	3
	USP	1142	1954	963	2
	GEN	757	1261	1707	10
	NEG	8	5	7	71

Cohen's κ = 0.53

confusion of SPC/GEN with USP is high

ACE-2005: agreement study

Problems of the ACE annotation guidelines

- **predicative uses** are marked
  - *John is a nice person. (specific)*
  - *John seems to be a nice person. (generic)*
- **noun modifiers** in compounds (9.5% of all mentions) are marked as generic: *subway system*

non-referential

- guidelines mix **genericity** and **specificity**

  (specificity = speaker has a particular referent in mind)
  - *Officials reported...*
  - not underspecified: not generic, but nonspecific

Our approach: motivation

Previous approaches:

range / mix of linguistic phenomena, focus on applications

many linguistically motivated schemes, but small corpora

Our approach:

motivated by **semantic theory** (Krifka et al. 1995)

study references to and statement about kinds

(Task NP, Task Cl, Task Cl+NP)

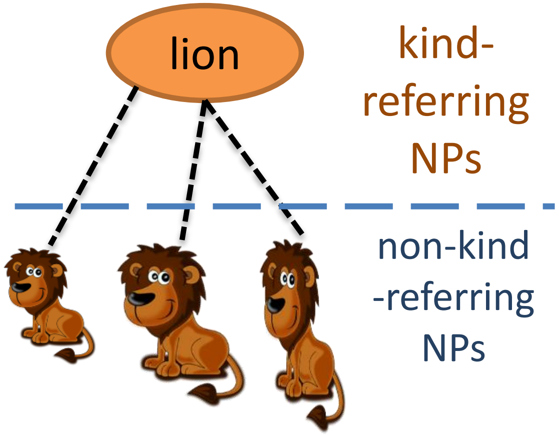
(other aspects of genericity → future work)

contribution of clauses to **discourse**:

characterizing statements ≠ particular events or states

→ relevant for processing temporal structure of discourse

Task NP: genericity of subject



generic: references to kind / class

The lion is a predatory cat.

Lions have manes.

A lion may eat up to 30kg in one sitting.

non-generic: references particular individual(s)

Simba flees into exile.

A lion must have eaten the rabbit. (nonspecific)

Task Cl: genericity of clause

generic: characterizing statements about kinds

subject must be **generic**.

The lion is a predatory cat.

Lions eat up to 30kg in one sitting. (habitual)

non-generic: statements about particular individuals or particular events.

John is a nice guy.

John cycles to work. (habitual)

Task Cl+NP: clause and subject

clause subject	generic	non-generic
generic	*Lions have manes.* *Lions eat meat.*	*The blobfish was voted the "World's Ugliest Animal".* *Dinosaurs died out.*
non-generic	-- --	*John is a nice guy.* *John cycles to work.*

Corpus data

Manually Annotated Subcorpus of the Open American National Corpus (**MASC**)

- essays, travel, letters, journal, jokes, blog, news, fiction
- 20136 clauses

102 Wikipedia texts (**WikiGenerics**)

- aim: balanced corpus (many generics)
- about animals, sports, politics, science, biographies, …
- 10279 clauses

Annotation process

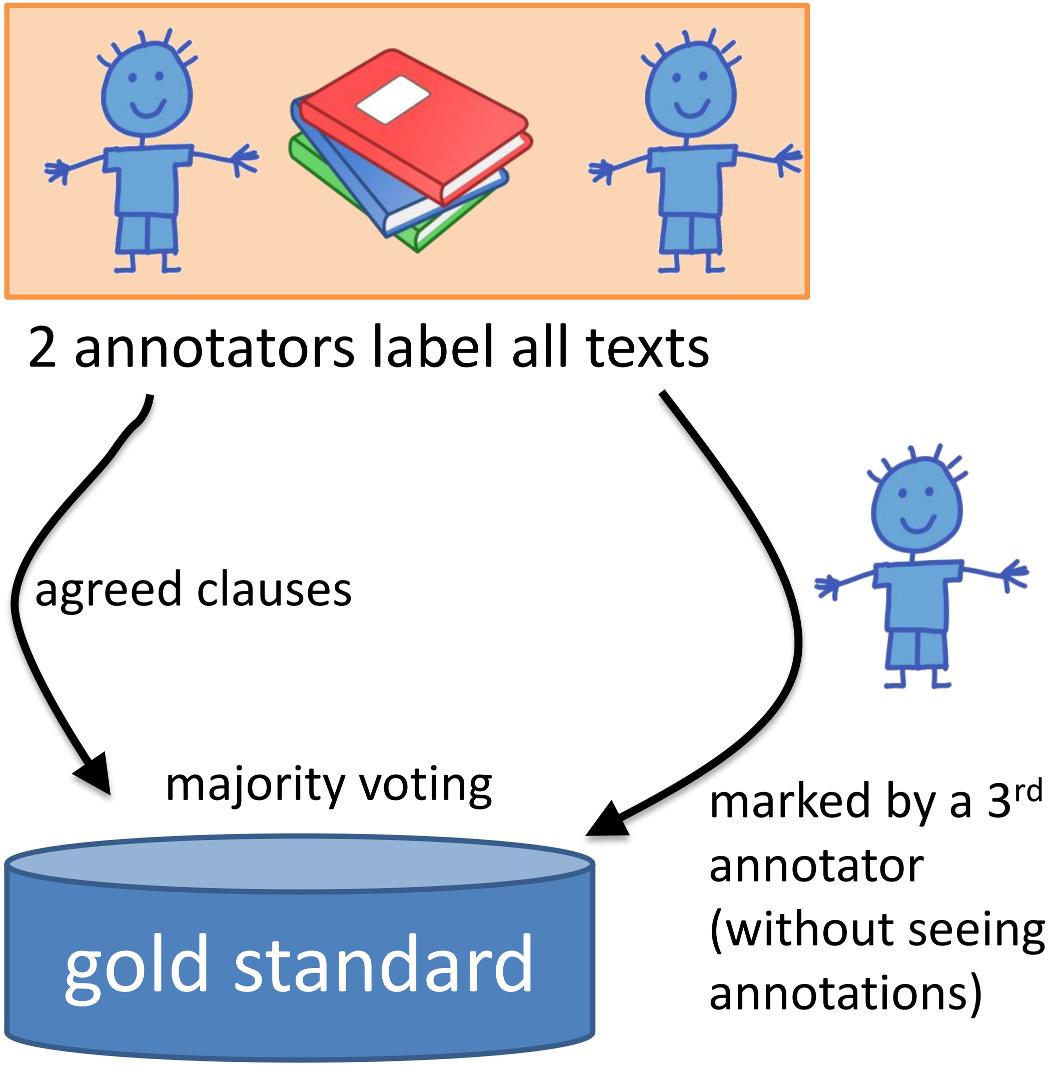
↓

SPADE system

↓

segmentation into clauses, subjects are not pre-marked

1) Lions are big cats
2) and eat meat.



2 annotators label all texts

agreed clauses

majority voting

marked by a 3rd annotator (without seeing annotations)

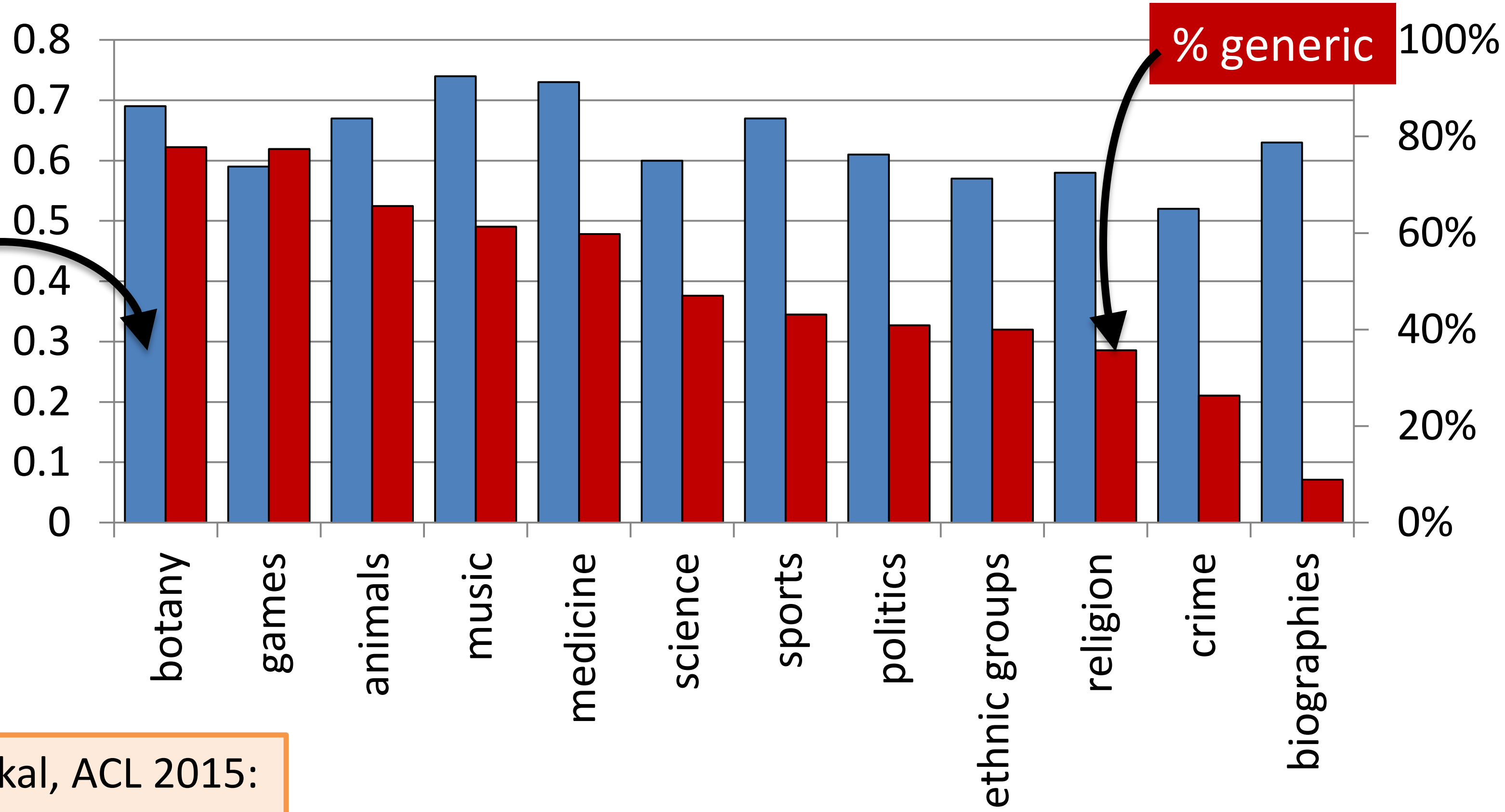
gold standard

Inter-annotator agreement: WikiGenerics

MASC, Cl+NP

Fleiss' κ

marked completely by 3 annotators



Friedrich & Pinkal, ACL 2015: **Discourse-sensitive Automatic Identification of Generic Expressions.**

Task NP	Task Cl	Task Cl+NP	% generic
0.69	0.72	0.68	50.1%

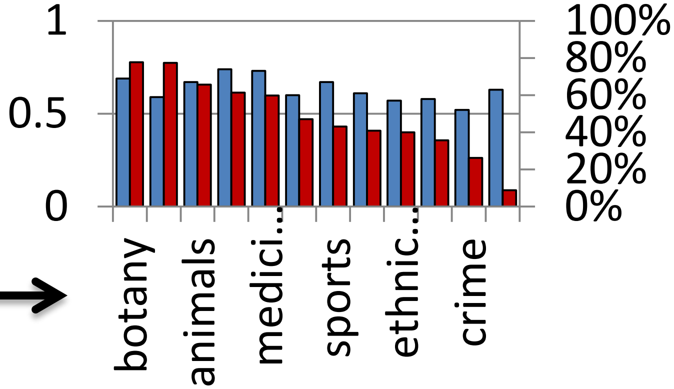
balanced corpus, substantial agreement

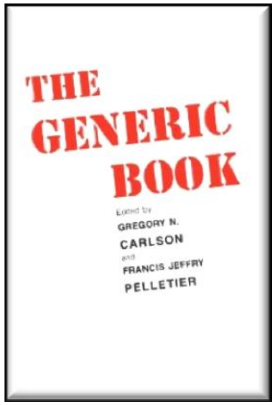
Annotating genericity

survey & our corpus:

K moderate substantial

interpretation in relation to label distribution




THE GENERIC BOOK

Edited by GREGORY N. CARLSON and FRANCIS JEFFRY PELLETIER

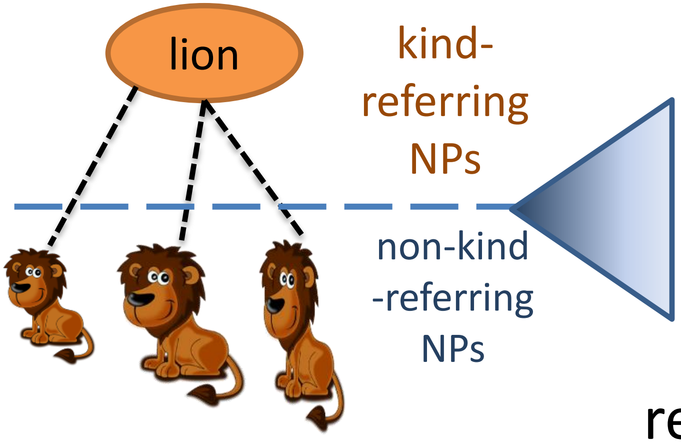
linguistically motivated 3-way **annotation scheme:** NP, Cl, Cl+NP

MASC, WikiGenerics

balanced

substantial agreement

www.coli.uni-saarland.de/projects/sitent



Students of Saarland university have lunch at mensa.

extensional (non-generic) vs. intensional (generic) reading

+ study related phenomena (e.g. habitual sentences)

+ extend to other languages

redefine USP label?

Thank you

Alexis Palmer

Melissa Peate Sørensen

Manfred Pinkal

Questions?

www.coli.uni-saarland.de/projects/sitent

References

ACE corpora: https://www.ldc.upenn.edu/collaborations/past-projects/ace

Bhatia, A. et al. **Automatic classification of communicative functions of definiteness**. (2014). In *Proceedings of COLING*, pp. 1059-1070. Dublin, Ireland.

Friedrich, A. & Pinkal, M. (2015). **Discourse-sensitive Automatic Identification of Generic Expressions**. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Beijing, China. *(to appear)*

Krifka, M. et al. (1995). **Genericity: an introduction**. *The Generic Book*, 1-124. University of Chicago Press.

Herbelot, A., & Copestake, A. (2010). **Annotating underquantification**. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 73-81). Association for Computational Linguistics.

Herbelot, A., & Copestake, A. (2011). **Formalising and specifying underquantification**. In *Proceedings of the Ninth International Conference on Computational Semantics.* Assocation for Computational Linguistics.

References (ctd)

Ide, N., Baker, C., Fellbaum, C., & Fillmore, C. (2008). **MASC: The manually annotated sub-corpus of American English.** In *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC*).

Louis, A., & Nenkova, A. (2011). **Automatic identification of general and specific sentences by leveraging discourse annotations.** In *IJCNLP 2011* (pp. 605-613).

Mathew, T. and Katz, G. (2009**). Supervised Categorization of Habitual and Episodic Sentences**. In *Sixth Midwest Computational Linguistics Colloquium*. Bloomington, Indiana: Indiana University.

Nedoluzhko, A. (2013). **Generic noun phrases and annotation of coreference and bridging relations in the Prague Dependency Treebank**. *LAW VII & ID*.

Poesio, M. (2004). **Discourse annotation and semantic annotation in the GNOME corpus**. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation* (pp. 72-79). Association for Computational Linguistics.

Reiter, N., & Frank, A. (2010, July**). Identifying generic noun phrases.** In *Proceedings of ACL (*pp. 40-49). Association for Computational Linguistics.

Soricut, R., & Marcu, D. (2003). **Sentence level discourse parsing using syntactic and lexical information.** ACL-HLT. (pp. 149-156). Association for Computational Linguistics.