

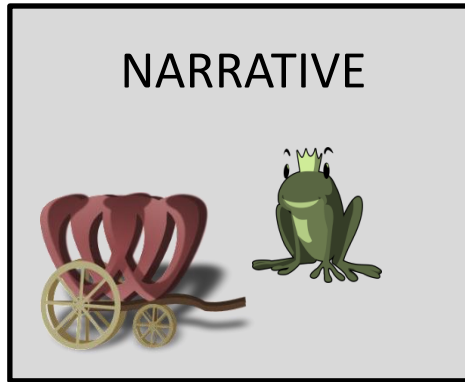
Annotating aspectual information

Annemarie Friedrich
joint work with: Alexis Palmer
Saarland University

Edinburgh May 6th, 2014

Modes of discourse [Smith 2003]

temporal progression



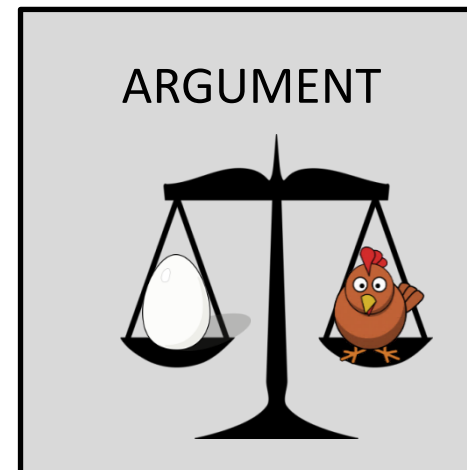
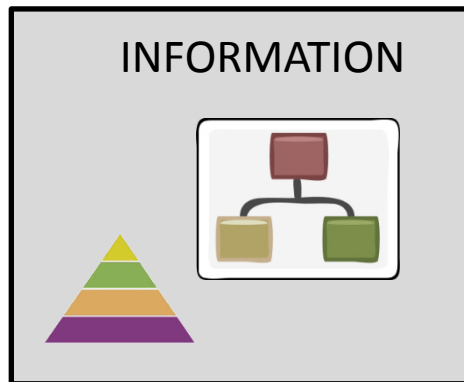
temporal progression,
related to speech time



time is static, spatial progression
through described scene

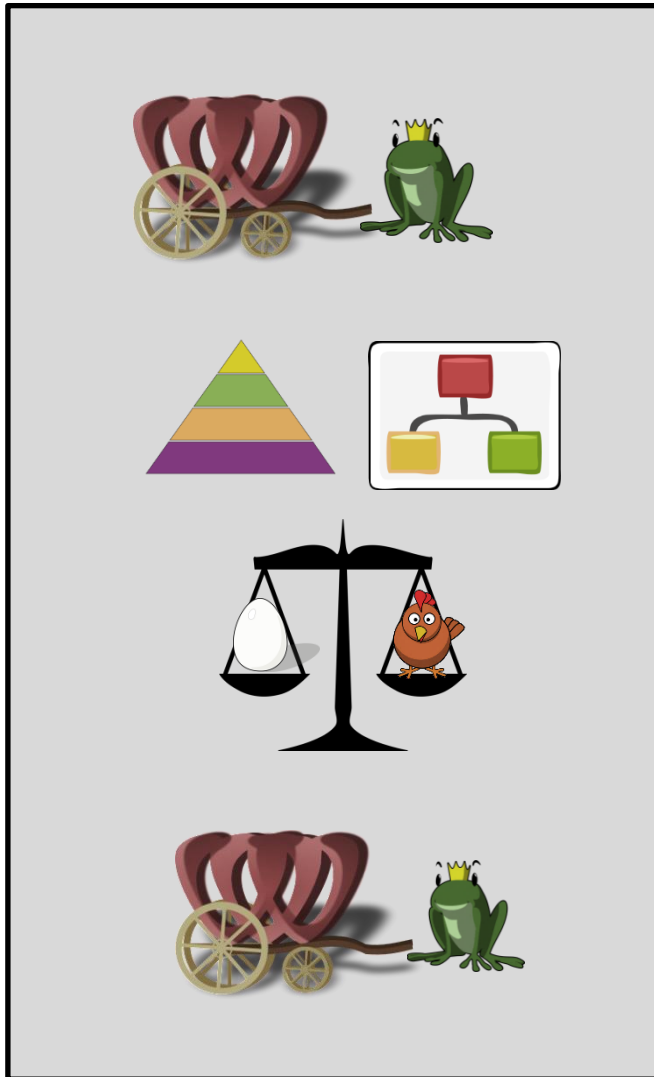


atemporal,
progression
metaphoric path
through domain
of text



atemporal,
progression
metaphoric path
through domain
of text

Modes of discourse \neq genre



one text \approx one genre

one text \neq one discourse
mode

Each passage of a text can have a
different discourse mode.

But: there are correlations
between the distribution of
situation entities and genre.
[Palmer & Sporleder, 2009]

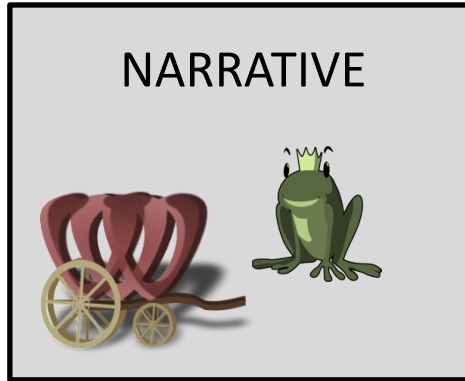
Situation entity (SE) types

Clauses introduce **situation entities** of different types to the discourse.

Eventualities	State	<i>Mary likes cats.</i>
	Event	<i>Mary fed the cats.</i>
	- Report	<i>..., Mary said.</i>
General Statives	Generalizing Sentence	<i>Mary often feeds my cats.</i>
	Generic Sentence	<i>Cats are always hungry.</i>
Abstract Entities	Fact	<i>I know <u>that Mary fed the cats.</u></i>
	Proposition	<i>I believe <u>that Mary fed the cats.</u></i>
Speech Acts	Question	<i>Does Mary like cats?</i>
	Imperative	<i>Don't forget to feed the cats!</i>

Modes of discourse [Smith 2003]

events, states



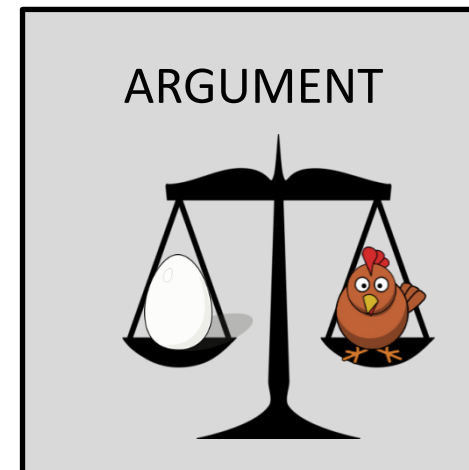
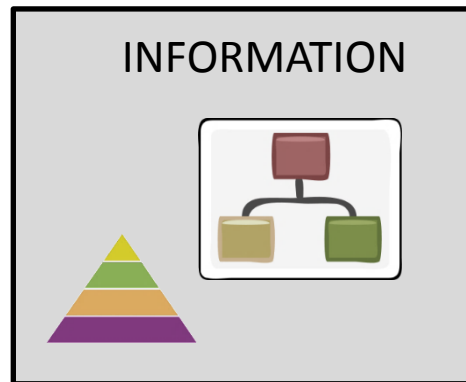
events, states,
general statives



events, states,
ongoing events



general
statives



fact,
proposition,
general
stative

Segmentation

situation \approx clause

most important for situation type: the verb & its arguments

Some of the strongest critics of our welfare system are the people
who have become dependent on it.
Their complaint: the system discourages working
because unemployment provides benefits
that many entry-level jobs don't include.
That's where you
– and Goodwill – come in.

SPADE discourse parser
[Soricut & Marcu 2002]
+ heuristic post-processing

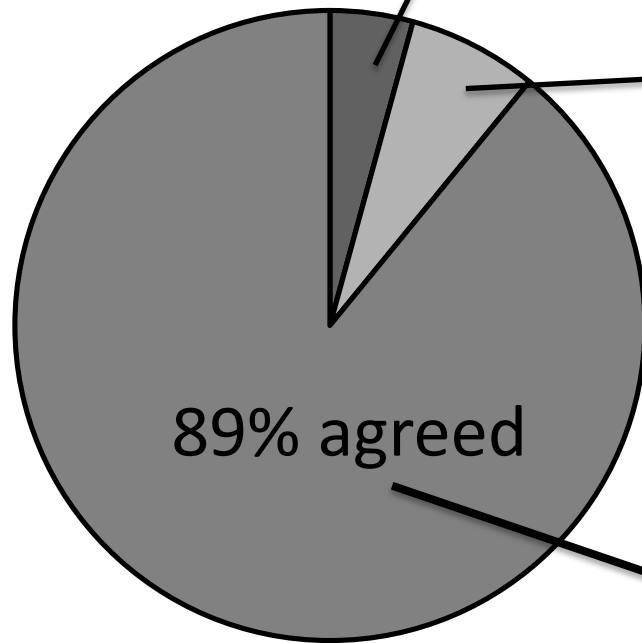
+ allow manual corrections
by annotators

segmentation problem

Segmentation: agreement

MASC news section
3 annotators

2823 automatically
created segments



4%
marked as **NO SITUATION**
by at least one annotator
(e.g. headlines, names, dates)

7%
merged to other segment
by at least one annotator

2515 segments for further agreement analysis

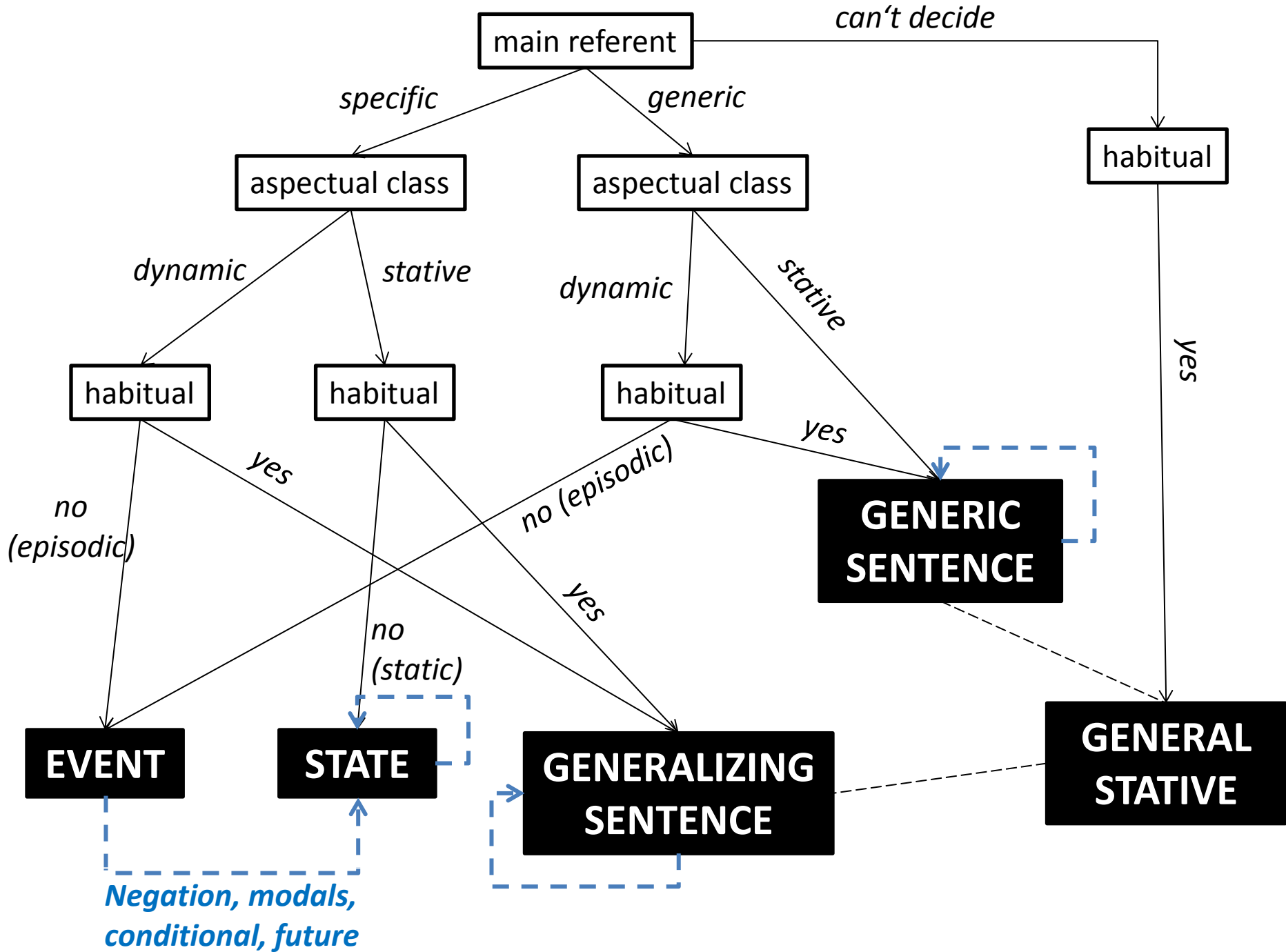
Feature-driven annotation scheme

Previous attempts (e.g. data used by Palmer et al. 2007):

- intuitive assignment of situation entity types
- no annotation manual

This approach: determining situation entity type involves

- Is the main referent specific (my car) or generic (red cars)?
Krifka et al. (1995)
- Is the aspectual class of the clause stative or dynamic?
following distinctions of Vendler/Bach/Mourelatos
- Is it a habitual (describes a regularity) or episodic (happens once)?
Carlson (e.g. 2005)



Situation Entity Types : Annotation



File: masc_essays_anth_essay_4_part1.txt



6	seg_prob	
7	seg_prob	Indigenous Revitalization Movements of Latin America:
8	seg_prob	
9	seg_prob	The Rediscovery of Language and Identity
10	seg_prob	
11	GEN_STAT, GENERIC	Humans, in the process of defining themselves, look to immediately recognizable factors
12	GEN_STAT, GENERIC	in order to differentiate themselves from others.
13	GEN_STAT, GENERIC	This process leads to the development of what can be broadly termed culture:
14	GEN_STAT, GENERIC	a group of people joined by characteristics, traits, and beliefs comes to view themselves as sharing a distinct identity, discernible from that of those "outside" them, those
15	GEN_STAT, GENERIC	who do not belong.
16	GEN_STAT, GENERIC	Naturally this can be signaled by any number of things - clothing, food, lifestyle choices.
17	GEN_STAT, GENERIC	One of the most dominant and easily noted factors is language.
18	GEN_STAT, GENERIC	The command of a language, while not affording guaranteed access to the community
19	ST	which speaks it,
20	GEN_STAT, GENERIC	is often the key to gaining entry.
21	ST	In turn communal identity can be formed around a language;
22	GEN_STAT, GENERIC	it is used as the focal point
23	GEN_STAT,	ground which the group gathers

Features

Main Referent

- specific
- generic
- expletive
- can't decide
- not the grammatical subject

Lexical Aspect

- stative
- eventive
- can't decide

Frequency

- episodic
- habitual/regularity
- static
- can't decide

Segmentation Problems

- no situation
- includes text that does not belong to this situation
- multiple situations
- no complete situation
 - belongs to previous situation
 - belongs to following situation
 - belongs to situation number:

Situation Entity Types

- STATE
- EVENT
 - REPORT
- GENERAL STATIVE
 - GENERALIZING SENT.
 - GENERIC SENTENCE
- ABSTRACT ENTITY
 - FACT
 - PROPOSITION
- SPEECH ACT
 - IMPERATIVE
 - QUESTION

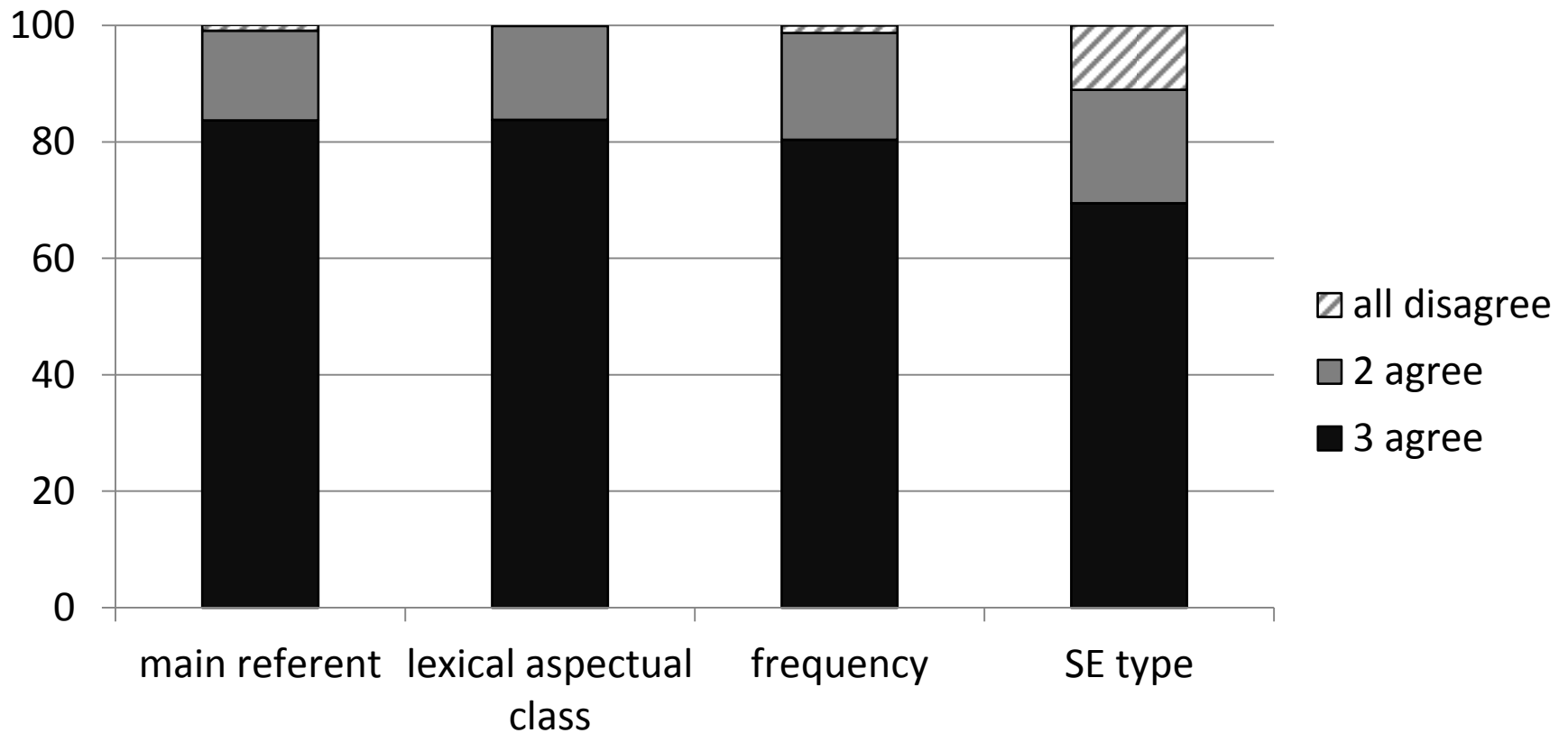
- Not done
- I'm not sure here

Comments:

Submit comment

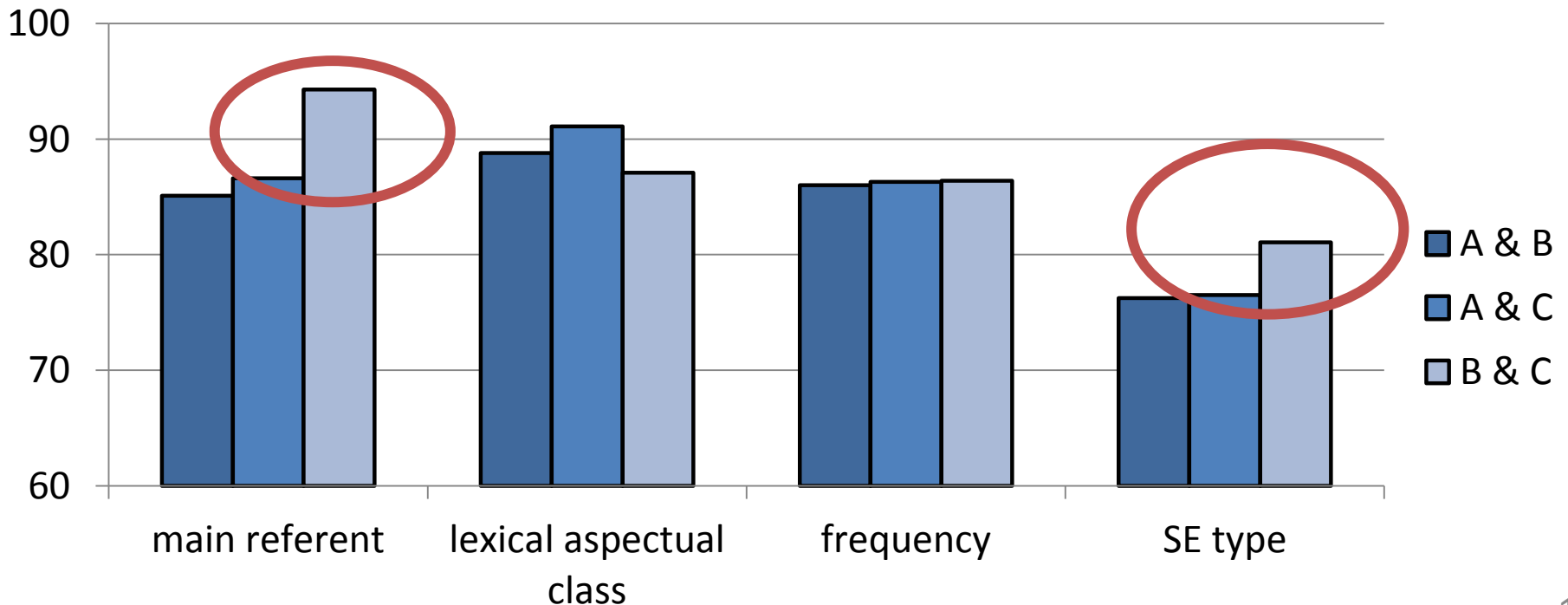
Agreement: news, 3 annotators

2515 segments where all 3 annotators agreed on segmentation



Agreement: news, pairs of annotators

- B&C achieve higher agreement than pairs involving A
- main difference vs. A: higher agreement on main referent type
 - disagreement with A: STATE vs. GENERIC SENTENCE
- by data / confusion matrix inspection:
 - A has tendency to mark difficult cases with GENERIC
 - B&C have tendency to mark difficult cases with SPECIFIC.



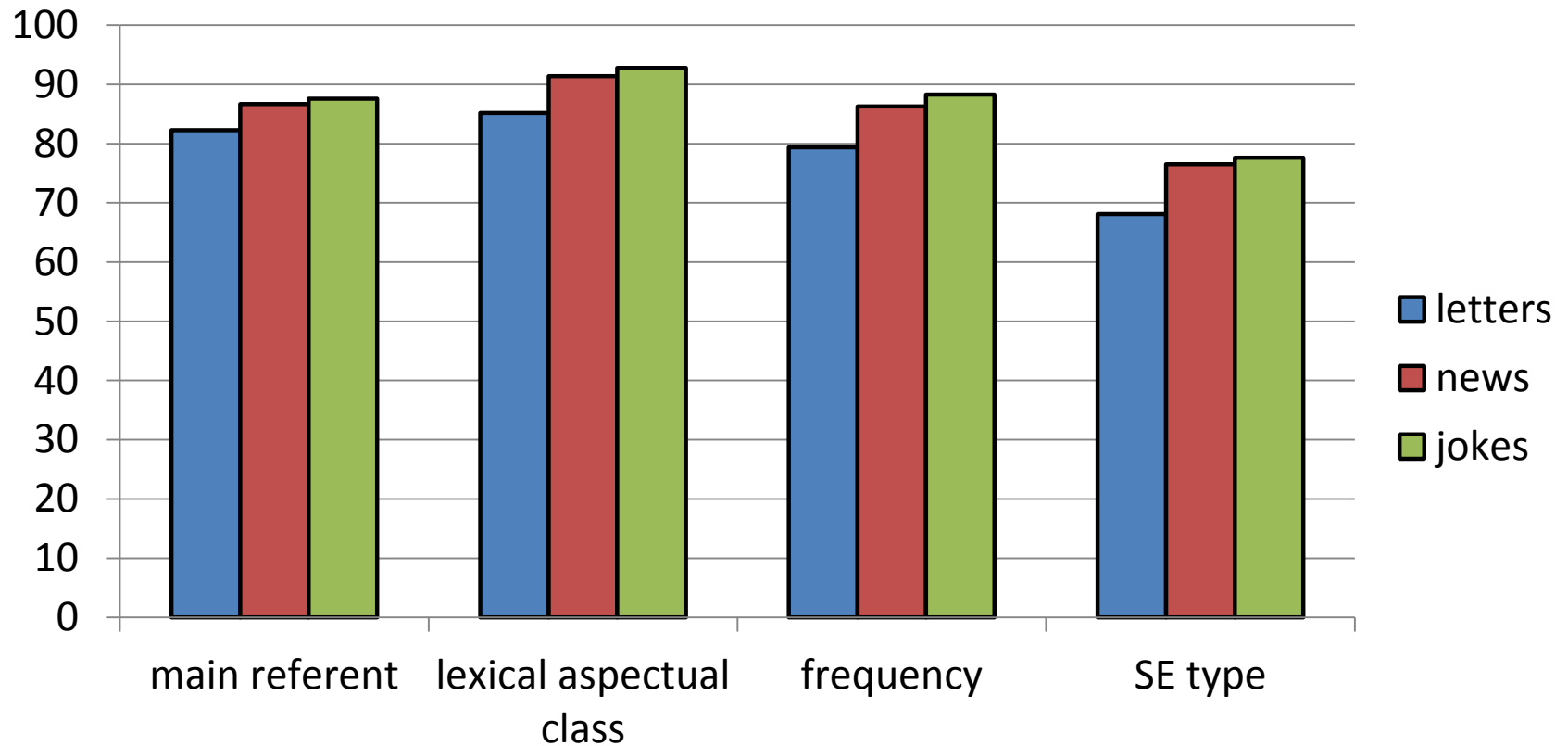
Disagreements on type of main referent

In many cases, both generic & specific readings are possible.

As governor, I'll make sure that every kid in New York has the same opportunity.

It shows that if you have the will and the dedication here you can learn a lot.

Agreement: 3 genres, 2 annotators

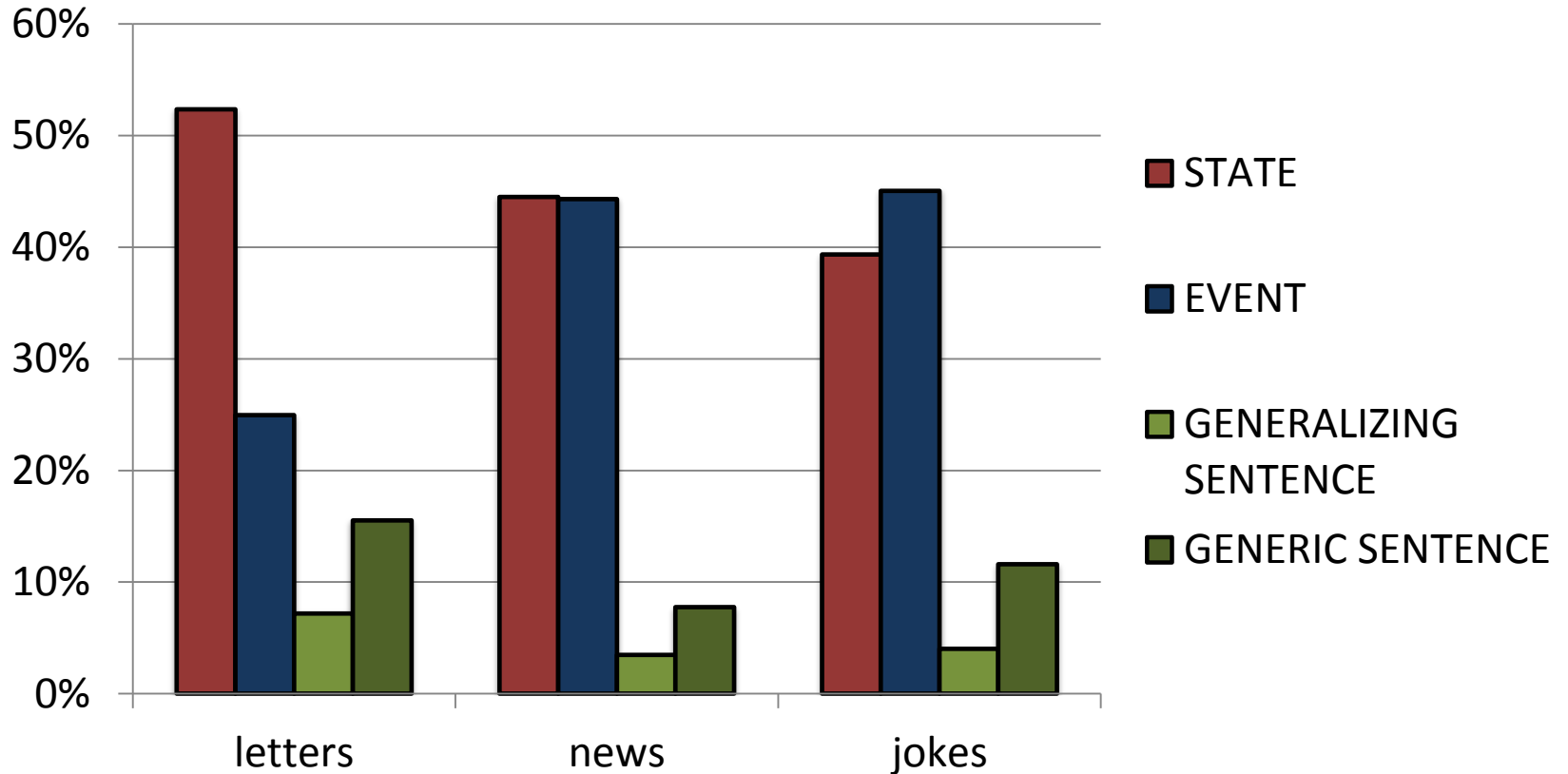


Cohen's $\kappa = 0.684$

on the four types:

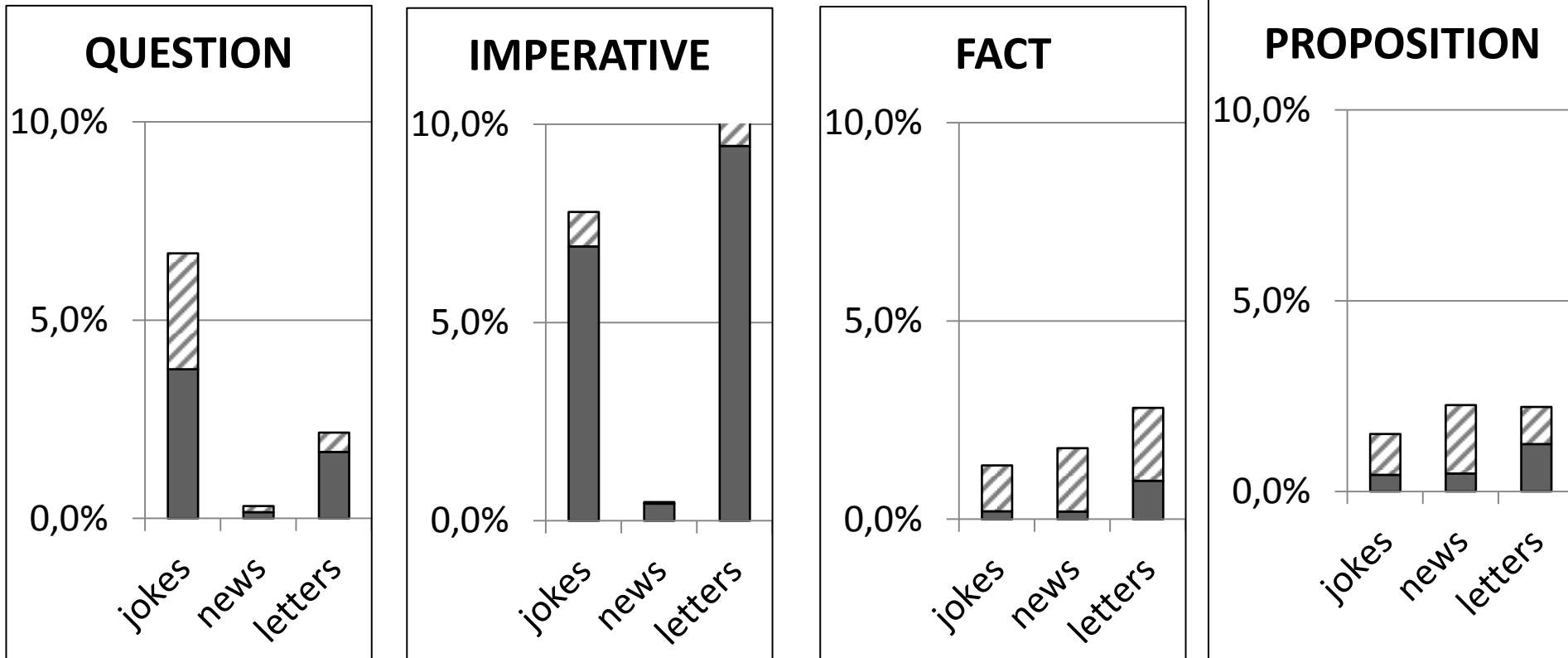
STATE, EVENT+REPORT, GENERALIZING SENTENCE, GENERIC SENTENCE

Distribution of SE types



agreement for **letters** is lower than for **news / jokes**
→ more General Statives in letters → “harder types”

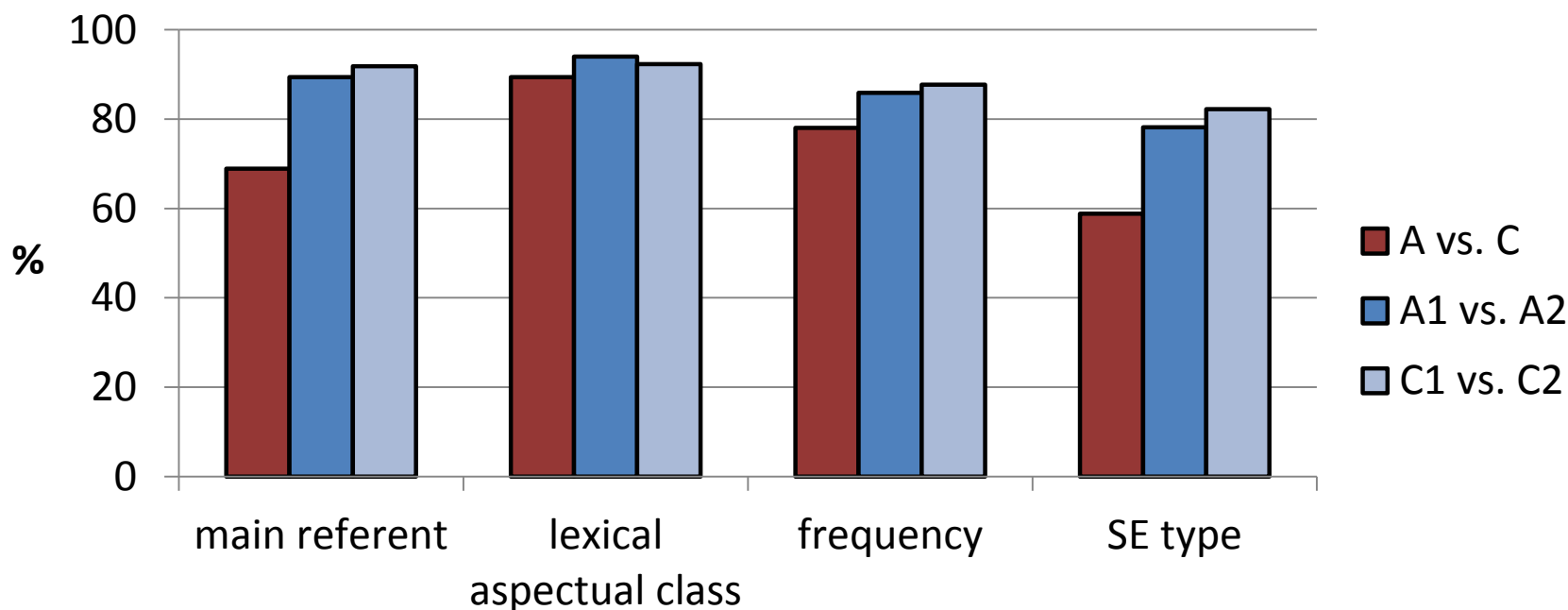
Agreement: 3 genres, 2 annotators, “surface SE types”



- **FACT & PROPOSITION:** no satisfying agreement yet.
 - Idea: highlight verbs potentially introducing these?

Intra-annotator consistency

11 (5 news, 5 letters, 1 jokes) documents (with lowest agreements on SE type) re-annotation by two annotators, after 3+ weeks after first pass.



→ agreement with oneself > agreement with other annotator
→ annotators have a different understanding of some cases

→ there are some noisy cases: annotators *do* disagree with themselves (this is not the full data set but the hardest part, total % of noise on SE type level << 20%)

Overall....

- agreement on “easy cases”
- disagreement mostly on cases where different readings are possible
- integrate this information into the learning methods