Bridging Relations in Polish: Adaptation of Existing Typologies

Maciej Ogrodniczuk
Institute of Computer Science
Polish Academy of Sciences
Jana Kazimierza 5
01-248 Warsaw, Poland
`maciej.ogrodniczuk@ipipan.waw.pl`

Magdalena Zawisławska
Institute of Polish Language
Warsaw University
Krakowskie Przedmieście 26/28
00-927 Warsaw, Poland
`zawisla@uw.edu.pl`

Abstract

The paper attempts at presenting initial verification of existing approaches to annotation of bridging relations by proposing a compiled model based on schemata used in previous annotation projects and testing its validity on a corpus of Polish. The categorization features structural relations, dissimilation, analogy, reference to label, class, entailment and attribution. Multiple categories can be assigned to model situations where several aspects of the relation play a part. The relations are organized hierarchically which allows varied granularity of processing depending on computational needs. The classification is confronted with existing annotation of other-than-identity relations in a portion of Polish Coreference Corpus. Results of manual annotation involving two annotators and adjudicator are presented. Findings from the process are intended to facilitate development of annotation guidelines of a new reference-related project.

1 Introduction

The term *bridging* (bridging anaphora, indirect anaphora, associative anaphora) refers to relations between non-coreferential expressions that influence the text coherence. In most cases these expressions are nominal (and we will limit our analysis to such cases in this paper), although bridging between events can be also distinguished.

Several classifications of bridging relations are available, both in theoretical approaches (Clark, 1975; Prince, 1981; Löbner, 1996; Asher and Lascarides, 1998) and previous annotation projects (Poesio et al., 1997; Poesio, 2000; Gardent et al., 2003; Poesio et al., 2004; Poesio and Artstein, 2008). Another source of inspiration can be found in ontologies such as Cyc (Lenat, 1995) or lexical databases such as WordNet (Fellbaum, 1998) Yet there seem to be no consensus over general classification of such phenomenon.

In the article we attempt to compile the existing taxonomies of bridging relations into a common model, validate it on corpus data and present findings from the process which are planned to help develop annotation guidelines for the new project involving annotation of referential relations in Polish.

2 Related Work

Clark's classic classification of indirect implicature (Clark, 1975) 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's classification (Poesio et al., 1997) is composed of six classes: *synonymy/hyponymy/meronymy*, *names*, *compound nouns*, *events*, *discourse topic* and *inference*.

Gardent et al. (2003) summarize bridging relations identified in the literature listing 13 cat-

egories (*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 PA-ROLE 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).

Poesio and Artstein's annotation scheme for AR-RAU (Poesio and Artstein, 2008) allows *part–of*, *set–membership* and *converse* relation, which probably results from successful annotation of such limited number of relations in GNOME (Poesio, 2000) and VENEX corpora (Poesio et al., 2004). The solution is similar to Recasens' annotation in CESS-ECE corpus (Recasens et al., 2007), using 3 basic relations and *rest* type with no further subtype specification.

Irmer's classification (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 et al., 1987; Iris, 1988; Vieu and Aurnague, 2007; Kleiber, 1999) which seem to differ in detail only.

Greek Coreference and Bridging Team's annotation guidelines (GCBT: Greek Coreference & Bridging Team, 2014) 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*.

Prague Dependency Treebank (PDT) in its present 3.0 version (Zikánová et al., 2015, chapter 4) 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*).

3 Compilation of Typology of Bridging Relations

The proposed initial classification unifying existing approaches is depicted in Figure 1. Each main branch represents the intended relation type; leaf relations are specified as examples only.

3.1 Metareference

The relation allows to model relations such as *has–model*, *has–name* or *has–label*. This covers e.g. PDT's meta-linguistic reference, a subtype of non-cospecifying anaphoric relation.

(1) *Byłem wczoraj w restauracji „Smaczna rybka", ale ich ryby mi w ogóle nie smakowały.*
'I was yesterday in a restaurant called "Delicious Fish" but I didn't like their fish at all.'

3.2 Class

Class-instance relation, for some seen as of privileged nature, is represented similarly to standard *part–whole* or *set–member*, so reference between class and instance can be modelled in a unified manner.

(2) *Kobiety mają prawo do takiej wolności. Dlatego dobrze, by Ewa przekonała się, że nie wszystko musi być tak, jak było w rodzinnym domu.*
'Women have the right to such freedom. It is all right then for Eve to get convinced that not everything must be as it was in her family home.'

3.3 Temporal Relation

Temporal relation will be used to represent near-identical temporal aspects of the object (e.g. 'prewar Warsaw' and 'Warsaw of today'). Note that traditional temporal expressions such as anaphoric references to the time when the antecedent situation takes place (e.g. 'this time' and an event; a subtype of PDT's non-cospecifying anaphoric relation) will not be marked as temporal bridging relations (due to nominal intention of the current typology).

(3) *Warszawa jest pięknym miastem, ale przedwojenna Warszawa była jeszcze*

Figure 1: Bridging relations: compiled version
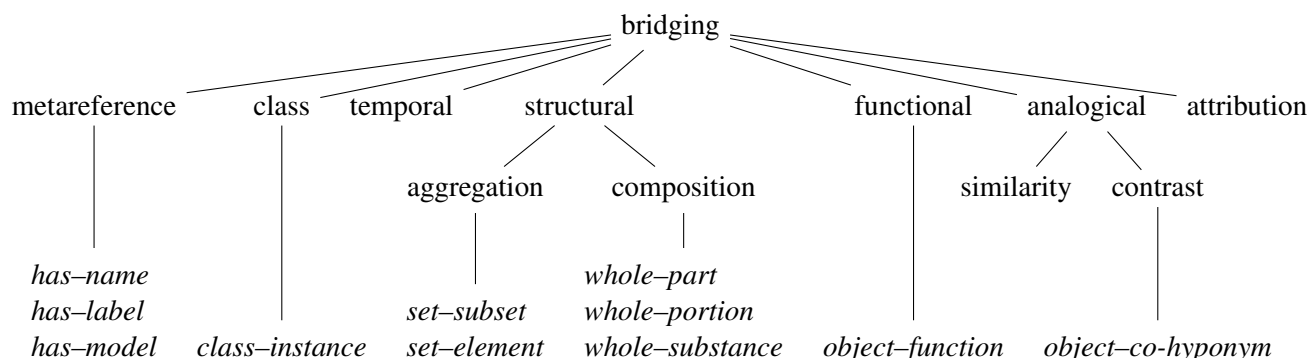
piękniejsza.
'Warsaw is a beautiful city, but pre-war Warsaw was even more beautiful.'

3.4 Structural Relation

Structural or meronymic relations are probably the least controversial part of the taxonomy, starting with Clark's necessary/probable/inducible parts through standard aggregation (*set–subset*, *set–element*) and composition (*whole–part*) to relations introducing inseparability such as *whole–portion* (also called *segment*, e.g. 'cake/slice') or *whole–substance* (e.g. 'cake/flour'). A ready-to-use sub-classification of meronymic relations can be found e.g. in (Winston et al., 1987).

(4) *Dobrze się czuję jako matka synów. Mój pierwszy syn nazywa się Adam.*
'I am feeling good as a mother of sons. My first son is Adam.'

3.5 Functional Relation

The basic *function–object* relation (as e.g. in PDT), causal relations from literature, Clark's necessary/optional role and Gardent's *thematic relation* can be interpreted as functional relations. Most PDT's 'other' relations such as *location–resident*, *event–argument* or *author–work* are also regarded as functional.

Clark's indirect reference by characterization also falls into this category, though it is mostly used for events and not objects.

The most interesting aspect of the functional relation is its correspondence with Recasens' near-identity (Recasens et al., 2012). In our opinion such weak near-identity cases as *representation* (e.g. between a manuscript and its content printed in a book) should be modelled as functional relations.

(5) *Intencja konkursu nie jest literacka, ale socjologiczna. Jeśli w wyniku wyłonią się jakieś talenty, będzie bardzo dobrze.*
'Intention of the contest is not literary but sociological. If any talents emerge as a result, we will be very fine.'

3.6 Analogical Relation

Both similarity relations (signaled by *such as* etc.) and contrast relations are intended to be marked as analogical.

(6) *Jego głowa była ogromna. Jak wielki balon.*
'His head was enormous. Like a big balloon.'

3.7 Attribution

Attribution is a type introduced to represent relations between an object and someone's opinion on the object (i.e., what is believed, doubted etc.) or indicate incomplete certainty about the nature of identity between two mentions.

(7) — *Jak się nazywa mąż Ani?*
— *Chyba Michał.*
'— What's the name of Anna's husband?
— Michał, I guess.'

In most projects this relation is annotated as coreference, but in general case (e.g., when several clashing opinions are represented in one discourse) such approach seems to be inappropriate.

		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**

Table 1: Results of the annotation experiment.

4 From Quasi-identity To Bridging Relations

The proposed classification was initially validated on the Polish Coreference Corpus (Ogrodniczuk et al., 2015, chapter 8). During its annotation, apart from marking direct identity-of-reference, annotators were asked to identify 'quasi-identity' relations, i.e. relations distorting or distinguishing properties of an object, metaphorical relations between substance and container, set-element relations and other relations not characterized by identity or non-identity. Over 5100 instances of such relations were marked, making a useful resource for corpus-based investigation of bridging.

## 4.1 Preliminary Corpus-based Verification

Randomly selected 5%, i.e. 255 relations, were reviewed to provide material for evaluation of the proposed taxonomy. The process was carried out by two annotators previously involved in classification of quasi-identity relations in the Polish Coreference Corpus. Cases incompatible with the current proposal of the typology were marked as 'other' with three subtypes: 1) coreference, for cases where original annotators of the Polish Coreference Corpus set quasi-identity type to a direct coreferential relation by mistake, 2) predicate, where relation was used to link mention with a predicate noun, and 3) error, for cases when no relation could be identified reasonably.

The results of this experiment are presented in Table 1. The annotation agreement was 0.50 (Cohen's $\kappa = 0.36$) which indicates that the typology is not precise enough to be used efficiently in practice.

The prevailing share of structural relations (60%) is compatible with Gardent's findings (Gardent et al., 2003, Figure 5) where 52% of the investigated relations were of meronymic type.

Figure 2: Bridging relations: updated version
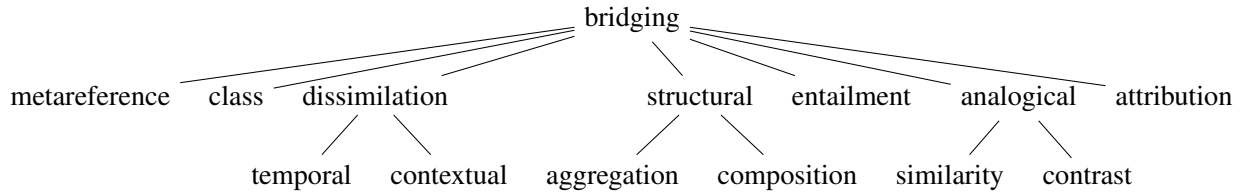
4.2 Error Analysis

The probable causes of divergence in the annotation are: 1) too extensive annotation categories, 2) too vague definition of some categories, 3) too many non-classified phenomena, 4) confusion of the coreference, near-identity and other semantic relations.

Some categories distinguished at the beginning turned out to be too extensive. Almost 44% examples were classified as of the *set* category. On the other hand, this category includes very diverse examples which calls for its division into subcategories in the prospective annotation.

The definitions of the *predicative* and *attribution* classes were not clear enough which led to confusion. Other difficult pairs were: *class* and *set*, *class* and *function*, *class* and *meta*.

In the proposed classification the category *other* was included for all doubtful examples. The annotations had shown that too many examples were classified as *other* and that there are quite distinct categories like: causality, connection of content or dissimilation.

In some cases making the distinction between definiteness and indefiniteness is virtually impossible. For example, when previous part of the text includes information on a merger of companies A and B and then someone comments that the idea of a merger of companies is cost-justified, it depends on interpretation whether it refers to this particular merger (and in such case makes an *composition* relation between *companies* and *A*) or it refers to a general statement which makes *A* an instance of companies referred to in the subsequent statement. Such cases are a frequent cause of disagreement in our annotation.

The data shows numerous coreferential links which are reported as other since only non-coreferential relations should be present in the annotated set. This can be explained with problems related to distinguishing other-than-coreferential relations from different linguistic means of expressing proper coreference, particularly in the initial phase of the annotation. A common observed mistake was treating mentions from indirect speech as non-coreferential with their direct speech equivalents — despite their identical reference targets.

Functional category calls for subclassification; several cases were commented as being best defined by WordNet's entailment relation (e.g. *to sleep* is entailed by *to snore*); a few others were marked as metonymy (e.g. *Ottawa* meaning *Canada*, also confused with a simple *part–whole* relation).

Temporal category needs to be confronted with Recasens' near-identity which defines more aspects of dissimilation.

5 Updated Version of the Typology

5.1 The Revised Model

Figure 2 presents the revised version of the typology of bridging relations based on findings from the annotation process. *Contextual dissimilation* can be used in cases when different realization or representation is being referred to in the process of refocusing (Fauconnier, 1994); *entailment* is mostly *effect* which corresponds to reason–cause relation (war–occupation, manure–smell, competition–result etc.) while function groups general role-casting relations such as place–inhabitant, writer–work etc.

Within the most coarse-grained and abundantly represented aggregation subclass several evident subcategories were identified: *collection*, *group* and *hyponymy–hypernymy*. Collections are ad hoc sets of generally unrelated objects, e.g. shopping items while elements of a group are related, e.g. members of the same organization. Hyponyms are collections of objects related by a common hypernym (e.g. animals vs. monkeys, elephants etc.)

Table 2 presents statistics of different relations ob-

served in the analyzed set (after adjudication and conversion of annotation results to the new typology).

Relation facet	Count
Structural	**122**
Aggregation	105
Collection	7
Group	63
Hyponymy	35
Composition	17
Class	**44**
Entailment	**14**
Effect	8
Function	6
Attribution	**13**
Analogical	**5**
Similarity	3
Contrast	2
Metareference	**3**
Dissimilation	**2**
Temporal	1
Contextual	1
Error	**52**
Coreference	17
Apposition	11
Predicate	9
Other	15

Table 2: Post-adjudication statistics of bridging relations.

5.2 Transitivity of Facets

An important aspect of referential associations which does not seem to be covered by existing approaches is transitivity of basic relations, i.e. ability to maintain a more distant but still decodable relation than just atomic link between a pair of referents. To illustrate the case, Example 8 shows a mixture of aggregation and composition: the link between a set and part of one element in the set is clear to understand yet reasonably complex: my sons → my son → my son's broken leg. Example 9 shows a similar mixture of functional relation and attribution.

(8)   *Moi synowie uwielbiają niebezpieczną jazdę na desce. Nawet złamana noga nie zniechęciła ich do startu w kolejnych zawodach.*
        'My sons love risky skateboarding. Even the broken leg did not discourage them from entering the next competition.'

(9)   *Oto Jean. Niektórzy mówią, przyszły prezydent Francji.*
        'This is Jean. Some say, the future president of France.'

6   Conclusions

The presented unified classification of bridging relations intends to be an initial step towards annotation of referential relations on a larger scale. The typology covers only relations available in existing models and preliminarily annotated data but several other aspects of referentiality should be verified against the corpus, e.g. the issue of definiteness, negation or natural ambiguity.

The experiment confirmed that clear identification of types of bridging relations is a difficult task, particularly when fine-grained distinctions are introduced. This leads to conclusion that shallow semantics is probably insufficient to describe such a complex phenomenon as reference. A new annotation guidelines taking into account discourse structure, lexical-semantic models and extra-linguistic knowledge are currently under preparation.

Acknowledgements

References

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Herbert H. Clark. 1975. Bridging. In *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP 1975, pages 169–174, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gilles Fauconnier. 1994. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press.

Christiane Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press, Cambridge, MA; London, May.

Claire Gardent, Hélène Manuélian, and Eric Kow. 2003. Which bridges for bridging definite descriptions? In *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, pages 69–76.

GCBT: Greek Coreference & Bridging Team. 2014. Coreference & Bridging Annotation Guidelines.

Madelyn Anne Iris. 1988. Problems of the Part-Whole Relation. In Martha Evens, editor, *Relational Models of the Lexicon*, pages 261–288. Cambridge University Press, New York, NY, USA.

Matthias Irmer. 2010. *Bridging Inferences in Discourse Interpretation*. Ph.D. thesis, University of Leipzig.

Georges Kleiber. 1999. Associative anaphora and part-whole relationship: The condition of alienation and the principle of ontological congruence. *Journal of Pragmatics*, 31(3):339–362.

Sebastian Löbner. 1996. Definite associative anaphora. In S. Botley, editor, *Approaches to Discourse Anaphora: Proceedings of the Discourse Anaphora and Resolution Colloquium (DAARC 96)*, pages 28–39, Lancaster University.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, November.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Massimo Poesio and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 1170–1174, Marrakech, Morroco. European Language Resources Association.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving Bridging References in Unrestricted Text. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 1–6.

Massimo Poesio, Rodolfo Delmonte, Antonella Bristot, Luminita Chiran, and Sara Tonelli. 2004. The VENEX corpus of anaphora and deixis in spoken and written Italian. Unpublished manuscript. Available at `cswww.essex.ac.uk/poesio/publications/VENEX04.pdf`.

Massimo Poesio. 2000. The GNOME annotation scheme manual. Technical report, University of Essex, United Kingdom.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Syntax and semantics: Vol. 14. Radical Pragmatics*, pages 223–255. Academic Press, New York.

Marta Recasens, Antònia Martí, and Mariona Taulé. 2007. Where Anaphora and Coreference Meet. Annotation in the Spanish CESS-ECE Corpus. In *Proceedings of RANLP 2007, Borovets, Bulgaria*, pages 504–509.

Marta Recasens, Antònia Martí, and Constantin Orasan. 2012. Annotating Near-Identity from Coreference Disagreements. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Laure Vieu and Michel Aurnague. 2007. Part-of relations, functionality and dependence. In Michel Aurnague, Maya Hickmann, and Laure Vieu, editors, *The categorization of spatial entities in language and cognition*, volume 20 of *Human Cognitive Processing*, pages 307–336. John Benjamins Publishing Company.

Morton E. Winston, Roger Chaffin, and Douglas J. Herrman. 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11(4):417–444.

Šárka Zikánová, Eva Hajičová, Barbora Hladká, Pavlína Jínová, Jiří Mírovský, Anja Nedoluzhko, Lucie Poláková, Kateřina Rysová, Magdaléna Rysová, and Jan Václ. 2015. *Discourse and Coherence. From the Sentence Structure to Relations in Text*. Institute of Formal and Applied Linguistics.