Bridging Relations in Polish: Adaptation of Existing Typologies

Maciej Ogrodniczuk | Institute of Computer Science, Polish Academy of Sciences

Magdalena Zawisławska | Institute of Polish Language, University of Warsaw

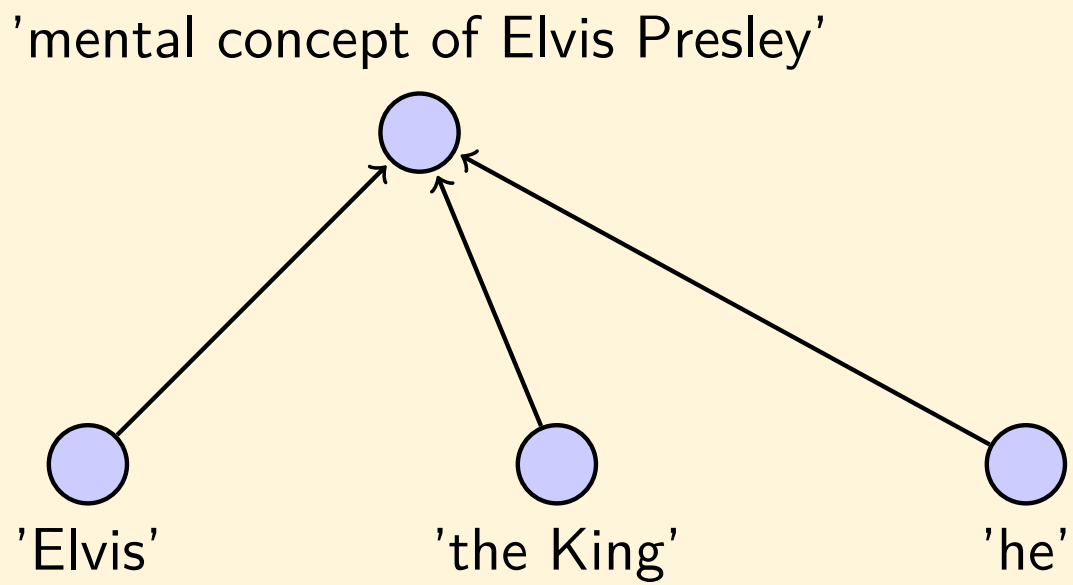
CORBON Workshop at NAACL 2016

San Diego, June 16, 2016

Coreference and bridging

Coreference:

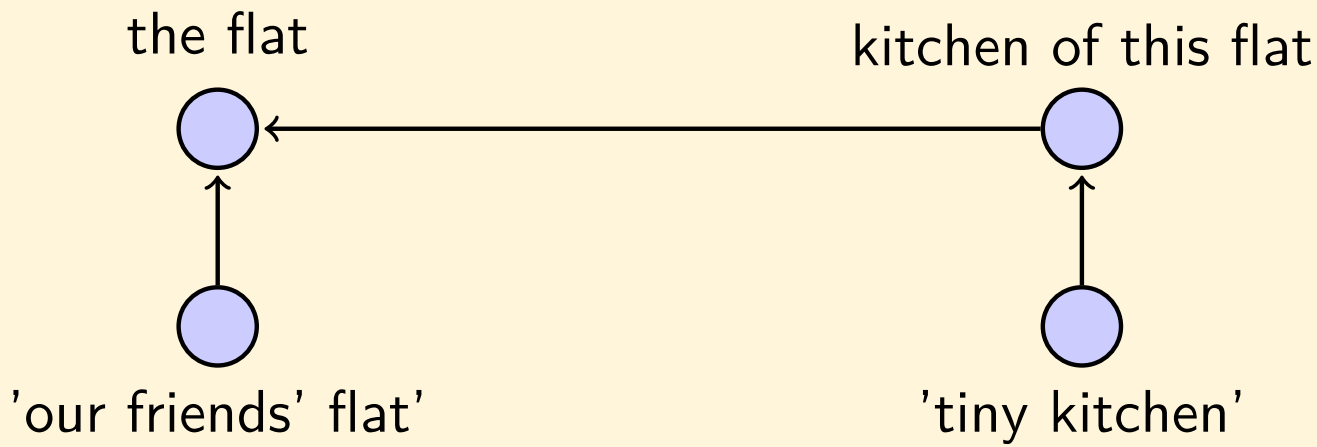
occurs when several textual expressions refer to the same discourse world object.



Coreference and bridging

Bridging:

(indirect reference, associative reference) occurs when some relation can be distinguished between targets of non-coreferential expressions and this relation influences coherence of the text.



Existing classifications of bridging

Clark, 1975:

Classic classification of indirect implicature lists *set membership*, *indirect reference by association* (necessary/probable/inducible parts) *indirect reference by characterization* (necessary/optional roles), *reason*, *cause*, *consequence* and *concurrence*.

Poesio, Vieira and Teufel, 1997:

Six classes: *synonymy/hyponymy/meronymy*, *names*, *compound nouns*, *events*, *discourse topic* and *inference*.

Existing classifications of bridging

Gardent, 2003:

Gardent summarizes bridging relations identified in the literature listing 13 categories (*set–subset*, *set–element*, *event–argument*, *individual–function*, *individual–attribute*, *whole–part*, *whole–piece*, *individual–stuff*, *collection–member*, *place–area*, *whole–temp.subpart*, *location–object* and *time–object*) and propose their own approach applied in annotation of PAROLE corpus, limited to: *set membership* (inclusion relation), *thematic relation* (thematic roles such as agent, patient etc.), *definitional relation* (attribute, meronymy etc.), *co-participant relation* and *non-lexical relation* (defined by discourse structure or world knowledge).

Existing classifications of bridging

Poesio and Artstein, 2008:

Annotation scheme for ARRAU allows *part–of*, *set–membership* and *converse* relation, which probably results from successful annotation of such limited number of relations in GNOME and VENEX corpora. The solution is similar to Recasens' annotation in CESS-ECE corpus, using 3 basic relations and *rest* type with no further subtype specification.

Existing classifications of bridging

Irmer, 2010:

Splits indirect references into mereological (*part-of*, *member-of*) and frame-related (thematic, causal, spatial, temporal) and offers a useful comparison of four other analyzed classifications (Winston, Iris, Vieu, Kleiber) which seem to differ in detail only.

GCBT, 2014:

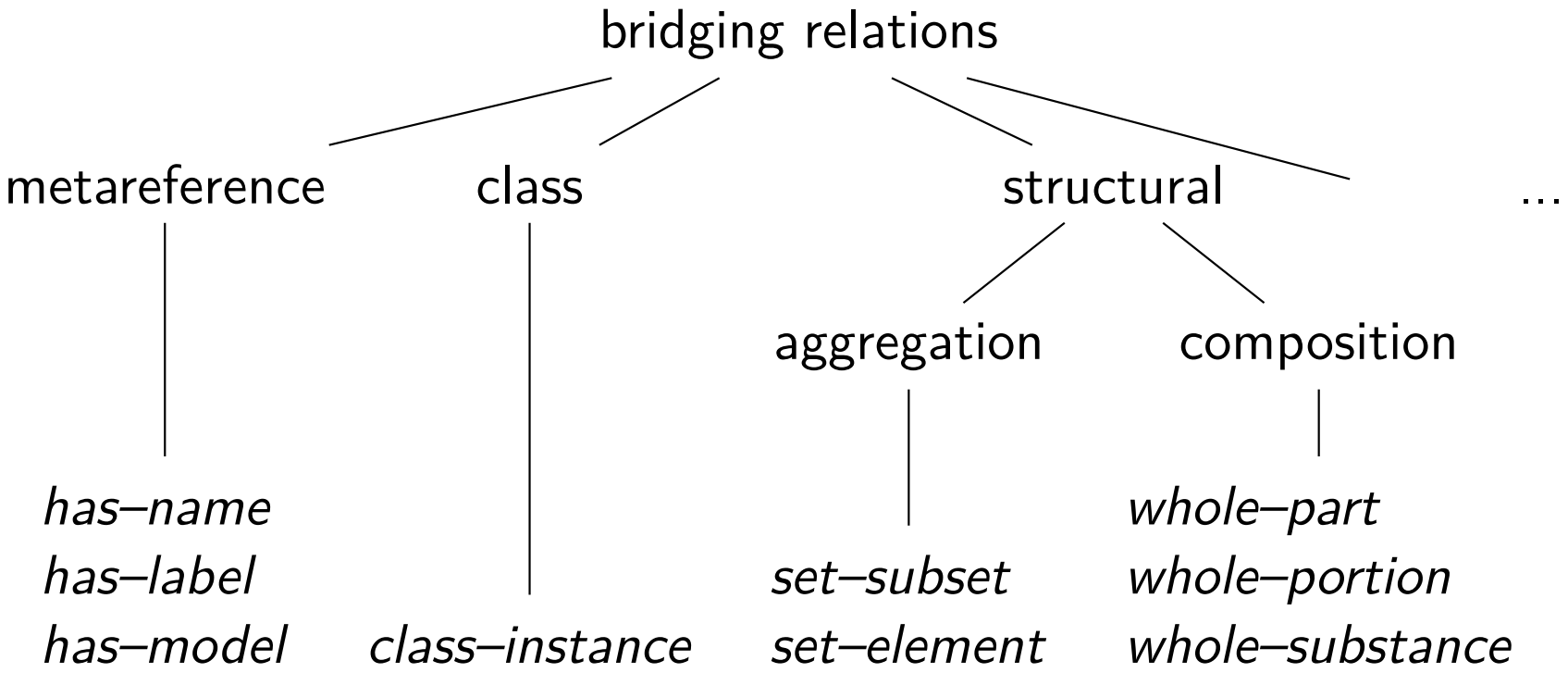
Greek Coreference and Bridging Team's annotation guidelines use *contrast*, *possession–owner*, two predicate relations, *entity–property* and *object–function* apart from traditional *set–subset* and *part–whole* relations. Other relations (spatial, temporal, generic–specific, thematic or situational association) are represented as *rest*.

Existing classifications of bridging

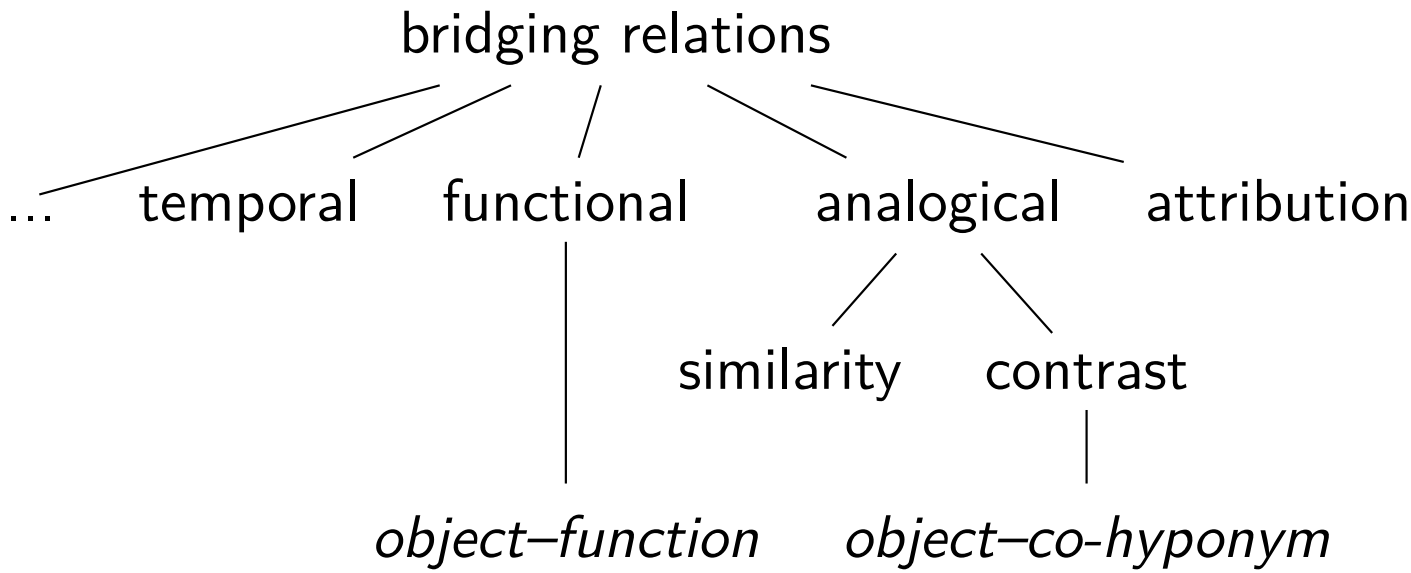
Prague Dependency Treebank, 2015:

In its present 3.0 version PDT uses six bridging relation types: *part–whole*, *set–subset/element*, *entity–singular function*, *contrast* (linking coherence-relevant discourse opposites), *non-coreferential explicit anaphoric relation* and *rest* (further unspecified group with *location–resident*, relations between relatives, *author–work*, *event–argument* and *object–instrument*).

Compiled classification: attempt 1



Compiled classification: attempt 1



The Polish Coreference Corpus

Bird's eve view:

- resulting from a national grant completed in 2015
- nominal direct coreference plus experimental annotation of near-identity
- the core: 1773 'short' plain texts (250–350 segments each, $>$ 500K segments in total)
- planned experimental near-identity annotation

Near-identity

Recasens' concept:

- a relation between two mentions when clear distinction between identity and non-identity is difficult
- two most frequent cases:
  - refocusing (e.g. "a child" vs. "an adult")
  - neutralization (e.g. "a book" vs. "a movie" with the same content).

Example:

'She hasn't seen "Gone with the wind", but she has read it.'
(this refers to both the book and the film)

Near-identity vs. quasi-identity

Our case:

Annotators were asked to identify 'other-than-identity' relations, without showing them the definition of near-identity.

Result:

Relations of different types were annotated, e.g. distorting or distinguishing properties of an object, metaphorical relations between substance and container ('quasi-identity'), but also set–element relations etc.

Corpus statistics

Text type	# mentions	# quasi-identity links
short	167,871	4,699
long	12,561	407
any	180,432	5,106

Text type	# singleton clusters	# non-singleton clusters
short	102,218	17,630
long	7,166	1,259
any	109,384	18,889

Preliminary corpus-based verification

From quasi-identity to bridging:

- randomly selected 5% (255) quasi-identity relations were reviewed
- two annotators previously involved in annotation of the corpus
- cases incompatible with the current proposal of the typology were marked as 'other':
  - coreference
  - predicate relations
  - errors (no relation)
- annotation agreement: 0.50 (Cohen's $\kappa = 0.36$)
- prevailing share of structural relations (60%).

Annotation statistics

		Metareference	Class	Temporal	Aggregation	Composition	Functional	Similarity	Contrast	Attribution	Coreference	Predicate	Other	ALL
1	**Metareference**	1	2		2								1	**6**
2	**Class**	1	15		7					1			1	**25**
3	**Temporal**		2	2										**4**
4	**Aggregation**	1	15		70	3	1			3	5	3	2	**103**
	Composition		1			8	1				2	2		**14**
5	**Functional**		3		5	1	9	2	1		3		1	**25**
6	**Similarity**							4						**4**
	Contrast				6									**6**
7	**Attribution**				2									**2**
8	**Coreference**		9		12	2	3	2		6	11	1	2	**48**
	Predicate				1	1				4		3		**9**
	Other		1		1	1	1			1			4	**9**
	ALL	**3**	**48**	**2**	**106**	**16**	**15**	**8**	**1**	**15**	**21**	**9**	**11**	**255**

Error analysis

Source of errors:

- too vague definition of some categories, e.g.
  - attribution
  - class vs. set
- extensive *other*: too many non-classified phenomena (entailment, metonymy etc.)
- confusion of the coreference, near-identity and other semantic relations (such as WordNet relations used to express direct coreference — and not bridging)
- changes in annotation guidelines.

Compiled classification: attempt 2

Relation	Count
Structural	**122**
Aggregation	105
Collection	7
Group	63
Hyponymy	35
Composition	17
Class	**44**
Entailment	**14**
Effect	8
Function	6
Attribution	**13**

Relation	Count
Analogical	**5**
Similarity	3
Contrast	2
Metareference	**3**
Dissimilation	**2**
Temporal	1
Contextual	1
Error	**52**
Coreference	17
Apposition	11
Predicate	9
Other	15

What comes next?

Questions:

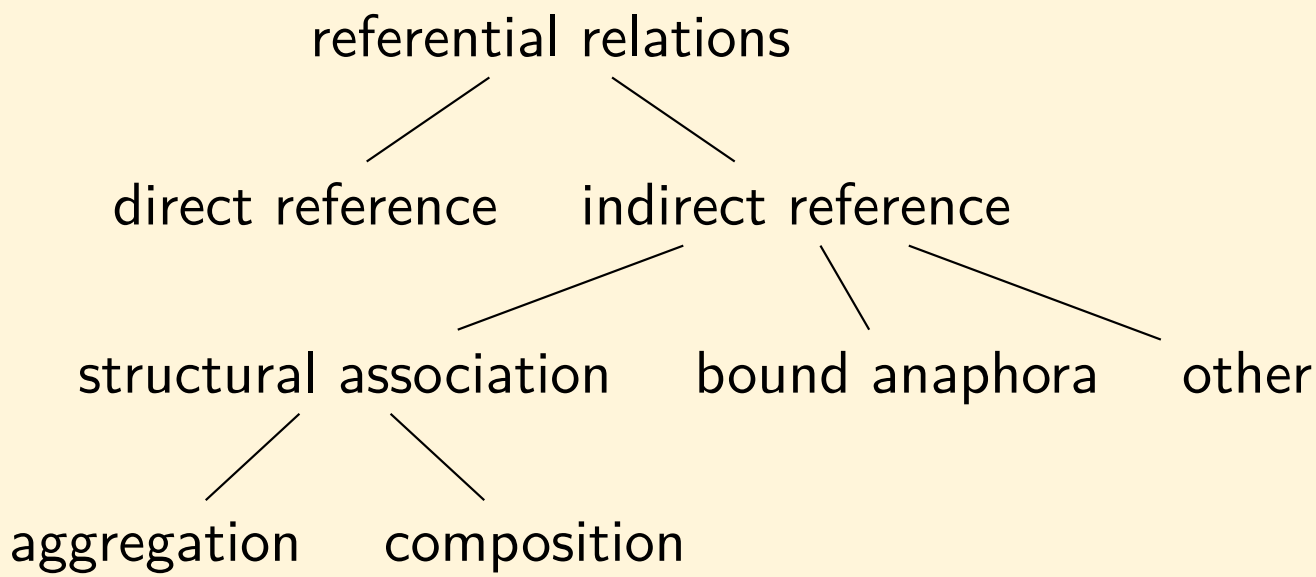
- which other factors are blurring the relation?
  cf. *A man started running towards me. Later it occurred it was Paul.*
- what do we do with non-obvious clues in the text?
  cf. *Paul painted it. [...] The author of the painting...*

Validation:

- more systematic annotation is needed
- a new national grant was acquired for this purpose
- but: can we use what we have as annotation guidelines?

Compiled classification: attempt 3

Referential relations:

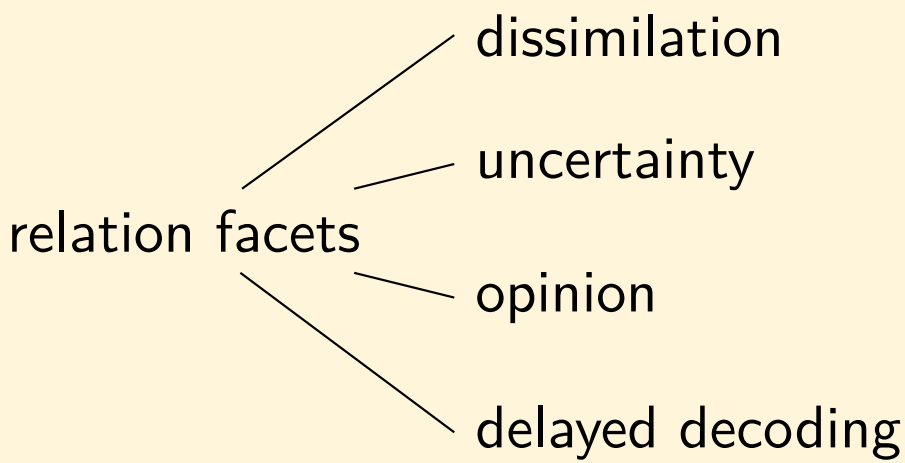


Compiled classification: attempt 3

Concept of a facet:

Relation facet is some property changing interpretation of the relation or signalling its incompleteness.

Relation facets:



Opinion

The idea:

Opinion (attribution) facet marks relations between an object and someone's opinion on the object (i.e., what is believed, doubted etc.) It assigns subjectivity to the link, as expressed by the speaker.

Example:

— *What's the name of Anna's husband?*
— *Michał, I guess.*

Uncertainty

The idea:

Uncertainty represents indeterminateness of pair of objects, if expressed by the speaker.

Example:

He is president but I am not sure whether it is the president of Warsaw or Cracow.

Delayed decoding

The idea:

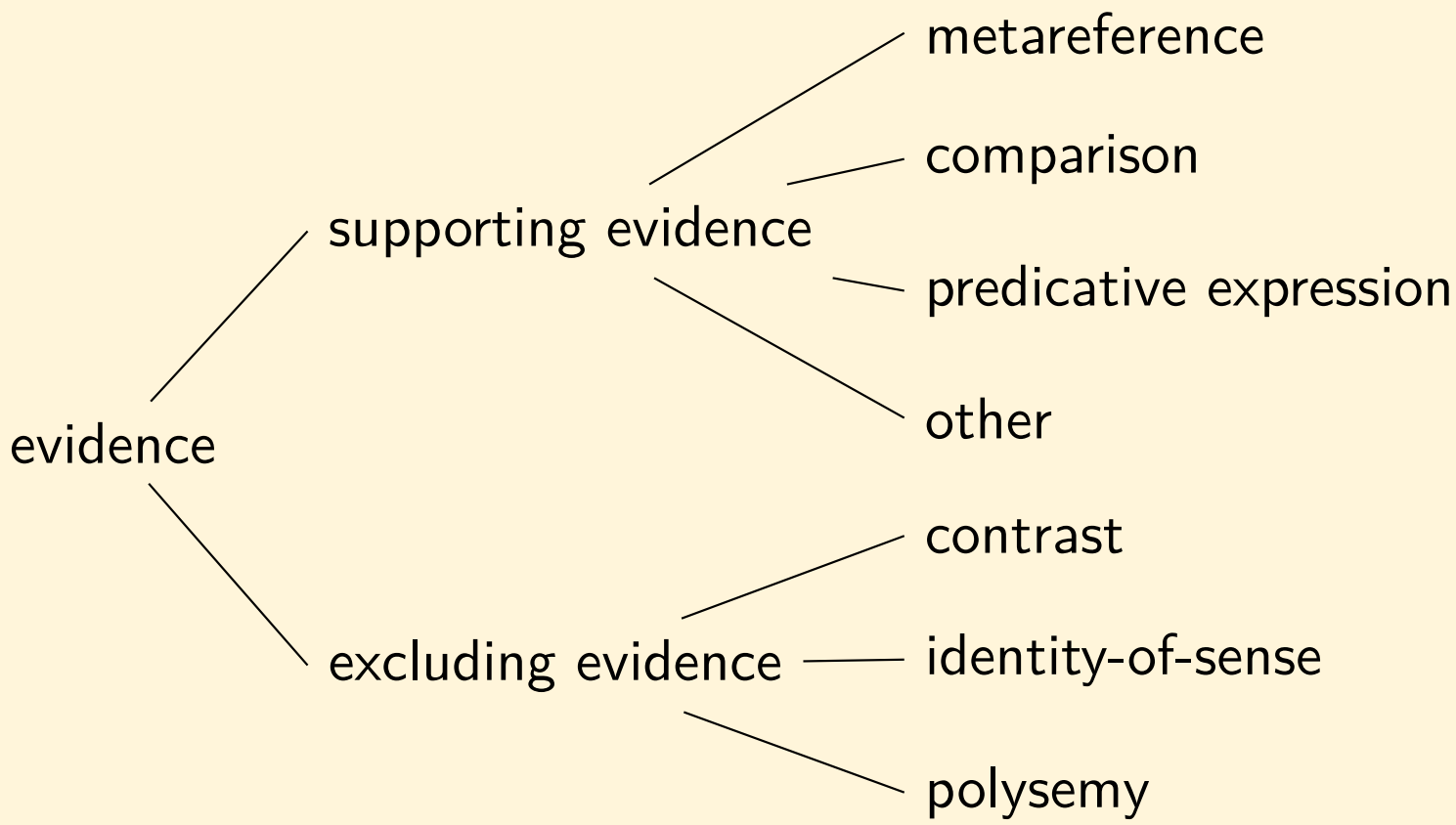
Delayed decoding facet indicates that the relation cannot be established when first mention is encountered in the text.

Example:

No one knew who the murderer was. [...] At the end of the day Peter pleaded guilty.

Compiled classification: attempt 3

Evidence:



Supporting evidence

Two examples:

His head resembled a big baloon. Suddenly the baloon guy took out the gun...

Peter lit the candle and gave the bouquet to his wife. – Blow it out, I don't feel like celebrating my birthday – said Eve.

Thank you!

The grant:

The work reported here was carried out within the research project financed by the Polish National Science Centre (contract number 2014/15/B/HS2/03435).

The purpose of the grant is to create:

- methods and tools to enable resolution of general referential relations
- a corpus manually annotated with bridging relations, predicates, non-nominal coreference...