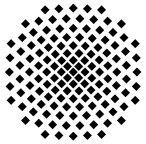
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Universität Stuttgart

Philosophische Fakultät
Seminar für Sprachwissenschaft

Sonderforschungsbereich 732
Institut für Maschinelle Sprachverarbeitung

Referential vs. Lexical Information Status

Annotating Corpora with Information Structure
ESSLLI 2014

Kordula De Kuthy and Arndt Riester

August 19, 2014

Overview *RefLex* – referential information status

Annotation units: DP, PP	
Label	**Description**
R-GIVEN	coreferential anaphor
R-GIVEN-DISPLACED	anaphor far from antecedent
R-GIVEN-SIT	symbolic deixis
R-ENVIRONMENT	gestural deixis
R-CATAPHOR	forward-looking anaphor
R-BRIDGING	non-coreferential and context-dependent
R-BRIDGING-CONTAINED	anchor is part of bridging anaphor
R-UNUSED-KNOWN	globally unique and known
R-UNUSED-UNKNOWN	globally unique and unknown
R-NEW	non-unique expression
+GENERIC	class reference, abstract/hypothetical entity

Referential vs. lexical GIVENNESS

- Recall that Schwarzschild (1999) distinguishes between (i) expressions of type *e* and (ii) "functional" expressions of type $\langle \alpha, \beta \rangle$.
  i. GIVENNESS = coreference anaphora
  ii. GIVENNESS = entailment / set inclusion (for words: repetition, synonymy or hypernymy)

- Halliday & Hasan (1976): different types of *lexical cohesion*
- Baumann & Riester (2012): R-GIVENNESS vs. L-GIVENNESS
- No use made of Schwarzschild's notion of *Existential F-closure*, i.e. on our account there are *new* expressions.

Referentially vs. lexically given

(1)

A man came in.	The	idiot	dropped a vase.
	R-GIVEN		
		L-NEW	

(2)

A student came in.	Another	student	greeted	him.
		L-GIVEN		
	R-NEW			R-GIVEN

(3)

A policeman came in.	Another	man	left.
		L-GIVEN	
	R-NEW		

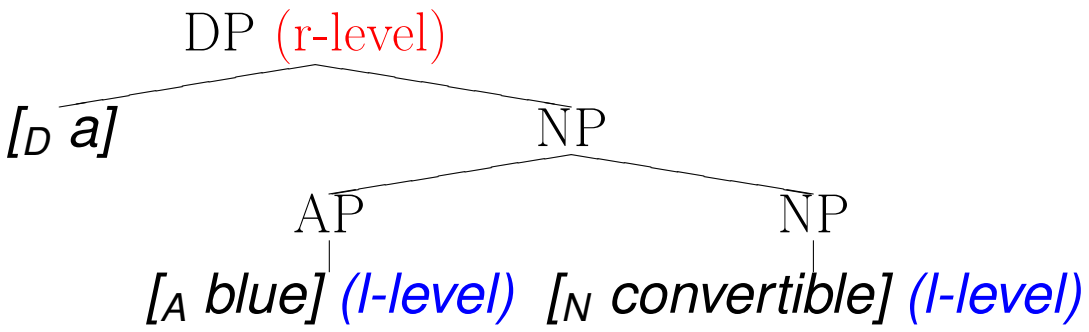
(4)

A woman came in.	The	woman	coughed.
		L-GIVEN	
	R-GIVEN		

Neither type of GIVENNESS is a prerequisite for the other!

Annotation conventions for lexical information status

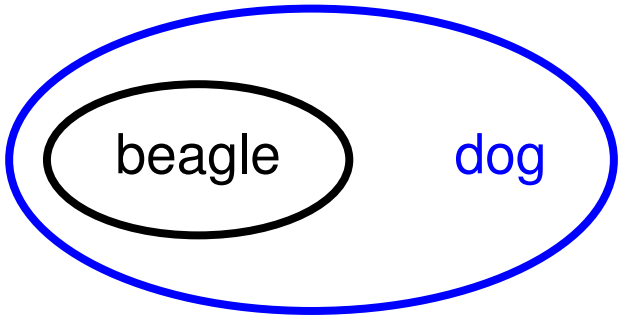
- From lexical givenness to **lexical information status**
- Lexical information status describes **semantic relations** between words and set-denoting phrases.
- In particular, we examine **content words**: nouns, verbs, adjectives, adverbs.



- Compounds are treated as word units
- Context window (5 clauses) to model “memory loss” and to make annotation feasible

I-given

- Markable is identical to, or a superset or holonym of, a previously mentioned expression



(5) Look at the funny beagle over there.

a. It makes me think of Anna's [beagle]. *I-given-same*

b. It makes me think of Anna's [dog]. *I-given-super*

(6) a. Where are your bags? Did you leave your [luggage] at the station?

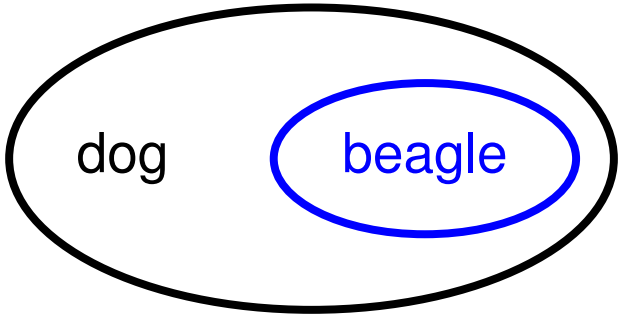
b. The Office for National Statistics said the inflation rate has slipped. The [ONS] cited motor fuels as a factor.

c. The PC is ready to obtain data and [receive] alarms from an external system. *I-given-syn*

(7) Florence is my favourite city in [Italy]. *I-given-whole*

l-accessible

- Markable is a subset or meronym of a previously mentioned expression, or related in a different way



(8) Look at the funny dog over there. It makes me think of Anna's [beagle]. *l-accessible-sub*

(9) In my hotel room the [ceiling] is 3m high and the [windows] won't open. *l-accessible-part*

(10) It is anticipated that complex financing schemes ([structured finances]) will be needed in an increasing measure to realise investment projects. *l-accessible-stem*

I-new

- A content word which is not semantically related to another expression within the current context window

(11) [Look] at the [funny] [dog] over there! It makes me [think] of [Anna's] [boyfriend].

(12) [Pakistan's] [highest] [court] has [declared] that the country's [prime minister] is [disqualified] from office.

Combining referential and lexical information status

(13) UN Special Envoy Ahtisaari is making the case for an independence of Kosovo under international control. This would be the only political and economic option for the future [of the Serbian province].

of the	Serbian	province
	L-NEW	L-GIVEN-SUPER
R-GIVEN		

(14) An earthquake has hit Central Japan. Also in the island state of Vanuatu in the Southern Pacific [two quakes] have been registered.

two	quakes
	L-GIVEN-SYN
R-NEW	

Snowden interview: RefLex combinations

60 [00:1	61 [00:19.0]	62 [00:1	63 [00:20.1]	64 [00:2	65 [00:20.7]	66 [00:21.5	67 [00:22.0*]	68 [00:2	69 [00:22.6]	70 [00:23.0]
the	GCHQ	is	collecting	an	incredible	amount	data	on	British	Citizens
DT	NP	VBZ	VBG	DT	JJ	NN	NNS	IN	JJ	NNS
r-given-displaced				r-new						
								r-new +generic		
	l-given-same		l-accessible-stem		l-new	l-new	l-given-same		l-given-same	l-given-same
	%							%		%
	*		*		*			*	*	

71 [00:2	72 [00:2	73 [00:2	74 [00:24.2]	75 [00:24.5]	76 [00:25.1]	77 [00:2	78 [00:26.1]	79 [00:26.6]	80 [00:27.3]	81 [00:2	82 [00:27.8*]	83 [00:2	84 [00:2	85 [00:28.4]
just	as	the	National	Security	Agency	is	gathering	enormous	amount	of	data	on	US	citizens.
RB	IN	DT	NP	NP	NP	VBZ	VBG	JJ	NN	IN	NNS	IN	NP	NNS
		r-given-displaced						r-new						
												r-new +generic		
			l-given-same				l-given-syn	l-given-syn	l-given-same		l-given-same		l-given-same	
					%		%				%			%
*			*	*			*	*					*	

Annotating complex constituents

- In principle, it is possible to also annotate larger set-denoting constituents.
- Entailment relations (as in Schwarzschild (1999) but without F-closure!)
- [blue convertible] $\models$ [car] $\rightarrow$ *l-given*
- [car] $\dashv$ [blue convertible] $\rightarrow$ *l-accessible*

Annotating complex constituents (cont.)

<table>
<tr><td>Ein</td><td>starkes</td><td>Erdbeben</td><td>hat</td><td>Zentral-Japan</td><td>erschüttert.</td></tr>
<tr><td>A</td><td>strong</td><td>earthquake</td><td>has</td><td>Central Japan</td><td>shaken</td></tr>
<tr><td rowspan="2"></td><td>L-NEW</td><td>L-NEW</td><td rowspan="2"></td><td>L-NEW</td><td>L-NEW</td></tr>
<tr><td colspan="2">L-NEW</td><td>R-UNUSED</td><td></td></tr>
<tr><td colspan="3">R-NEW</td><td colspan="3">L-NEW</td></tr>
<tr><td colspan="6">L-NEW</td></tr>
</table>

<table>
<tr><td>Die</td><td>Behörden</td><td>gaben</td><td>eine</td><td>Tsunami-Warnung</td><td>für den</td><td>Südwesten</td><td>heraus.</td></tr>
<tr><td>The</td><td>authorities</td><td>issued</td><td>a</td><td>tsunami warning</td><td>for the</td><td>southwest</td><td>–</td></tr>
<tr><td></td><td>L-NEW</td><td>L-NEW</td><td></td><td>L-NEW</td><td></td><td>L-NEW</td><td>L-NEW</td></tr>
<tr><td colspan="2" rowspan="2">R-BRIDGING</td><td rowspan="2"></td><td rowspan="2"></td><td rowspan="2"></td><td>R-BRIDGING</td><td rowspan="2"></td><td rowspan="3"></td></tr>
<tr><td colspan="3">R-NEW</td></tr>
<tr><td colspan="7">L-NEW</td></tr>
<tr><td colspan="8">L-NEW</td></tr>
</table>

“Information structure light” :o)

Empirical results from read text

Baumann & Riester (2013) Percentages of (i) nuclear pitch accents, (ii) pre-nuclear pitch accents, (iii) post-nuclear prominences and (iv) deaccentuation on short referring expressions in German read text, for different *RefLex* combinations.

Annotation units: N, V, A, Adv	
Label	**Description**
L-GIVEN-SAME	word identity
L-GIVEN-SYN	synonym
L-GIVEN-SUPER	hypernym
L-GIVEN-WHOLE	holonym
L-ACCESSIBLE-SUB	hyponym
L-ACCESSIBLE-PART	meronym
L-ACCESSIBLE-STEM	same word stem
L-NEW	unrelated

References

Baumann, S. & A. Riester (2012). Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. In G. Elordieta & P. Prieto (eds.), *Prosody and Meaning*, Berlin: Mouton de Gruyter, vol. 25 of *Interface Explorations*, pp. 119–162.

Baumann, S. & A. Riester (2013). Coreference, Lexical Givenness and Prosody in German. *Lingua* 136, 16–37. Special Issue 'Information Structure Triggers' ed. by Jutta Hartmann, Susanne Winkler and Janina Radó.

Halliday, M. & R. Hasan (1976). *Cohesion in English*. London: Longman.

Schwarzschild, R. (1999). GIVENness, AvoidF, and Other Constraints on the Placement of Accent. *Natural Language Semantics* 7(2), 141–177.