Proceedings of the Eighth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation

isa-8

October 3–5, 2012

Pisa, Italy

University of Pisa, Faculty of Foreign Languages and Literatures

and Istituto di Linguistica Computazionale “Antonio Zampolli”

Harry Bunt, editor

Table of Contents:

Preface

This slender volume contains the accepted long and short papers that were submitted to the Eigth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, isa-8, which was organized in Pisa, Italy, October 3-5, 2012.

isa-8 is the eigth edition of joint workshops on the International Organization for Standards ISO and the ACL Special Interest Group in Compuuttional Semantics, Working Group "The Representation of Multimodal Semantic Information (`http://sigsem.uvt.nl`). The isa workshops are often organized on the occasion of meetings of ISO projects concerned with the establishment of international standards for semantic annotation and representation. The main focus of these workshops is on the presentation and discussion of approaches, experiments, experiences, and proposals concerning the construction or application of interoperable linguistic resources with semantic annotations.

The isa-8 workshop co-occurs with meetings of several subprojects of the ISO project 24617, "Semantic annotation framework (SemAF)", in particular those concerned with the annotation of spatial information, the annotation of semantic roles, the annotation of discourse relations, and basic issues in semantic annotation.

I would like to thank the members of the isa-8 Programme Committee for their careful and quick reviewing, and the members of the isa-8 organizing committee for their wonderful support and cooperation, in particular Nicoletta Calzolari, Giovanna Marotta, Paola Baroni, Sara Goggi and Monica Monachini.

Harry Bunt
isa-8 chair

Programme Committee:

Jan Alexandersson
Harry Bunt (chair)
Thierry Declerck
Alex C. Fang
Robert Gaizauskas
Koti Hasida
Nancy Ide
Michael Kipp
Kiyong Lee
Alessandro Lenci
Inderjeet Mani
Martha Palmer
Volha Petukhova
Andrei Popescu-Belis
Rashmi Prasad
James Pustejovsky
Laurent Romary
Claudia Soria
Thorsten Trippel
Piek Vossen

Organizing Committee:

Harry Bunt (Tiburg University)
Kiyong Lee (Korea University, Seoul)
Nicoletta Calzolari (ILC-CNR, Pisa)
Kiyong Lee (Korea University, Seoul)
Giovana Marotti (University of Pisa)
Paola Baroni (ILC-CNR, Pisa)
Sara Goggi Pulman (ILC-CNR, Pisa)
Monica Monachini (ILC-CNR, Pisa)
James Pustejovsky (Brandeis University, Waldham, MA)
Laurent Romary (CNRS, Berlin)

"ISA in *P*isa"

8th Joint ACL-ISO Workshop on Interoperable Semantic Annotation

Workshop Programme

Wednesday, October 3, 2012
University of Pisa, Faculty of Foreign Languages and Literatures,
Aula Magna, Via Santa Maria 85

08:45 - 09:15	On-site registration
09:15 - 09:20	Welcome, opening
09:20 - 12:15	**Session: Semantic Relations in Discourse**
09:20 - 09:55	Sandrine Zufferey, Liesbeth Degand, Andrei Popescu-Belis & Ted Sanders: *Empirical validations of multilingual annotation schemes for discourse relations*
09:55 - 10:15	Ted Sanders, Kirsten Vis & Daan Broeder: *Project notes CLARIN project DiscAn*
10:15 - 10:35	coffee break
10:35 - 11:10	Silvia Pareti: *The Independent Encoding of Attribution Relations*
11:10 - 12:15	Project ISO 24617-8: Semantic Relations in Discourse **Discussion of results of NP ballot and comments on Working Draft** (Rashmi Prasad and Harry Bunt)
12:15 - 14:00	lunch break
14:00 - 15:30	**Session: Project ISO 24617-7: Spatial Information ("ISO-Space")** Discussion of results of NP ballot and comments on Working Draft (James Pustejovsky)
15:30 - 15:50	tea break
15:50 - 17:00	**Session: Semantic Resources, Part 1**
15:50 - 16:20	Alex Chengyu Fang, Jing Cao, Harry Bunt & Xioyuao Liu: *The annotation of the Switchboard corpus with the new ISO standard for Dialogue Act Analysis*
16:20 - 16:50	Elisabetta Jezek: *Interfacing typing and role constraints in annotation*
16:50 - 17:15	General discussion

Thursday, October 34, 2012
University of Pisa, Faculty of Foreign Languages and Literatures,
Aula Magna, Via Santa Maria 85

09:00 - 12:15	**Session: Semantic Role Annotation**
09:00 - 09:35	Susan Windisch Brown & Martha Palmer *Semantic Annotation of Metaphorical Verbs Using VerbNet: A Case Study of 'Climb' and 'Poison'*
09:35 - 10:00	Johan Bos, Kilian Evang & Malvina Nissim : *Semantic role annotation in a lexicalised grammar environment*
10:00 - 10:35	Weston Feely, Claire Bonial & Martha Palmer: *Evaluating the coverage of VerbNet*
10:35 - 10:55	coffee break
10:55 - 12:15	Project ISO 24617-4: Semantic Roles **Discussion of results of NP ballot and comments on Working Draft** (Martha Palmer)
12:15 - 14:00	lunch break
14:00 - 15:15	**Session: Semantic Resources, Part 2**
14:00 - 14:25	Aina Peris & Mariona Taule: *IARG AnCora: Annotating the AnCora Corpus with Implicit Arguments*
14:25 - 15:00	Massimo Moneglia, Gloria Gagliardi, Alessandro Panunzi, Francesca Frontini, Irene Russo, and Monica Monachini: *IMAGACT: Deriving an Action Ontology from Spoken Corpora*
15:00 - 16:35	**Session: Basic Issues in Semantic Annotation**
15:00 - 15:25	Camilo Thorne: *Studying the Distribution of Fragments of English Using Deep Semantic Annotation*
15:25 - 15:45	tea break
15:45 - 16:00	Project ISO 24617-6: "ISO-Basics", Status Report (Harry Bunt)
16:00 - 16:35	Kiyong Lee and Harry Bunt: *Counting Time and Events*
16:35 - 16:40	Status report on new initiative concerning the annotation of veridicality (on behalf of Annie Zaenen)
16:45 - 17:15	ISO TC 37/SC 4/WG 2 Plenary meeting Chair: Kiyong Lee

Friday, October 5, 2012
Istituto di Linguistica Computazionale "Antonio Zampolli",
Via Giuseppe Moruzzi 1, Pisa

09:00 - 16:45	**Session: Web Service Exchange Protocols (ISO PWI 24612-2**
09:00 - 09:30	Overview (Nancy Ide)
09:30 - 09:50	Donghui Lin: Language Grid Service Ontology
09:50 - 10:10	Alessio Bosca, Milen Kouylekov and Marco Trevisan (for Luca Dini): LinguaGrid exchange protocols
10:10 - 10:30	Núria Bel: PANACEA project protocols
10:30 - 11:00	coffee break
11:00 - 11:20	Bernardo Magnini: EXCITEMENT project protocols
11:20 - 11:40	Carsten Schnober: ISO TC37/SC4/WG6 PNWI Corpus Query Lingua Franca
11:40 - 12:00	Nancy Ide for Richard Eckart de Castilho: UIMA
12:00 - 12:15	Steve Cassidy: AusNC/DADA (video presentation)
12:15 - 12:30	Sebastian Hellman: NLP Interchange Format (video presentation)
12:30 - 14:00	lunch break
14:00 - 15:30	Group discussion following presentations: Are there commonalities, clear directions,..? Generation of outline of content for the ISO New Work Item (NWI)
15:30 - 15:45	tea break
15:45 - 16:45	Finalization of NWI proposal outline and next steps
17:00	Closing

Annotating semantic roles in a lexicalised grammar environment

Johan Bos
CLCG
University of Groningen
The Netherlands
johan.bos@rug.nl

Kilian Evang
CLCG
University of Groningen
The Netherlands
k.evang@rug.nl

Malvina Nissim
Linguistics and Oriental Studies
University of Bologna
Bologna, Italy
malvina.nissim@unibo.it

Abstract

Annotating text with abstract information such as semantic roles is costly. In previous efforts, such as PropBank, this process was aided with the help of syntactic trees, manually correcting automatically produced annotations. We argue that when using a lexicalised approach the annotation effort can be made simpler, avoiding the need to explicitly select two entities for each role. Our model is demonstrated by the Groningen Meaning Bank, using Combinatory Categorial Grammar as syntactic formalism, and Discourse Representation Theory as a formal semantic backbone.

1 Introduction and background

Annotating thematic roles is a time-consuming business: given an annotation scheme, for each role two entities need to be identified in the text, and the relation between them selected. This is often carried out with the help of syntactic trees and complex annotation aids. Perhaps this process can be made easier if it is considered as part of a larger semantically-oriented annotation effort. In this paper we argue that this is indeed the case.

Viewed from a simple but global perspective, annotation of thematic roles could be carried out on the surface (token) level, syntactic level, or semantic level. Perhaps, intuitively speaking, annotating semantic roles should take place at the semantic level (a logical form of some kind), because that's eventually where semantic roles belong. But reading and editing logical forms can be hard and requires extensive training for non-semanticists. Human annotation on the surface level, on the other hand, seems attractive but turns out to be a tiresome process without the aid of part-of-speech and requires sophisticated tools to select entities and specify relations between them.

There has been ample interest in semantic roles recently in the Natural Language Processing community. The main resource encoding subcategorisation frames and semantic roles over verb classes is VerbNet (Kipper Schuler, 2005). FrameNet (Baker et al., 1998) also encodes semantic roles, and it does so at a more detailed level than VerbNet, including adjuncts too, but has a much more limited coverage. NomBank (Meyers et al., 2004) provides semantic roles for nouns rather than verbs. The primary corpus annotated for semantic roles is PropBank (Palmer et al., 2005), which was annotated by hand-correcting the output of a rule-based tagger over constituency-based syntactic trees.

The evident need for joint modelling of syntactic dependencies and semantic roles has prompted a revision of PropBank for the CoNLL-2008 Shared Task on "Joint Parsing of Syntactic and Semantic Dependencies" (Surdeanu et al., 2008). One extension is the annotation of roles for the arguments of nouns as well, exploiting NomBank. The other, major, amendment is the translation of the original constituent-based structures into dependency-based ones, as a dependency grammar framework is believed to model more appropriately the syntax-semantics interface for the annotation of semantic roles (Johansson and Nugues, 2008).

Our claim is that the annotation of semantic roles is best done with the help of a *lexicalised grammati-*

cal framework. In a lexicalised grammar, verbs (and nouns) encode all their arguments inside their lexical category. This has some pleasant consequences: tokens can be easily divided into those that trigger (a finite, ordered set of) semantic roles and those that do not. Annotation then boils down to assigning the correct roles to each token. There is no need to select entities. Roles can be derived from existing resources such as VerbNet and FrameNet, depending on the desired granularity and taking into account coverage issues.

Thus, we propose a strongly lexicalised model where roles are assigned to verbs and modifiers, deriving them from external resources, and are subsequently inherited by the arguments and adjuncts directly through syntactic composition. Our experiments are implemented as part of the Groningen Meaning Bank (GMB, henceforth), a project that aims to annotate texts with formal semantic representations (Basile et al., 2012). The syntactic formalism used in the GMB is Combinatory Categorial Grammar (CCG), a lexicalised framework where syntactic categories are composed out of a few base categories (S, NP, N, PP), and slashes of complex categories indicate the direction of arguments (e.g., S\NP is a complex category looking for an noun phrase on its left to complete a sentence). The semantic formalism adopted by the GMB is Discourse Representation Theory, with a neo-Davidsonian view on event semantics.

2 Annotation Model

Semantic relations are relations between two entities, of which one is the internal and one the external entity. In the GMB semantic relations are two-place relations between discourse referents. The internal entity is usually an event, triggered by a verb; the external entity is usually triggered by a noun phrase. External entities are realised by arguments or adjuncts – annotation of roles differs with respect to whether external entities are arguments or adjuncts.

We will outline our model using the VerbNet inventory of roles for the verb *to build*. Let's first consider the annotation of roles whose external entities are introduced by arguments. In the GMB corpus various CCG categories are assigned to *build*, corresponding to different subcategorisation frames. The verb *build* is listed in two VerbNet classes: build-26.1-1 (WordNet sense 1); base-97.1 (WordNet sense 8).

Table 1 shows that *build* could be mapped to (at least) seven different VerbNet frames. However, the different CCG categories assigned to *build* already aid in disambiguating: the intransitive form S\NP maps to one VerbNet frame, the transitive form (S\NP)/NP to just three of the possible seven VerbNet frames. Whenever a CCG-category for a given verb could be mapped to more than one VerbNet frame, annotators will be presented with the relevant *roleset* (Palmer et al., 2005), i.e. the set of available role values to choose from associated to that verb usage. In the case of (S\NP)/NP, for example, Agent, Material, or Asset could be selected for the subject NP, while the object would be Product in any case.

The last column of Table 1 shows how the VN roles are inserted in the CCG categories. This, in turn, allows us to introduce the roles in the lexical DRSs for the verb. For instance, the lexical entry for the transitive form of *build* is illustrated in Figure 1. Note that VerbNet also provides the WordNet sense of a verb. This is also included in the lexical DRS as part of the symbol representing the building event (build-1). See Section 3 for the way WordNet senses can be used in the model.

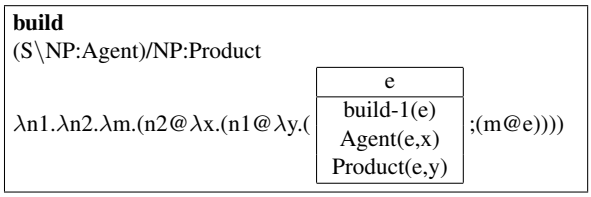

Figure 1: Lexical DRS for *build*.

CCG categories corresponding to passive verb forms lack the subject NP of the corresponding active forms. Active forms are distinguished by passive forms by features on the S category. In order to map passive CCG categories to VN entries one needs to bear in mind the correspondences below:

S_{pss}\NP:**X** ⇔ (S\NP:**Y**)/NP:**X**

(S_{pss}\NP:**Z**)/PP:**Y** ⇔ ((S\NP:**X**)/PP:**Y**)/NP:**Z**

This is how roles are assigned to arguments in the annotation model. For the roles that are introduced

Table 1: Mapping VerbNet roles to CCG categories, for *build*.

Category	Class	Sense	VerbNet frame	Enhanced CCG category
S\NP	build-26.1	1	Agent V	S\NP:agent
(S\NP)/NP	build-26.1	1	Agent V Product	(S\NP:agent)/NP:product
	build-26.1	1	Material V Product	(S\NP:material)/NP:product
	build-26.1-1	1	Asset V Product	(S\NP:asset)/NP:product
((S\NP)/PP)/NP	build-26.1	1	Agent V Product {from} Material	((S\NP:agent)/PP:material)/NP:product
	build-26.1-1	1	Agent V Product {for} Asset	((S\NP:agent)/PP:asset)/NP:product
	base-97.1	8	Agent V Theme {on} Source	((S\NP:agent)/PP:source)/NP:theme

by adjuncts we need a different strategy. In CCG, adjuncts are represented by categories of the form X/X or X\X, where X is any CCG category, possibly enhanced with further subcategorisation information (for instance in the case of prepositions). This will allow us to assign roles at the token level. This idea is shown in Figure 2 for a preposition (VP modifier).

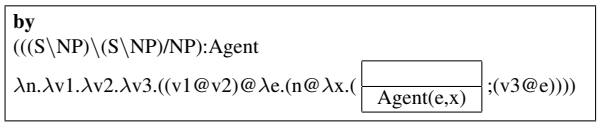

Figure 2: Lexical DRS for *by*.

It is important to see that, in this annotation model, semantic roles are annotated at the token level. Given a set of tokens corresponding to a sentence, each token is associated with an ordered, possibly empty, set of tokens. The number of elements in this set is determined by the CCG category. Categories corresponding to adjuncts introduce one role, the number of roles for categories associated with verbs is determined by the number of arguments encoded in the CCG category. This makes annotation not only easier, it also makes it more flexible, because one could even annotate correct roles for a clause whose syntactic analysis is incorrect.

3 Implementation

The GMB implements a layered approach to annotation. On the token level, there are separate layers, each with its own tag-set, for part-of-speech, named entities, numeral expressions, lexical categories, word senses, among others (Figure 3). These layers all contribute to the construction of the semantic representation of the sentence, and eventually that of a text, in the form a DRS. For semantic roles of VerbNet a further annotation layer is

The	**contractor**	**builds**	**houses**	**for**	**$100,000**
DT	NN	VBZ	NNS	IN	CD
0	1	1	1		0
NP/N	N	((S\NP)/PP)/NP	NP	PP/NP	NP
[ ]	[ ]	[Agent,Product]	[ ]	Asset	[ ]

DRS:

x e y z			
contractor(x)	houses(y)	$100,000(z)	
build-1(e)	Agent(e,x)	Product(e,y)	Asset(e,z)

Figure 3: Annotation layers in the GMB and corresponding semantic representation.

added. Note that for different inventory of roles, such as FrameNet, a further annotation layer could be included (Bos and Nissim, 2008). As we have shown in the previous section, the roles turn up in the DRS for the sentence, following the compositional semantics determined by the syntactic analysis, as two-place relation between two discourse referents (see Figure 3).

The manual annotation could be performed in three possible modes. The *open* mode lets the annotator choose from all possible VerbNet frames available for a given verb. In a *restricted* mode, the annotator can choose to activate specific constraints which limit the number of frames to choose from. For example, by activating the constraint relative to the syntactic category of the verb, for instance (S\NP)/NP, the annotator could reduce the number of possible frames for *to build* from seven to just three (see Table 1). Another constraint could be the WordNet sense: in the GMB, verb sense disambiguation is dealt with by a separate layer using the senses of WordNet, and WordNet senses are also used in VerbNet. Using the WordNet constraint, only VerbNet frames associated to a given WordNet sense would be available to choose from. For

example, if sense 8 of *to build* is selected there is only one option available (see Table 1). Alternatively, the WordNet sense could be used for detecting a possible error — for example if "source" is used in combination with sense 1 of *to build*, a warning should be issued as "source" can only be used with sense 8. In the *automatic* mode, the system will produce the annotation automatically on the basis of the correspondences and constraints which we have described, and the human annotator will be able to subsequently amend it through the GMB annotation interface. Whenever there is more than one option, such as assigning the appropriate VerbNet frame to an instance of build with category (S\NP)/NP, choice strategies must be devised (see Section 4).

4 Further Issues

There are a couple of further issues that need to be addressed. First, the choice of roleset depends on the sense assigned to a verb (or noun). In the GMB, word senses and roles are implemented by two different annotation layers. The question remains whether to permit inconsistencies (supported by a system of warnings that notices the annotator might such contradictions arise) or instead implement a system that constrains the choice of roleset on the basis of the selected word sense.

As we have seen, and as it is also noted by (Palmer et al., 2005), the same verb can be listed more than once with the same subcategorisation frame to which are however associated different roles. While in *open* and *restricted* modes the annotator will select the appropriate one, in *automatic* mode decision strategies must be devised. Another issue is to do with missing frames in VerbNet, such as for *build-8* with a NP V PP structure as in "He also seeks to build on improvements". An appropriate frame, such as Agent V Theme or Agent V Source, does not exist in VerbNet for *to build*, unlike e.g. for *to rely*. To address such cases, the interface should also let annotators choose from the whole inventory of VerbNet frames.

In the CoNLL 2008 shared task, data from NomBank is integrated with PropBank to get a wider range of arguments to be annotated for semantic roles, including thus nouns beside verbs. The lexicalised framework we have presented here can easily be extended to cover NomBank data as well.

Finally, this annotation model also has consequences for predicting semantic roles by machines. This is because, in a lexicalised framework such as the one that we propose, the process of semantic role labelling is essentially transformed to a classification task on tokens. Whether this could lead to better performance in semantic role labelling is a question left for future research.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, pages 86–90. ACL.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3196–3200, Istanbul.

Johan Bos and Malvina Nissim. 2008. Combining Discourse Representation Theory with FrameNet. In R. Rossini Favretti, editor, *Frames, Corpora, and Knowledge Representation*, pages 169–183. Bononia University Press, Bologna.

Richard Johansson and Pierre Nugues. 2008. Dependency-based semantic role labeling of propbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 69–78, Stroudsburg, PA, USA. ACL.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA. ACL.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England, August.

The Annotation of the Switchboard Corpus with the New ISO Standard for Dialogue Act Analysis

Alex C. Fang
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Hong Kong SAR
acfang@cityu.edu.hk

Jing Cao
College of Foreign Languages
Zhongnan University of Economics and Law
Wuhan, China
c_jinhk@yahoo.cn

Harry Bunt
Tilburg Center for Cognition and Communication
Tilburg University
The Netherland
harry.bunt@uvt.nl

Xiaoyue Liu
The Dialogue Systems Group
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Hong Kong SAR
xyliu0@cityu.edu.hk

Abstract

This paper is the description of a semantic annotation project that aims at the re-annotation of the Switchboard Corpus, previously annotated with the SWBD-DAMSL scheme, according to a new international standard for dialogue act analysis. A major objective is to evaluate, empirically, the applicability of the new ISO standard through the construction of an interoperable language resource that will eventually help evaluate the pros and cons of different annotation schemes. In this paper, we shall provide an account of the various aspects of the annotation project, especially in terms of the conversion between the two analytical systems, including those that can be fully automated and those that have to be manually inspected and validated. A second objective is to provide some basic descriptive statistics about the newly annotated corpus with a view to characterize the new annotation scheme in comparison with the SWBD-DAMSL scheme.

1 Introduction

The Switchboard Corpus is a valuable language resource for the study of telephone conversations. The Switchboard Dialogue Act Corpus, which is distributed by the Linguistic Data Consortium (LDC) and available online at http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2001T61, provides extensive added value because of its annotation of the component utterances according to an adapted DAMSL scheme for dialogue act (DA) analysis. More recently, the NXT-format Switchboard Corpus has been created (Calhoun et al. 2010). It combines orthographic transcriptions with annotations for dialogue act, syntax, focus/contrast, animacy, information status, and coreference in addition to prosodic and phonemic markings.

This paper describes a new development in the annotation of the Switchboard Dialogue Act Corpus. In this new version, each component utterance has been additionally annotated according to a new international standard, namely, ISO 64217-2:2012 (Bunt et al. 2010, 2012; ISO 2012). A major objective for the re-annotation of the corpus is to produce a new language resource where the same linguistic material is annotated according to two different schemes in order to facilitate a comparative study of different analytical frameworks. A second major objective is to verify the applicability of the new international standard through the practical annotation of authentic data and also to verify if the new scheme represents theoretical and practical advancement in real terms.

The basic principles for the project include the following:

1) The new DA scheme should be empirically applicable to a corpus of authentic conversations.
2) The re-annotation of the corpus should be realized by converting as much as possible

from its previous annotation in order to retain maximal data reliability.

3) The conversion should be optimized for automatic conversion, and manual mapping should be applied only when necessary.

A direct outcome for the project is a new language resource, which comprises transcribed real-life conversations and two different sets of DA annotations. As Figure 1 indicates, such a resource is especially well suited for comparative studies of DA annotation schemes and also in-depth investigation of the corpus through parallel annotations according to different schemes. As far as we know, such a resource is the first of its kind in the area of dialogue act analysis.

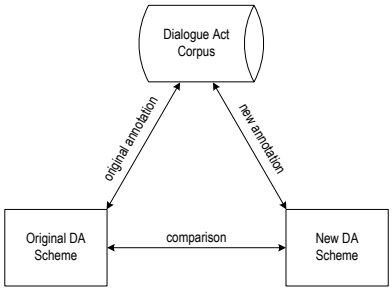

Figure 1: A new resource for DA research

The rest of this paper will describe how the original SWBD-DAMSL scheme has been converted to the new ISO DA standard. It will also describe the newly constructed corpus, including the ISO DA tags and their dimensions. The paper will then discuss some of the issues in the conversion and outline some future work.

2 SWBD-DAMSL

SWBD-DAMSL is a version of DAMSL (Dialogue Act Markup in Several Layers; Allen and Core 1997) that was specially adapted for the annotation of the Switchboard Corpus. The SWBD-DAMSL scheme consists of 220 DA types and has facilitated past studies such as Jurafsky et al. (1997) and Stolcke et al. (2000).[1]

To follow the practice of standoff markup, the original 1,155 annotated telephone conversations were re-processed, each slash-unit was coded and the utterance and its corresponding DA tag were separated and stored in individual files. Consider Example (1) below extracted from the file named `sw_0052_4378.utt`.

(1) sd B.7 utt1: *{C And,} {F uh,} <inhaling> we've done <sigh> lots to it. /*

Such an utterance, which is annotated as `sd` (*statement-non-opinion*), resulted in two files, where `SBD` stands for SWBD-DAMSL:

File 1: `sw00-0052-0010-B007-01.utt`
Content: *{C And,} {F uh,} <inhaling> we've done <sigh> lots to it.*
File 2: `sw00-0052-0010-B007-01-SBD.da`
Content: `sd`

As a general rule, the transcribed utterance is stored in a file with the `.utt` suffix and its SWBD-DAMSL tag in `*-SBD.da`. Similarly, `*-ISO.da` represents the set of files containing the ISO DA tags and `*-ISO.di` their corresponding dimensions.

3 Conversion to the ISO DA Standard

The ISO scheme contains 56 core DA tags, representing a tagset size comparable to that of the SWBD-DAMSL scheme of 59 combined tags. The tags are grouped according to 9 core dimensions and additionally described by a number of qualifiers designed to provide additional information about subtleties of communication functions. To maximally facilitate the conversion from SWBD-DAMSL to SWBD-ISO, four types of relation between the SWBD-DAMSL scheme and the ISO scheme were identified, namely, exact matches, many-to-one matches, one-to-many matches and unique SWBD-DAMSL tags. In the project, we performed the first two types of conversions automatically, and the one-to-many conversion was mapped manually. The treatment of the last group of tags, i.e., those unique to SWBD-DAMSL, will be discussed in section 3.4.

3.1 Automatic Mapping

The automatic mapping was performed on exact matches and many-to-one matches between the two schemes. In this process, 46 SWBD-DAMSL tags were matched to 22 ISO DA types, with a

[1] It should be pointed out that for the sake of enough instances, some original SWBD-DAMSL DA types have been combined together, which resulted in 42 different DA types in Jurafsky et al. (1997). The current study uses the 59 DA tags in Fang et al. (2011).

total number of 187,768 utterances, or 83.97% of the corpus, which accounts for 94.29% of the whole corpus in terms of tokens.

3.2 Manual Mapping

Six SWBD-DAMSL DA types were observed to have multiple destinations in the ISO scheme. These include `accept`, `accept-part`, `reject`, `reject-part`, `action directive`, and `other answer`. A user-friendly GUI was specially constructed and all the utterances concerned manually inspected and assigned an ISO tag.

For this task, three postgraduate students majoring in linguistics were invited to perform the annotation. They were provided with the manual of SWBD-DAMSL and the ISO standard. The training session included three phases: First, the annotators got familiar with the two DA schemes through trail annotation of 2 files for each of the six DAs. During the second phase, supervised annotation was carried out with 10 additional files for each DA. Finally, unsupervised annotation was conducted with another set of 10 files for each DA, and the inter-annotator agreement test was calculated based on the unsupervised samples.

Results show that in most cases a predominant ISO DA type could be identified. In some cases, an annotator favoured just one particular ISO DA type, which creates the bias and prevalence problems for the calculation of the kappa value (e.g. Di Eugenio and Glass, 2004). To solve this problem, the prevalence-adjusted and bias-adjusted kappa (PABAK) was proposed by Byrt et al. (1993) and used in quite a few past studies such as Sim and Wright (2005), Chen et al. (2009), Cunningham (2009) and Hallgren (2012). The adjusted kappa is defined as:

$$PABAK = \frac{kP_{obs} - 1}{k - 1}$$

where k is the number of categories and P_{obs} the proportion of observed agreement. At the end of the training session, PABAK was calculated pair-wised and the mean was taken as the final result. The average PABAK value is 0.69. According to Landis and Koch (1977), the agreement between the three annotators is substantial and therefore judged acceptable for subsequent manual annotation.

The actual manual annotation saw six SWBD-DAMSL tags mapped to 26 different ISO DA tags, which involves 12,837 utterances (i.e. 5.74% of the corpus) and covers 2.03% of the corpus in terms of tokens.

Altogether, through both automatic and manual annotation, 200,605 utterances in the corpus (i.e. 89.71%) were treated with ISO DA tags. Table 1 presents the basic statistics of the SWBD corpus annotated with the ISO DA scheme, including the types of ISO DAs, the number of utterances and tokens, and their corresponding percentage and accumulative percentage. The ISO DA types are arranged according to the number of utterances in descending order.

ISO DA Type	Utterance			Token		
	#	%	Cum%	#	%	Cum%
inform	120227	53.767	53.77	1266791	82.962	82.96
autoPositive	46382	20.743	74.51	66506	4.355	87.32
agreement	10934	4.890	79.40	20598	1.349	88.67
propositionalQuestion	5896	2.637	82.04	39604	2.594	91.26
confirm	3115	1.393	83.43	3698	0.242	91.50
initialGoodbye	2661	1.190	84.62	9442	0.618	92.12
setQuestion	2174	0.972	85.59	15841	1.037	93.16
disconfirm	1597	0.714	86.31	3392	0.222	93.38
answer	1522	0.681	86.99	8154	0.534	93.91
checkQuestion	1471	0.658	87.64	11053	0.724	94.64
completion	813	0.364	88.01	3188	0.209	94.85
question	680	0.304	88.31	5068	0.332	95.18
stalling	580	0.259	88.57	3004	0.197	95.37
choiceQuestion	506	0.226	88.80	4502	0.295	95.67
suggest	369	0.165	88.96	3320	0.217	95.89
autoNegative	307	0.137	89.10	798	0.052	95.94
request	278	0.124	89.22	1644	0.108	96.05
disagreement	258	0.115	89.34	689	0.045	96.09
acceptApology	112	0.050	89.39	366	0.024	96.12
instruct	106	0.047	89.44	961	0.063	96.18
acceptSuggest	99	0.044	89.48	195	0.013	96.19
apology	79	0.035	89.52	317	0.021	96.21
thanking	79	0.035	89.55	221	0.014	96.23
offer	71	0.032	89.58	590	0.039	96.27
acceptRequest	65	0.029	89.61	96	0.006	96.27
signalSpeakingError	56	0.025	89.64	75	0.005	96.28
promise	41	0.018	89.66	279	0.018	96.30
correction	29	0.013	89.67	210	0.014	96.31
acceptOffer	26	0.012	89.68	40	0.003	96.31
turnTake	18	0.008	89.69	28	0.002	96.31
alloPositive	17	0.008	89.70	21	0.001	96.31
correctMisspeaking	14	0.006	89.70	38	0.002	96.32
selfCorrection	8	0.004	89.71	41	0.003	96.32
acceptThanking	6	0.003	89.71	6	0.000	96.32
declineOffer	3	0.001	89.71	5	0.000	96.32
declineRequest	3	0.001	89.71	3	0.000	96.32
turnRelease	2	0.001	89.71	2	0.000	96.32
declineSuggest	1	0.000	89.71	1	0.000	96.32
other	23001	10.29	100.00	56175	3.679	100.00
Total	223606	100.00		1526962	100.00	

Table 1: Basic stats of the SWBD-ISO corpus

Other in Table 1 glosses together all the SWBD-DAMSL tags that cannot be matched to the ISO DA scheme. These represent 10.29% of the total

number of utterances in the corpus or 3.679% of all the tokens. They will be discussed in detail later in Section 3.4.

3.3 Dimensions

A feature of the ISO DA standard is that each utterance is also marked with dimension information. Consider Example (1) again. According to the ISO annotation scheme, it is annotated with the DA type *inform*, which belongs to the ISO dimension of *Task*. As a matter of fact, out of the nine ISO dimensions, eight are identified in the newly created SWBD-ISO corpus except for the dimension of *Discourse Structuring*.[2] Table 2 lists the eight dimensions and their corresponding ISO DA types, together with the percentage of the utterances they cover. Note that only those DA types observed in the corpus are listed in the table. The DA tag `alloNegative`, for instance, is missing from Table 2 since the corpus does not contain any utterance analysed as such. *Other** in Table 2 actually refers to the portion of utterances in the corpus that do not have an appropriate ISO DA tag and hence no dimension information. According to the table, those account for 10.29% of the total number of utterances in the corpus. The original SWBD-DAMSL analysis of the utterances is described in detail in Section 3.4 below and summarised in Table 3.

ISO Dimension	%	ISO DA Type
Task	66.85	inform; agreement; propositionalQuestion; confirm; setQuestion; disconfirm; answer; checkQuestion; question; choiceQuestion; suggest; request; disagreement; instruct; acceptSuggest; offer; acceptRequest; promise; correction; acceptOffer; declineOffer; declineRequest; declineSuggest
Auto-Feedback	20.88	autoPositive; autoNegative
Social Obligations Management	1.31	initialGoodbye; acceptApology; apology; thanking; acceptThanking
Time Management	1.19	stalling
Partner Communication Management	0.37	completion; correctMisspeaking
Own Communication Management	0.03	signalSpeakingError; selfCorrection
Allo-Feedback	0.01	alloPositive
Turn Management	0.01	turnTake; turnRelease
*Other**	*10.29*	**See Table 3 for a detailed breakdown*
Total	100.00	

Table 2: Basic stats for ISO dimensions

[2] In the current project, the dimension of *Discourse Structuring* is not explicitly treated since it most often overlaps with the more general *Task* dimension.

In addition, a particular feature of the ISO standard for DA annotation is that an utterance can be associated with more than one dimension, known as multi-dimensionality of DA. Example (1) has two dimensions, namely, *Task* and *Time Management*, for which the following files would be created:

File 3: `sw00-0052-0010-B007-01-ISO-21.da`
Content: `inform`
File 4: `sw00-0052-0010-B007-01-ISO-21.di`
Content: `task`
File 5: `sw00-0052-0010-B007-01-ISO-22.da`
Content: `stalling`
File 6: `sw00-0052-0010-B007-01-ISO-22.di`
Content: `timeManagement`

In our annotation scheme, `.da` files contain the name of the ISO DA types, while `.di` the name of the ISO dimensions. The first digit following `ISO-` (i.e. 2 in file names above) indicates the number of dimensions that a certain utterance is contextually associated with, while the second digit indicates the current number in the series.

Of the 200,605 mapped utterances, 144,909 utterances are annotated with one dimension, 44,749 with 2 dimensions and 10,947 with 3 dimensions.

3.4 Unmatched SWBD-DAMSL Tags

The conversion process left 13 SWBD-DAMSL tags unmatched to the ISO scheme. They account for 23,001 utterances and 56,175 tokens, representing respectively 10.29% and 3.68% of the corpus. See Table 3 for the basic statistics.

SWBD-DAMSL Tag	Utterance			Token		
	#	%	Cum%	#	%	Cum%
abandoned	12986	5.81	5.81	35363	2.32	2.32
non-verbal	3730	1.67	7.48	77	0.01	2.33
uninterpretable	3131	1.40	8.88	5729	0.38	2.71
quoted material	1058	0.47	9.35	8114	0.53	3.24
other	820	0.37	9.72	1603	0.10	3.34
transcription errors	649	0.29	10.01	3028	0.20	3.54
conventional opening	225	0.10	10.11	529	0.03	3.57
exclamation	136	0.06	10.17	282	0.02	3.59
3rd party talk	118	0.05	10.22	508	0.03	3.62
self talk	106	0.05	10.27	630	0.04	3.66
double quoted	27	0.01	10.28	189	0.01	3.67
explicit performative	9	0.00	10.28	81	0.01	3.68
other forward function	6	0.00	10.29	42	0.00	3.68
Total	23001	10.29		56175	3.68	

Table 3: Basic stats of unique SWBD tags

It is noticeable that a majority of these tags (e.g. *abandoned*, *non-verbal*, *uninterpretable*, and *quoted material*) are not defined on the basis of the communicative function of the utterance. Only two tags, i.e., *exclamation* and *explicit performative*, are clearly defined in functional terms and yet could not be matched to any of the DA types in the ISO standard.

3.5 Unmatched ISO Tags

An examination of the converted corpus has revealed that some ISO DA tags cannot be empirically observed in the corpus. See Table 4 for the specific ISO DA tags along with their corresponding dimensions.

ISO DA Type	ISO Dimension
addressRequest; addressSuggest; addressOffer	Task
alloNegative	Allo-Feedback
turnAccept; turnAssign; turnGrab; turnKeep	Turn Management
pausing	Time Management
interactionStructuring; opening	Discourse Structuring
initialGreeting; returnGreeting; initialSelfIntroduction; returnSelfIntroduction; returnGoodbye	Social Obligations Management
retraction	Own Communication Management

Table 4: DA Tags unique to ISO scheme

As is worth noting here, Table 4 should not be taken to suggest that the corpus does not contain any utterance that performs those communicative functions specified in the new ISO standard. Bear in mind that the ISO annotation of the corpus is achieved through mapping the original SWBD-DAMSL tags. Hence, the non-observation of the ISO tags listed in Table 4 only suggests that there is no direct mapping between the SWBD-DAMSL and ISO tagsets as far as these particular ones are concerned. Annotation of these unique ISO tags can be realized by considering the actual content of the utterances. Secondly, it should also be noted that the unmatched tags in the *Task* dimension include the mother nodes (e.g. *addressRequest*) of some more specific DAs (e.g. *acceptRequest* and *declineRequest*) and the utterances concerned have been annotated with the more specific daughter nodes as requested by the manual of annotation.

4 Conclusion

This paper described a project to re-annotate the SWBD DA corpus with the new ISO standard for DA analysis and reported some of the basic statistics concerning the conversion between SWBD-DAMSL and SWBD-ISO. A significant contribution of the current work is the creation of an interoperable language resource which can serve as the test-bed for the evaluation of different DA annotation schemes. The same resource can also be used for the exploration and verification of the contribution of different DA taxonomies to the automatic identification and classification of DAs. Our immediate future work will include a comparative study of the SWBD-DAMSL and ISO DA schemes. It is also expected that attempts will be made to address the treatment of the unmatched DA tags with a view how best to accommodate empirically encountered dialogue phenomena that were not considered in the drafting process of the standard. At the same time, we are performing some preliminary research to assess the performance of an automatic classifier of ISO dialogue acts with the specific intent to construct a DA model from the SWBD-ISO Corpus to be applied to other linguistic resources for dialogue studies. An issue that is of particular interest at this stage is the prospect of applying the ISO DA standard to dialogue resources in the Chinese language.

Acknowledgement

The project described in this article was supported in part by grants received from the General Research Fund of the Research Grants Council of the Hong Kong Special Administrative Region, China (RGC Project No. 142711) and City University of Hong Kong (Project Nos 7008002, 7008062, 9041694, 9610188, and 9610226). The authors would like to acknowledge the academic input and the technical support received from the members of the Dialogue Systems Group (http://dsg.ctl.cityu.edu.hk) based at the Department of Chinese, Translation and Linguistics, City University of Hong Kong.

References

Alex C. Fang, Harry Bunt, Jing Cao and Xiaoyue Liu. 2011. Relating the Semantics of Dialogue Acts to Linguistic Properties: A machine learning perspective through lexical cues. In Proceedings of the 5th IEEE International Conference on Semantic Computing, September 18-21, 2011, Stanford University, Palo Alto, CA, USA.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul

Taylor, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3): 339–371.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa Statistic: A Second Look. Journal of Computational Linguistics, 30(1): 95-101.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-discourse-function Annotation Coders Manual, Draft 13. University of Colorado, Boulder Institute of Cognitive Science Technical Report 97-02.

Guanmin Chen, Peter Faris, Brenda Hemmelgarn, Robin L. Walker, and Hude Quan. 2009. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. BMC Medical Research Methodology. 9: 5.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kyong Lee, Volha Petukhova, Andrel Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO Standard for Dialogue Act Annotation. In Proceedings of the Seventh International Conference on Language Resources and Evaluation. Valletta, MALTA, 17-23 May 2010.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrel Popescu-Belis, and David Traum. 2012. A Semantically-based Standard for Dialogue Annotation. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). Istanbul.

ISO Standard 24617-2. 2012. Language resource management – Semantic annotation framework (SemAF), Part 2: Dialogue acts. ISO, Geneva, 2012.

James Allen and Mark Core. 1997. DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1*)*. Technical Report, Multiparty Discourse Group. Discourse Resource Initiative, September/ October 1997.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1): 159-174.

Julius Sim and Chris C. Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy, 85 (3): 257-268.

Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology. 8(1): 23-34.

Michael Cunningham. 2009. More than Just the Kappa Coefficient: A Program to Fully Characterize Inter-Rater Reliability between Two Raters. SAS Global Forum 2009.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language Resources and Evaluation. 44(4): 387-419.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and Kappa. Journal of Clinical Epidemiology. 46(5): 423-429.

Evaluating the Coverage of VerbNet

Weston Feely
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213-3891
WFeely@cs.cmu.edu

Claire Bonial, Martha Palmer
Department of Linguistics,
University of Colorado at Boulder
Hellems 290, 295 UCB
Boulder, CO 80309-0295
{CBonial, MPalmer}@colorado.edu

Abstract

This research presents a comparison of the syntactic behavior of verbs represented in an online verb lexicon, VerbNet, and the actual behavior of the verbs in the SemLink corpus. To complete this comparison, each verbal instance of the SemLink corpus is reformulated into a syntactic frame, e.g. Noun Phrase – Verb – Noun Phrase, and compared to syntactic frames listed in VerbNet. Through this effort, the coverage and accuracy of VerbNet is extended with the addition of new syntactic frames and thematic roles such that VerbNet is a more complete reflection of language in use.

1 Introduction

VerbNet (VN) (Kipper et al., 2008) is an online verb lexicon that provides valuable information on the relational semantics of approximately 6200 English verbs. VN is an extension of Levin's (1993) classification, in which verbs are organized according to their compatibility with certain syntactic, or "diathesis," alternations. For example, the verb *break* can be used transitively (*Tony broke the window*) or intransitively (*The window broke*). This represents one diathesis alternation, and other verbs that share the ability to alternate between these two syntactic realizations could be classified with *break*. Although the primary basis of Levin's classification is syntactic, the verbs of a given class do share semantic regularities as well. Levin hypothesized that this stems from the fact that the syntactic behavior of a verb is largely determined by its meaning; thus, there is a fundamental assumption that syntactic behavior is a reflection of semantics.

VN has extended Levin's work and the lexicon has proved to be a valuable resource for various NLP applications, such as automatic semantic role labeling (Swier & Stevenson, 2004), semantic inferencing (Zaenen, 2008), and automatic verb classification (Joanis et al., 2007). However, the utility of VN relies heavily on its coverage and accurate representation of the behavior of English verbs. Although VN is theoretically motivated, the coverage and accuracy of the lexicon has not been comprehensively investigated, with the exception of examinations of VN's representation of certain syntactic constructions (Bonial et al., 2011c; Bonial et al., 2012). This work compares the representations of syntactic behavior found in VN to actual syntactic behaviors found in the SemLink corpus. There are two primary purposes of this comparison: 1) coverage: to what extent does VN capture all syntactic realizations of a given verb? 2) accuracy: to what extent is VN's syntactic representation an accurate reflection of realization possibilities and probabilities? The findings herein can be used to improve both coverage and accuracy, thereby improving the utility of VN overall.

2 Background

In order to evaluate VN's coverage, syntactic and semantic information available in the verb lexicon VN and the annotated corpus SemLink were compared. These two resources are described in the next sections.

2.1 VerbNet Background

Class membership in VN is based on a verb's compatibility with certain syntactic frames and alternations. For example, all of the verbs in the Spray class have the ability to alternate the Theme or Destination as a noun phrase (NP) object or as a

prepositional phrase (PP): *Jessica loaded the boxes into the wagon*, or *Jessica loaded the wagon with boxes*. VN's structure is somewhat hierarchical, comprised of superordinate and subordinate levels within each verb class. In the top level of each class, syntactic frames that are compatible with all verbs in the class are listed. In the lower levels, or "sub-classes," additional syntactic frames may be listed that are restricted to a limited number of members. In each class and sub-class, an effort is made to list all syntactic frames in which the verbs of that class can be grammatically realized. Each syntactic frame is detailed with the expected syntactic phrase type of each argument, thematic roles of arguments, and a semantic representation. For example:

Frame NP V NP PP.Destination
Example Jessica loaded boxes into the wagon.
Syntax Agent V Theme Destination
Semantics Motion(during(E), Theme)
Not(Prep-into (start(E), Theme, Destination))
Prep-into (end(E), Theme, Destination)
Cause(Agent, E)

2.2 SemLink Background

The SemLink corpus (Palmer, 2009; Loper et al., 2007) consists of 112,917 instances of the Wall Street Journal, each annotated with its corresponding VN class. Each instance is further annotated with PropBank (Palmer et al., 2005) arguments, which are numbered arguments that correspond to verb-specific roles. For example, these are the potential roles to be assigned for the verb *load*:

Roleset ID: load.01, *cause to be burdened*,
VN class: 9.7-2
Roles:
Arg0: loader, agent (VN role: 9.7-2-agent)
Arg1: beast of burden (VN role: 9.7-2-destination)
Arg2: cargo (VN role: 9.7-2-theme)
Arg3: instrument

Note that each verb sense, or "roleset," is mapped to its corresponding VN class, and each of the PropBank roles are mapped to VN thematic roles where possible. This roleset also demonstrates a sort of mismatch between PropBank and VN's treatment of *load*: PropBank treats the instrument as a numbered argument, whereas VN doesn't list an instrument as a semantic role for this verb.

Within the SemLink corpus, these mappings are made explicit such that with each instance, both PropBank and VN thematic roles are given for each argument. SemLink also contains mappings between PropBank rolesets, VN classes and FrameNet (Fillmore et al., 2002) frames, as well as corresponding mappings between PropBank arguments, VN thematic roles and FrameNet frame elements. Thus, SemLink is a resource created with the intent of allowing for interoperability amongst these resources.

2.3 Investigating VerbNet Using SemLink

The motivation for this project is to compare the set of syntactic frames listed in each VN class to the set of syntactic frames that actually occur in usage in the class's corresponding SemLink entries. Such a comparison is challenging because VN is a largely theoretical verb lexicon, which is still strongly rooted in Levin's original classification. SemLink, on the other hand, is an annotated corpus of real language in use, which often shows far more syntactic variability than assumed by theoretical linguistics. Thus, a comparison of VN with SemLink could provide a greater range of syntactic frames for most VN classes, simply because unexpected syntactic frames present themselves in the SemLink data.

This additional syntactic variation in the SemLink data should facilitate the primary goal of this project, which is to increase the coverage of VN's syntactic and semantic information. This is accomplished by using the empirically-derived information in the SemLink data to validate the class organization of VN by demonstrating which of VN's syntactic frames are present in the SemLink corpus for a given class, and which syntactic frames are present in the corpus that are not listed among the options for a given VN class. The additional syntactic frames detected can increase the coverage of each verb class's syntactic information, by augmenting each class's previous set of syntactic frames with empirically derived alternatives.

Additionally, the SemLink data will provide frequency information for syntactic frames, so that each syntactic frame in a VN class can be listed with how often it occurs in corpus data. This is especially important, because our empirical validation of the class organization of VN can be extended to: which syntactic frames are highly frequent in SemLink and present in a given VN class; which frames are highly frequent but missing from a given class; which frames are infrequent and present in a given class; and which frames are infrequent but missing from a given class.

3 Methods

The SemLink data for this project includes 70,270 SemLink instances, which are all the instances of the total 112,917 with a currently valid VN class assignment. Each of the SemLink instances included in the project data was processed for the necessary information to compare it to VN frames. This included the extraction of each SemLink instance's VN class assignment, the instance's PropBank roleset assignment, the syntactic frame from the Treebank parse, and the VN semantic roles for each constituent in the frame. After gathering this information from SemLink, frequencies were calculated for each syntactic frame type given its VN class assignment. The syntactic frames from SemLink were created using a Penn Treebank application-programming interface that automatically retrieved the syntactic constituents immediately dominating the part-of-speech tag for each of the words that were marked as arguments to the main verb in the SemLink instances. The rest of the information taken from SemLink was extracted directly from the SemLink Wall Street Journal annotations, using regular expressions.

The VN data for this project includes the frames (e.g. NP V NP) and corresponding semantic role argument structures (e.g. Agent V Theme) for all VN classes. These frames and argument structures were taken directly from the VN XML class files using regular expressions, with some small modifications to each frame. In order to facilitate matching with the SemLink frames, the constituents in each of VN's flat syntactic frames were stripped of additional tags, such as: redundant thematic roles (e.g. PP.Location; all roles are listed again in a separate line, e.g. Agent V Theme Location), syntactic alternation tags (e.g. NP-Dative), and other tags extraneous to the purpose at hand.

3.1 Frame Creation Method

The syntactic frames extracted from SemLink for this project were formed based on the linear order of syntactic constituents, as retrieved from the linear order of thematic role annotations in SemLink. In the case of arguments of the verb that were syntactically null elements, the last element in a movement chain was taken to form the frame, unless the null element was a passive or reduced relative clause marker, in which case the constituent one level above the trace constituent was taken. As an example, consider the following question: *Whom did she see?* In the Penn Treebank treatment of this sentence, there would be an object trace after *see* with an index indicating that the object trace corresponds to the question word *whom: Whom-1 did she see *T*-1?* The arguments identified for *see* would use the trace as the object position, resulting in the following frame: NP V NP, as opposed to the position of the realized constituents: NP NP V. In order to avoid interpreting passives as verb-initial frames, the passive and reduced relative constructions are treated differently and identified as such. Passives are currently excluded from this study as discussed in greater detail in Section 4.1.

3.2 Matching Conditions

After extracting the data from SemLink and VN, the data from each SemLink instance was matched against the set of [frame, argument structure] pairs in the corresponding VN class. This matching process was done using regular expressions in a three-step process.

First, the frame from the SemLink instance was checked against each of the frames in its corresponding VN class. If there was a match, the instance was counted as having matched a VN frame, and if the [VN class, frame] pair for this SemLink instance had not previously been matched, it was added to a list of frame types that matched VN. For example, consider the following SemLink instance, shown with its PropBank arguments and VN thematic role labels:

1. *The explosion of junk bonds and takeovers has…loaded corporations…with huge amounts of debt.*
Load, PropBank load.01, VN class Spray-9.7-2:
[The explosion of junk bonds and takeovers]$_{\text{ARG0, AGENT}}$ has…loaded$_{\text{RELATION}}$ [corporations…]$_{\text{ARG1, DESTINATION}}$ [with huge amounts of debt…]$_{\text{ARG2, THEME}}$.

This SemLink instance would be assigned the frame NP V NP PP, which matches a frame listed in its associated VN class:

Frame NP V NP.Destination PP.Theme
Example Jessica loaded the wagon with boxes.
Syntax Agent V Destination {with} Theme

Thus, this instance would be considered a frame match to VN.

Second, if the frame from the SemLink instance did not match any of the frames in the corresponding VN class, then the argument structure for the instance was checked against each of the argument structures in the corresponding VN class. If there was a match, the instance was counted as having matched VN, and if the [VN class, frame] pair for the SemLink instance had not

previously been matched, it was added to a different back-off list of frame types that matched VN. The following instance is an example of this type of match:

2. *It doesn't mean unanimous…*
Mean, PropBank mean.01, VN class Conjecture-29.5:
$It_{ARG0,\ AGENT}$ $does[n't]_{NEGATIVE}$ $mean_{RELATION}$ $unanimous_{ARG1,\ THEME}$…

This frame syntactically is of type NP V ADJP, and VN only represents Themes realized as NPs. Thus, this frame was matched via arguments (Agent V Theme) rather than syntactic frames. It was quite common for a SemLink instance to include an unexpected constituent type such as the ADJP here, and it is this constituent information that can be used to expand the constituent types for frames in VN, discussed in Section 5. This particular instance also brings to light a problematic aspect of the SemLink corpus and the interoperability between VN and PropBank: PropBank has much more coarse-grained rolesets or senses than those found in the VN classes. Thus, this roleset, which would include instances of the sense of intentional "meaning" found in the Conjecture class, also includes this sense of unintentional "meaning". As a result, "It" above is treated as an Agent, although the status as an Agent is questionable.

Third, if the frame and its argument structure from the SemLink instance did not match any of the frames in the corresponding VN class, it was added to a final list of frame types that did not match VN. Consider the following unmatched examples of the relation *remain,* which belongs to the VN class Exist-47.1:

3. *Like just about everything else, that remains to be seen.*
$[Like\ just\ about\ everything\ else,]_{ADVERBIAL}$ $that_{ARG1,\ THEME}$ $remains_{RELATION}$ $[to\ be\ seen]_{ARG3}$ – NP V S

4. *The crowd remained good-natured, even bemused.*
$[The\ crowd]_{ARG1}$ $remained_{RELATION}$ $[good\text{-}natured,\ even\ bemused]_{ARG3}$ – NP V ADJP

These examples demonstrate a potential gap in VN's representation of verbs like *remain* in the Exist class. While the PropBank argument structure includes an Arg3 role that corresponds to "attribute" arguments for more abstract usages of *remain*, the VN class contains only the roles Theme and Location, and did not include frames with sentential complements or adjective phrases that could capture these attributes. This suggests one way that VN can be improved based on this empirical investigation of verbal behavior: the addition of an attribute argument to the Exist class for abstract usages.

The end result of this matching process was three counters and three lists. The counters are the portion of the total SemLink instances that 1) matched a VN frame, 2) did not match a frame but did match a VN argument structure, or 3) did not match VN at all. These token counters were converted into token percentages in Table 1 in Section 4 below. The lists contain frame types for each matching condition: frame types that were in VN, frame types that had argument structures that were in VN, and frame types that were not in VN. These type lists were converted into type percentages in Table 3 in Section 4.

This matching process was repeated for three frequency subdivisions of the SemLink frame types: high frequency, middle frequency, and low frequency. These frequency categories were defined as the top 30%, middle 40%, and bottom 30% of the SemLink frame types for each VN class, ranked by frequency. For this second matching process using frequency information, the SemLink frames that matched VN by frame and by argument structure were combined into one category of frame types that matched VN. The SemLink frames that did not match VN by frame or argument structure were left in a separate category of frame types that did not match VN. In the same manner as the first matching process, the end result was a set of counters for the frame tokens that matched VN, and a set of lists for the frame types that matched VN, subdivided by these frequency categories. The percentages of the SemLink frame tokens for each of these frequency subdivisions are in Table 2 of Section 4, and the percentages of the SemLink frame types for each of these frequency subdivisions are in Table 4 of Section 4.

3.3 Loose Match

For the particular instances in SemLink that contained WH-movement or topicalization, looser matching criteria were used: there was a successful argument structure match when the set of argument roles matched any set of argument roles in the corresponding VN class (ordering was not considered). This was done because transformations like topicalization and WH-movement allow variable movement of syntactic constituents along the syntactic parse, so this separate matching condition that disregards the linear order of argument roles was needed. Because these transformations are possible for all verbs, they are not the type of distinctive syntactic alternations that VN lists. For example:

5. *"It's been a steadily improving relationship," says Mr. Carpenter.*
Say, PropBank say.01, VN class Say-37.7:
["It's been a steadily improving relationship"]-1 says$_{\text{RELATION}}$ [**Trace*-1]$_{\text{ARG1, TOPIC}}$ [Mr. Carpenter]$_{\text{ARG0, AGENT}}$

This instance was recognized as a syntactic frame of the type V S NP, which VN does not include in the Say class. Since the frame did not match, the instance was tested for an argument match: V Topic Agent. However, this argument structure is also not represented in VN for the Say class. Nonetheless, the loose match condition recognizes the topicalization transformation, and with instances containing such movement, allows for a match based on sets of arguments. Because the roles of Agent and Topic are present in the class and the transformation was recognized, this instance was considered a match.

3.4 Semantic Role Updates

After the frame retrieval process, it was also necessary to update the set of semantic roles in each SemLink instance. This is due to the fact that the SemLink Wall Street Journal annotations are currently outdated, and awaiting an update in the near future. However, at the time of this writing the SemLink data used for this project was created using an old set of VN roles that are not current with the 3.2 version of VN (for a description of the recent VN semantic role updates, see Bonial et al., 2011a, Bonial et al., 2011b). Therefore, before the frame matching process could begin, the semantic roles in the argument structures retrieved from SemLink had to be updated using a type-to-type mapping of old VN roles to new VN roles. This update was done automatically.

4 Findings

The results of the matching process are discussed in the following sections.

4.1 Passives

Passive sentences in the Wall Street Journal section of SemLink were removed from the matching process to be considered separately, since previous attempts to include passives in the matching process created the largest source of error for the project. This is due to the fact that VN does not include passive versions of its frames in the frame listing for each verb class. This omission is purposeful, because common syntactic transformations like passivization and WH-movement are not considered to be syntactic alternations distinctive of verb classes, following Levin's original verb classification. Passives made up 26.7% of the original data set of 70,270 instances, and after removing them a set of 51,534 frame tokens remained to be considered for the matching process. Passive frames were included in a separate list of frames, potentially to be used for future augmentation of VN.

4.2 Matches

SemLink tokens that...	% of total SemLink frame tokens (51534)
Matched a VN Frame	51.23%
Matched a VN Argument Structure	24.30%
Did not match corresponding VN class	24.46%

Table 1: Results of Matching Process for SemLink Frame Tokens

If we focus on tokens, we see that the majority of frame tokens in SemLink match frames in VN. However, this needs to be qualified because the matches are highly skewed towards the high frequency frame token matches. This is shown in the following table.

Match/No match grouping	Frequency	% of total SemLink frame tokens
Matched VN	High Frequency (top 30%)	54.49%
	Middle Frequency (middle 40%)	20.63%
	Low Frequency (bottom 30%)	0.41%
Did not match VN	High Frequency (top 30%)	17.67%
	Middle Frequency (middle 40%)	5.47%
	Low Frequency (bottom 30%)	1.32%

Table 2: Results of Matching Process for SemLink Frame Tokens, Divided by Frequency

This demonstrates that the most frequent frame tokens make up the majority of the frame token matches. This is because a small number of highly frequent frame types bias the token matches towards the high frequency match category. For example, 34% of all frame tokens are NP V NP, the most frequent frame type. Therefore, it is important to also consider the SemLink frame type matches, which are available in the following tables.

SemLink frame types that...	% of total SemLink frame types (3721)
Matched a VN Frame	12.92%
Matched a VN Argument Structure	20.29%
Did not match corresponding VN class	66.78%

Table 3: Results of Matching Process for SemLink Frame Types

Match/No match grouping	Frequency	% of total SemLink frame types
Matched VN	High Frequency (top 30%)	18.57%
	Middle Frequency (middle 40%)	9.78%
	Low Frequency (bottom 30%)	4.86%
Did not match VN	High Frequency (top 30%)	19.99%
	Middle Frequency (middle 40%)	29.16%
	Low Frequency (bottom 30%)	17.63%

Table 4: Results of Matching Process for SemLink Frame Types, Divided by Frequency

When considering frame types, it is clear that the majority of unique syntactic frame types in SemLink do not match VN. Among the frame types that did match VN, the majority of these were high frequency, although the highest frequency frame types in each class only match VN frames of the class 18.57% of the time. This indicates that a wider set of constituents is needed in VN syntactic frames and possibly a wider range of semantic roles in several VN classes in order to account for abstract usages that will better match SemLink data.

5 Discussion

This research demonstrated that while the majority of frame tokens in SemLink match frames in VN, the frames listed in VN need a wider set of constituents because the prototypical constituents for a particular role (e.g. NP-Agent) are not always reflective of the prototypical syntactic realizations in SemLink. In this way, both coverage and accuracy of VN frames could be improved simply by expanding the constituent types that can make up a given frame. To address this issue, a supplementary resource has been created that lists all constituent types found in SemLink that match a particular frame type. For example, this frame exists in the Remove class:

Frame NP V NP
Example Doug removed the smudges
Syntax Agent V Theme

The drawback of this frame is that it assumes that the Agent and Theme roles will be realized as NPs for all verbs in the class in all cases. This investigation of SemLink shows that the Agent V Theme frame can truly be realized with each of the following orderings of constituents:

S_V_NP
NP_V_SBAR
NP_V_NP

The first two possibilities are likely not canonical usages, but in order for VN to fully capture verbal behavior, the resource should reflect both theoretically expected usage and actual usage. The mapping resource created through this research will, however, greatly increase the coverage of VN by including all possible constituent types. Additionally, this resource will help to facilitate interoperability between VN and corpus resources by allowing the information in VN to be more easily compared and applied to that of parsed corpora.

5.1 Assessment of Coverage

Overall, VN currently describes the prototypical syntactic and semantic behavior of many English verbs, but its coverage of a large text corpus like SemLink is fairly low. This is demonstrated by the figures in Table 3, which show that only 12.92% of the frame types in

SemLink are covered by VN's syntactic frames. An additional 20.29% of the frame types in SemLink can be covered using VN's thematic role labels, but this still leaves 66.78% of the syntactic frame types in SemLink unmatched to VN. This is a strong indication that there is a great amount of variability in the syntactic frame types that occur in real usage, which is not currently covered by VN.

When considering the impact of these results, it is important to remember that the organization of VN is based upon Levin's framework and hypothesis that semantic similarity underlies syntactic similarity. Accordingly, VN has focused on representing what can be thought of as typical, distinguishing frames and diathesis alternations of the verbs in a given class. The fact that these verbs participate in other syntactic behaviors not included in the classification is neither surprising nor does it necessarily undermine Levin's hypothesis, given that her classification was not originally intended to give a full enumeration of all behaviors, rather only distinctive behaviors. For the purposes of improving VN as a resource for NLP, the importance of coverage has become clear and is therefore the focus of this research. However, the focus of this research could easily be shifted to an examination of the frequency with which verbs participate in key diathesis alternations, and therefore an examination of Levin's hypothesis.

5.2 Increasing Coverage & Accuracy

Analysis of the SemLink instances that did not match VN frames revealed several classes that could be improved by the addition of a frame or thematic role, or both. In addition to the examples (3 & 4) of *remain* and its associated Exist class, which would require an additional Attribute role based on this study (discussed in Section 3.2), we found that a variety of other verbs and classes were characterized by roles and syntactic behaviors common to SemLink but not represented in VN. Unlike the examples of *remain,* some of these verbs represent new senses that may require entirely new classes. Consider these typical SemLink examples of the verb *add*, which take the following PropBank roleset:

Roleset id: add.03 , *achieve or gain*
Arg1: *Logical subject, patient, thing rising/gaining*
Arg2: *EXT, amount risen*
Arg4: *end point*

6. …*Nippon Mining added 15 to 960.*

...[Nippon Mining] $_{ARG1}$ added [15]$_{ARG2}$ [to 960]$_{ARG4}$

7. *Meanwhile, the broad-based Financial Times 100-share index added 30.4 points to end at 2142.6.*
[Meanwhile]$_{ARGM\text{-}TEMPORAL}$ [the broad-based Financial Times 100-share index]$_{ARG1}$ added$_{RELATION}$ [30.4 points]$_{ARG2}$ [to end at 2142.6]$_{ARG4}$

The verb *add* falls into several VN classes, Mix, Multiply and Say, of which the Multiply class is the closest fit. However, the Multiply class contains only the roles Agent, Theme, and Co-Theme, with frames such as:

Frame NP V NP PP
Example I multiplied x by y.
Syntax Agent V Theme {by} Co-Theme

This class does not reflect the realizations of the type seen in SemLink, which are particular to the financial domain. Thus, this study has revealed a gap in VN's coverage where the addition of a new (sub)class would be necessary to cover this sense of *add.*

The following table gives other examples of verbs, classes and actions required to increase the coverage of VN based on this study.

Verb	VN Class	Recommended Action
consent	Settle-89	Add NP V S frame: *Triton and Mr. Chase consented **to finding**…*
gain, rise, increase, climb	Calibratable-cos-45.6-1	Add Source/Result roles for beginning and final states: *Sales…rose 3% **to $29.3 million from $28.4 million**.*
get	-	Add class for *cause to do/be* sense: *We **can get that brought down to parity**…*
seek	Hunt-35.1	Add NP V S frame: *Cuba may seek **to postpone some sugar shipments**.*
stay, remain	Exist-47.1	Add Attribute role and frame NP V ADJP: *Oil prices stay **stable***
struggle	-	Add (sub)class for *try* sense: *The Sunday evening show **struggles to stay afloat***

5.3 Surprising Factors

One important factor revealed in the results of the frame

matching process is the large number of frame mismatches that were the result of the frame creation process itself. In the case of null elements, the frame creation method described in 3.1 was largely based on anaphora, rather than cataphora. Examples such as the one below, which include cataphoric co-reference, caused the creation of erroneous frames:

8. **Null subject* to further load the stakes, Mr. Lane dreamed up a highly improbable romance…*
Load, PropBank Load.01, VN class Spray-9.7-2:
[**Null subject**]-1 to further$_{\text{ARGM-EXTENT}}$ load$_{\text{RELATION}}$ [the stakes]$_{\text{ARG1, DESTINATION}}$, [Mr. Lane]-1$_{\text{ARG0, AGENT}}$ dreamed up a highly improbable romance …

The frame retrieved from this example was V_NP_NP, with the argument structure V Destination Agent. Neither of these matched the expected syntactic frame, as shown in the VN entry below.

Frame NP V NP.Destination
Example Jessica sprayed the wall.
Syntax Agent V Destination

This mismatch occurred because the argument to the verb was considered to be the realized constituent "Mr. Lane," rather than its previous null subject index. The algorithm for the frame matching process was designed to prefer realized subjects over null subjects, which in many cases was quite successful. However, examples such as these show that sometimes null elements are preferable when forming syntactic frames from a parse, in cases of cataphora. This is an area of improvement that needs to be considered when updating the frame matching process for future work.

6 Conclusion

This comparison of syntactic behavior in SemLink and the syntactic facts represented in VN has allowed for an expansion of the coverage and accuracy of VN. Although the frame matching method described herein requires further refinement, this method has provided data that can be used to compare VN with real language use. This will be of great value to VN as a lexical resource, since many verb classes can be improved by the insights gained from examining the frame mismatches from this project. The supplementary resource described in Section 5 will expedite such a task because it can be used to directly compare the syntactic frames available in SemLink for a particular verb's argument structure with the syntactic frames already available to a VN class. However, this resource is still limited by the erroneous frames generated during the matching process, such as in the cataphora example in Section 5.3. Further revisions to the method of forming syntactic frames from a given parse could better reflect these types of usage.

7 Future Work

As stated in the sections above, the frame matching process described in this paper is still in need of some refinement, to handle all the syntactic variations that occur in SemLink. In particular, the passive syntactic frames will need to be added back into the frame matching process, after further consideration on how to handle such frames. It may be necessary to add passives to the loose matching condition that was applied to cases of topicalization and WH-movement. In addition, the frame retrieval process needs to be revised to account for cataphoric co-reference with a null subject, and other cases of null elements that cause problematic syntactic frames to be generated. Finally, the forthcoming new version of SemLink will be updated with the latest set of VN thematic roles and expanded, which should prove helpful when re-implementing the frame matching process described in this paper.

Once the frame matching process has been further refined, a more in-depth analysis of the impact of these findings will be undertaken. Specifically, while this research has focused on adding syntactic frames to VN in order to increase coverage, future research should focus on the extent to which verbs participate in the key diathesis alternations represented in both VN and Levin's classes. A focus on this question would allow for valuable discoveries in the validity of Levin's hypothesis that syntactic behavior stems from semantics.

The syntactic frame data generated by this project will also be useful for future work in automatic verb clustering. The syntactic frames alone may prove to be a great feature for predicting verb classification, and such an automatically structured classification could be usefully compared to the VN classification to further evaluate it. Perhaps most importantly, the results of this research should increase the value of VN as a NLP resource. The addition of new syntactic constituent types and thematic roles to VN classes based on the SemLink syntactic frames and argument structures should allow for VN to more accurately and comprehensively reflect English verbal behavior, which makes VN more practical for a range of NLP tasks.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation Grant NSF-IIS-1116782, A Bayesian Approach to Dynamic Lexical Resources for Flexible Language Processing, DARPA/IPTO funding under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022, VN Supplement, and funding under the BOLT and Machine Reading programs. HR0011-11-C-0145 (BOLT) FA8750-09-C-0179 (M.R.) Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bonial, C., Brown, S.W., Corvey, W., Palmer, M., Petukhova, V., and Bunt, H. 2011a. An Exploratory Comparison of Thematic Roles in VN and LIRICS. *Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-6)*, 39-44.

Bonial, C., Corvey, W., Palmer, M., Petukhova, V., & Bunt, H. 2011b. A hierarchical unification of lirics and VN semantic roles. In *Proceedings of the ICSC Workshop on Semantic Annotation for Computational Linguistic Resources (SACL-ICSC 2011).*

Bonial, C., Brown, S.W. Hwang, J.D., Parisien, C., Palmer, M., and Stevenson, S. 2011c. Incorporating Coercive Constructions into a Verb Lexicon. *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics,* Portland, Oregon, June 2011.

Bonial, C., Feely, W., Hwang, J.D., and Palmer, M. 2012. Empirically Validating VN Using SemLink. In *Proceedings of the ICSC Workshop on Semantic Annotation for Computational Linguistic Resources* (SACL-ICSC 2012), May, 2012.

Fillmore, C. J., Johnson, C. R., & Petruck, M. R. L. (2002). Background to Framenet. *International Journal of Lexicography,* 16(3):235–250.

Joanis, E., Stevenson, S. & James, D. (2007). A general feature space for automatic verb classification. *Natural Language Engineering,* 14(3):337–367.

Kipper, K. Korhonen, A., Ryant, N. and Palmer, M. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42, pp. 21-40.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Loper, E., Yi, S. & Palmer, M. (2007). Combining lexical resources: Mapping between PropBank and VN. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7),* Tilburg.

Palmer, M. (2009). SemLink: Linking PropBank, VN and FrameNet. In *Proceedings of the Generative Lexicon Conference, GenLex-09,* Pisa, Italy.

Palmer, M., Gildea, D., and Kingsbury, P. 2005. The Proposition Bank: An annotated Corpus of Semantic Roles. *Computational Linguistics*, 31:1, pp. 71-105.

Swier, R.S. and Stevenson, S. 2004. Unsupervised semantic role labeling. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 95-102.

Zaenen, A., Condoravdi, C., and Bobrow, D. G. (2008). The encoding of lexical implications in VN. In *Proceedings of LREC 2008,* Morocco, May.

Acquiring typed predicate-argument structures from corpora

Elisabetta Jezek
Università di Pavia
Dipartimento di Studi Umanistici
Sezione di Linguistica
jezek@unipv.it

Abstract

In this note, I illustrate the methodology we are currently using to acquire typed predicate-argument structures from corpora, with the aim of compiling a repository of corpus-based patterns for Italian verbs and obtaining an empirically sound inventory of argument type shiftings in context for linguistic research and NLP applications. The note is organized as follows. I first introduce the resource, then focus on the annotation of type mismatches between argument fillers and verb selectional requirements and their linguistic classification, based on Generative Lexicon Theory (Pustejovsky et al. 2008). Finally, I outline the ongoing attempt to combine typing annotation with standard coarse-grained (high-level) thematic role information (Bonial et al. 2011a, 2011b) carried out in collaboration with Senso Comune (Vetere et al. 2011). A discussion of ongoing evaluation and improvements follows.

1 The resource

Typed predicate-argument structures are corpus-derived verb frames[1] with the specification of the expected semantic type for each argument position (e.g. [[Human]] mangia [[Food]], [Human]] guida [[Vehicle]], [[Human]] partecipa a [[Event]]), populated by lexical sets (Hanks 1986), i.e. the statistically relevant list of collocates that typically fill each position (e.g. [[Event]-iobj of partecipare] = {gara, riunione, selezione, manifestazione, seduta, cerimonia, conferenza, votazione ...}). The repository of corpus-based patterns for Italian verbs is a manually annotated resource under development at the University of Pavia in collaboration with the Faculty of Informatics at Masaryk University (Brno) and FBK (Trento). It currently consists of a nucleus of about 300 lexical units (verbs). In the resource, each lexical unit is linked to a set of frames extracted from the corpus following the Corpus Pattern Analysis technique (CPA, Hanks - Pustejovsky 2005). Each frame is associated with a corpus-derived verb sense (expressed in the form of an implicature linked to the typing constraints) and with a set of corpus instances (a sample of 250 occurrences for each verb), that represent more prototypical and less prototypical instantiations of the frame. Each corpus instance is tagged with information about pattern number and anomalous arguments, i.e. arguments that do not satisfy the typing constrains specified in the frame. At present, in compiling the patterns, we are using a list of shallow semantic types ([[Human]], [[Artifact]] etc.) borrowed from the English project (Pattern Dictionary of English Verbs (PDEV), project page at http://deb.fi.muni.cz/pdev/). The reference corpus for the Italian project is a reduced version of itWaC (Baroni & Kilgarriff 2006). We plan to make the resource available once the goal of analyzing 1000 "verbi a polisemia media" (average polysemous verbs) is reached.

[1] By verb frame we mean the relational semantic structure associated with the verb, specifying information about the linguistically relevant participants in the event encoded by the predicate.

2 Type mismatches

In acquiring the patterns from corpus data, the annotator retrieves the corpus instances, identifies the relevant structure, analyses the lexical set for each grammatical relation and associates a typing assignment to each argument position in the pattern.[2] Once the pattern is identified, each corpus instance is tagged with the associated pattern number. One recurrent problem that arises in this phase is the identification of mismatches between pattern type (assigned by the verb) and instance type (inherent in the argument filler) within the same grammatical relation.

3 Mismatch classification

Mismatches may be classified according to the following parameters:

- *Verb class* (Levin 1993, VerbNet, ...): aspectual verbs, communication verbs, perception verbs, directed motion verbs, verbs of motion using a vehicle ...
- *Targeted grammatical relation*: SUBJ_OF, OBJ_OF, COMPL ...
- *Shift type* (domain-preserving vs. domain-shifting): Artifact as Event, Artifact as Human, Artifact as Sound, Event as Location, Vehicle as Human ...
- *Elasticity/flexibility of noun class*: Artifacts vs. Naturals ... (Lauwers and Willems 2011).

Assuming a qualia-based lexical representation for nouns, as in Generative Lexicon, mismatches may be further classified according to which quale/qualia is/are exploited or introduced in composition. Besides the four standard roles, i.e.

- *Formal* (F): encoding taxonomic information about the lexical item (the `is-a` relation);
- *Constitutive* (C): encoding information on the parts and constitution of an object (`part-of` or `made-of` relation);
- *Telic* (T): encoding information on purpose and function (the `used-for` or `functions-as` relation);
- *Agentive* (A): encoding information about the origin of the object (the `created-by` relation).

we may assume that lexical representations include values for the following relations (Pustejovsky and Jezek Forth.):

- *Natural Telic* (NT): property that is necessarily associated with a natural kind (no intentionality). For example: *river*NT=$flow$, *heart*NT=$pump\_blood$.
- *Conventionalized Attribute* (CA): property/activity routinely or systematically associated with an object, but not strictly part of the identified Qualia roles. For example: *dog*CA=$bark$, *car*CA=$park$, *food*CA=$digest$.

3.1 Data

What follows is a list of examples of mismatches classified according to the parameters introduced above: a) verb class, b) targeted grammatical relation (in italics), c) type of shift (instance type *as* pattern type) and d) targeted Quale of the noun (both relation and value). In the examples, the instances are being matched to the semantic types derived from a CPA study of these verbs.[3]

(1) *Aspectual Verbs*

Arriva Mirko e interrompe *la conversazione*. 'Mirko arrives and interrupts the conversation' (matching)
Il presidente interrompe *l'oratore*. 'The president interrupts the speaker' (Human as Event; T=parlare 'speak')

(2) *Communication Verbs*

Lo speaker annuncia *la partenza*. 'The speaker announces the departure' (matching)
Il maggiordomo annuncia *gli invitati*. 'The butler announces the guests' (Human as Event,

[2] Types are conceived as abstractions over the lexical sets found in the argument slots in the corpus.

[3] This study was used as a base to build the dataset for the SemEval-2010 shared task on coercion (see below).

CA=arrivare ‘arrive’)[4]
L’altoparlante annunciava l’arrivo del treno. ‘The loudspeaker announces the arrival of the train’ (Artifact as Human; T=usare ‘use’(human, tool))
Una telefonata anonima avvisa la polizia. ‘An anonymous telephone call alerted the police’ (Event as Human; AG=telefonare ‘phone’(human1, human2))

(3) *Avoid Verbs*

Abbiamo evitato *l’incontro*. ‘We avoided the meeting’ (matching)
Meglio evitare *i cibi fritti*. ‘It is best to avoid fried food’ (Artifact as Event; T=mangiare ‘eat’)

(4) *Forbid Verbs*

Nell’Italia di allora la legge vietava *l’aborto*. ‘At that time in Italy law prohibited abortion’ (matching)
La Francia vieta *il velo* a scuola. ‘France bans the headscarf in schools’ (Artifact as Event; T=indossare ‘wear’)

(5) *Verbs of desire (Bos 2009)*

Preferisco *bere* piuttosto che *mangiare*. ‘I prefer drinking to eating’ (matching)
Preferisco *la birra al vino*. ‘I prefer beer to wine’ (Artifact as Event; T=bere ‘drink’)

(6) *Perception verbs*

Rilassarsi ascoltando *il rumore della pioggia*. ‘Relax while listening to the sound of rain’ (matching)
Ascoltava *la radiolina* con la cuffia. ‘He listened to the radio with his earphones’ (Artifact as Sound: T=produrre_suono ‘produce_sound’)
Rimasi a lungo ad ascoltare *il suo respiro*. ‘I stayed for a long while listening to his breath’ (Event as Sound; NT=produrre_suono ‘produce_sound’)
Non ho potuto ascoltare *tutti i colleghi* ‘I could not listen to all colleagues’ (Human as Sound; CA=parlare ‘speak’)

(7) *Directed motion verbs*

Abbiamo raggiunto *l’isola* alle 5. ‘We reached the island at 5’ (matching)
Ho raggiunto *il semaforo* e ho svoltato a destra. ‘I reached the traffic light and turned right’ (Artifact as Location; CA=essere_a ‘be_at ’(location))
Gli invitati arrivano *al concerto* in ritardo. ‘The guests arrive late at the concert’ (Event as Location; CA=aver luogo_a ‘take place_at’(location))

(8) *Motion using a vehicle*

Il nostro aereo atterra alle 21. ‘Our plane lands at 9pm’ (matching)
Il pilota è regolarmente atterrato senza problemi. ‘The pilot landed regularly with no problems’ (Human as Vehicle; T=pilotare ‘pilot’(human, vehicle))
Tutti i voli civili sono atterrati. ‘All civilian flights landed’ (Event as Vehicle; $ArgStr$ Exploitation?)

(9) *Vehicle Verbs*

Luca ha parcheggiato sotto casa. ‘Luca parked near the house’ (matching)
L’ambulanza ha parcheggiato lontano. ‘The ambulance parked far away’ (Vehicle as Human; T=guidare ‘drive’(human, vehicle))

4 Mismatch tagging

At present, we treat the entire NP as a markable. Following the CPA procedure, regular choices of types within the same argument position are coded as type alternations. Common alternations in subject position are for instance [[Human|Institution]] and [[Human|Body Part]], for example: [[Human|Body Part]] sanguina ‘bleeds’. “Non-canonical lexical items breaking a particular statistical threshold are coerced into honorary membership of a semantic type in particular contexts”. Honorary members are tagged as “a” = anomalous arguments.

[4] As noted by one reviewer, this example may be analyzed as an instance of a different sense of *annunciare* with different constraints. We propose instead that the sense is one and the same, and that the interpretation of the specific combination is achieved by exploiting one of the events conventionally associated with the noun.

5 Improving coercion annotation

Ongoing work focuses on improving the annotation of corpus instances in regard to three areas:

- annotating instance types,
- annotating the targeted quale/qualia in V-ARG composition,
- interfacing typing and semantic role annotation.

Each of these points is examined below.

5.1 Annotating instance types

Based on Pustejovsky et al 2008, 2010 (SemEval Coercion Task) and previous attemps to annotate metonymic relations in text (Markert and Nissim 2007), in Jezek and Frontini 2010 we finalized a scheme to annotate type mismatches in the resource. The scheme foresees three layers of semantic annotation:

- the Pattern Type, which records the semantic type that is inherited by the pattern for each argument position;
- the Argument Filler, which contains the lexical material that instantiates the semantic position in the instance;
- the Instance Type, which needs to be added when the argument filler instantiates a type that does not match with the Pattern Type, otherwise it is inherited from the pattern.

The following is an example:

(10) I ragazzi hanno bevuto una pinta insieme.
'the boys drank a pint together'
[[Human]-subj] beve [[Liquid]-obj]

<instance tid=102> <argument id=a1 pattern_id=p15 instance_sem_type=HUMAN instance_syn_role=subj> I ragazzi </argument> <verb pattern_id=p15> hanno bevuto </verb> <argument id=a2 pattern_id=p15 instance_sem_type=MEASURE_UNIT instance_syn_role=obj> una pinta </argument> insieme. </instance>

5.2 Annotating the targeted quale in V-ARG composition

In Jezek, Quochi, Calzolari 2009 and Jezek and Quochi 2010 we explored how to integrate qualia specification (relation and/or value) in the coercion annotation task, in addition to type specification. This may be attained in two ways:

- as online specification during the annotation,
- retrieving it from a pre-existing resource (e.g. SIMPLE, QS gold standard, noun-frame repository ....).

5.3 Interfacing types with semantic role annotation

In the resource, typing information is sometimes complemented with fine-grained semantic roles. In principle, the semantic type captures the Formal quale of the argument, which is an intrinsic property of nouns normally found in that argument slot (e.g. person, substance, artefact etc.). On the other hand, the semantic role captures an extrinsic property of the nouns in the same slot, namely one that specifies how the referent is involved in the event (e.g. as an intentional agent, an affected entity, a created entity, and so forth). This is illustrated below:

(11) [[Human 1 = Legal Authority]] arresta 'arrest' [[Human 2 = Suspect]]

Ongoing work focuses on improving role annotation with systematic coarse-grained roles annotation. In the context of the Senso Comune initiative (www.sensocomune.it), we designed a set of 27 coarse-grained (high-level) semantic roles based on VerbNet (VN) and LIRICS (Petukhova and Bunt 2008) and the on-going attempt to create a unified standard set for the International Standard Initiative (ISO) (Bonial et al. 2011a, b).[5] We conflated some LIRICS roles (e.g., Medium and Instrument), adopted some suggestions from Bonial et al. 2011a (e.g., the use of co-Agent and co-Patient rather than the unique Partner), and used some classical semantic roles like Experiencer rather than LIRICSs ambiguous Pivot. We adopted the hierarchy in Bonial

[5] Besides the author of this note, the group working at role annotation in Senso Comune includes Fabio Massimo Zanzotto, Laure Vieu, Guido Vetere, and Alessandro Oltramari.

et al. 2011b, but distinguished between *participants* and *context*.

We performed a pilot experiment on 400 usage examples (about 6% of the entire corpus) associated with the sense definitions of 25 fundamental verb lemmas of the Senso Comune resource to release the beta-version of the annotation scheme.

The annotation task involves tagging the usage instances with syntactic and semantic information about the participants in the frame realized by the instances, including argument/adjunct distinction. In semantic annotation, annotators are asked to attach a semantic role and an ontological category to each participant and to annotate the sense definition associated with the filler. We provide them with the hierarchical taxonomy of roles based on Bonial 2011b, together with definitions and examples for each role. The TMEO methodology (cf. Vetere et al. 2011) is used to help them selecting the ontological category in Senso Comune's top-level. For noun sense tagging, the annotator exploits the senses already available in the Senso Comune resource. Drawing on the results of previous experiments on "ontologization" of noun senses (Chiari et al. 2011), we allow multiple classification, that is, we allow the annotators to tag each slot with more than one semantic role, ontological category and sense definition. For example in the context in (12), the subject may be tagged with both Agent and Experiencer, if the annotator assumes that the participant shares entailments which belong to both roles.

(12) [I turisti AG EXP / Human] ammirano i quadri.
'The tourists admire the paintings'

The pilot experiment confirms our expectation that in category assignment, annotators are influenced by the inherent semantic properties of the referents filling the argument positions. For example, in (13) they annotate the referent of the object argument as Human, even though it is metonymically reinterpreted as Document in the context of *leggere* 'read'. Interestingly, the inherent semantic properties of the argument's referents appear to play a role also in semantic role assignment. For example, in the coercive environment in (13), the annotator hesitates whether he/she should annotate the mismatch in object position also at the role level, i.e. assigning Source instead of Theme (the latter is the role chosen for such contexts as *leggere una lettera, il giornale* 'read a letter, the newspaper' and so forth).

(13) leggere [un autore ?SOURCE / Human]
'read an author'

This appears to hold true also when annotations of role and ontological category are performed as separate sub-tasks. That is, if annotators are asked to annotate the semantic role only (besides grammatical relations), semantic role assignment still appears to be performed (also) on the basis of the perceived inherent category of the argument filler. We are currently exploring how to approach this issue (both in theory and in annotation practice), that appears to involve several classes of phenomena, including Instruments (and other kinds of Artifacts) in Subject position, as in (2) and (9) above.

6 Conclusions

In this note, I described the effort of creating a repository of corpus-based patterns for Italian verbs for purposes of linguistic research and NLP application. This involves creating a corpus-based inventory of metonymic shifts as a by-product. Ongoing work focuses on improving mismatch annotation and on examining the interplay between typing and role constraints to argument selection, focusing on coercive environments.

Acknowledgments

I would like to thank James Pustejovsky, Alessandro Lenci, Francesca Frontini and Valeria Quochi for coercion-related discussion, and Patrick Hanks for his scientific support for the Italian CPA project. I also acknowledge Guido Vetere, Laure Vieu, Fabio Zanzotto and Alessandro Oltramari for the ongoing collaborative work on semantic role annotation within the Senso Comune initiative. Thanks also to the participants of the "Semantic and pragmatic annotation of corpora" Workshop, held at Heinrich-Heine-Universität Düsseldorf, July 16, 2012, and to four anonymous reviewers for their fruitful comments.

7 References

Baroni, M. and A. Kilgarriff 2006. Large Linguistically-Processed Web Corpora for Multiple

Languages. In *EACL 2006 Proceedings*, 87-90.

Bonial, C., S.W. Brown, W. Corvey, V. Petukhova, Palmer M., Bunt H. 2011a. An Exploratory Comparison of Thematic Roles in VerbNet and LIRICS. In *Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Bonial, C., W. Corvey, Palmer M., V. Petukhova, Bunt H. 2011b. A hierarchical unification of LIRICS and VerbNet semantic roles. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, IEEE Computer Society Washington, DC, USA, 483-489.

Bos, J. 2009. Type Coercion in the Contexts of Desire. In P. Bouillon et al. (eds) *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pisa, ILC-CNR, Sept. 17-19, 2009.

Hanks, P. 1986. Contextual Dependencies and Lexical Sets. In *International Journal of Corpus Linguistics* 1:1, 7598.

Hanks, P. and J. Pustejovsky 2005. A Pattern Dictionary for Natural Language Processing. In *Revue franaise de linguistique appliquée*, 10 (2), 63-82.

Chiari, I. Oltramari, A. Vetere, G. 2011. Di cosa parliamo quando parliamo fondamentale? In S. Ferreri (ed.) *Atti del Convegno della Società di linguistica italiana*, Roma, Bulzoni, 221-236.

Jezek E., V. Quochi and N. Calzolari 2009. Relevance of Qualia Relations in Coercive Contexts. In P. Bouillon et al. (eds) *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pisa, ILC-CNR, Sept. 17-19, 2009, 128-136.

Jezek E. and F. Frontini 2010. From Pattern Dictionary to PatternBank. In G.M. de Schryver (ed) *A Way with Words: Recent Advances in Lexical Theory and Analysis*, Kampala, Menha Publishers, 215-239.

Jezek E. and V. Quochi 2010. Capturing Coercions in Texts: a First Annotation Exercise. In Nicoletta Calzolari et al. (eds) *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta (May 19-21, 2010), Paris: European Language Resources Association (ELRA), 1464-1471.

Lauwers, P. and D. Willems 2011. Coercion: Definition and Challenges, current approaches and new trends. In *Linguistics* 49:6, 1219-1235.

Markert K. and M. Nissim. 2007. SemEval-2007 task 8: Metonymy resolution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, Prague, Czech Republic. Association for Computational Linguistics.

Petukhova,V., Bunt H. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 28–30.

Pustejovsky, J., Rumshisky, A., Moszkowicz, J.L., Batiukova, O. 2008. GLML: A Generative Lexicon Markup Language. ms presented at GL workshop, Pisa, Istituto di Linguistica Computazionale (CNR), Sept. 2008.

Pustejovsky J., Rumshisky A., Plotnick A., Jezek E. Batiukova O., Quochi V. 2010. SemEval Task 7: Argument Selection and Coercion. In Proceedings of the Fifth International Workshop on Semantic Evaluation Uppsala University, Sweden, July 1116 2010.

Pustejovsky, J. and E. Jezek (Forth.). *Generative Lexicon Theory: A Guide*, Oxford, Oxford University Press.

Vetere, G., Oltramari, A., Chiari I., Jezek E. Vieu L., Zanzotto F. 2011. Senso Comune: An Open Knowledge Base for Italian. In *Traitement Automatique des Langues* (TAL), 52:3.

Counting Time and Events

Kiyng Lee
Korea University
Deaprtment of Linguistics
Seoul, Korea
ikiyong@gmail.com, klee@korea.ac.kr

Harry Bunt
Tilburg University
TiCC: Center for Cognition and Communication
Tilburg, The Netherlands
harry.bunt@uvt.nl

Abstract

Recurring events (e.g., *John calls twice everyday*) involve both temporal and event quantification. To annotate such events, there are two main approaches: one approach is represented by Pustejovsky et al. (2010a,b) and the other one, by Bunt and Pustejovsky (2010) and Bunt (2011a,b). In the framework of ISO-TimeML (2012), the first approach encodes information on quantification directly into both temporal and event entities, by introducing the attribute `@quant` into the element <EVENT> as well as the element <TIMEX3>. The second approach views quantification as a set of properties of the way a predicate applies to a set of arguments, or relates two sets of arguments, such as sets of events and their time of occurrence, and therefore annotates aspects of quantification as part of a relational link, such as <TLINK> or <TIME_ANCHORING>. In this paper, we discuss alternatives and explore possibilities to reach general consensus on the annotation of quantification over time and events and its extendibility to other entities.

1 Introduction

In January 2012, ISO-TimeML (2012) was published as an ISO's international standard: ISO 24617-1(E):2012 *Language resource management - Semantic annotation framework - Part 1: Time and events (SemAF-Time, ISO-TimeML)*.[1] Although it was officially published as an international standard, this document still contains a couple of issues that remain to be resolved, especially those related to the annotation of recurring time and events that involves quantification, distributivity, and scopes.

These issues involve for example the annotation of the following sorts of expressions:

(1) Sample Data for Recurring Time and Events
   a. Type 1: John called *twice*.
   b. Type 2: John calls *every day*.
   c. Type 3: John calls *twice a day*.

In ISO-TimeML (2012), the predicate modifiers, italicized in (1), are all treated as referring to temporal entities, thus all are annotated into the element, named <TIMEX3>, and almost in the same manner.

Detailed analysis shows, however, that each of them should be annotated differently, involving temporal and event quantification and scopes. This difference requires minor or major modifications in the current version of ISO-TimeML (2012). Far before its publication, some issues on quantification had been known and much discussed. On these issues, we note at least two proposals besides the published standard itself: (1) Pustejovsky et al. (2010a,b) and (2) Bunt and Pustejovsky (2010) and Bunt (2011a,b).

The first proposal is a minimally modified version of ISO-TimeML (2012) with its representation scheme, which we tag `<isoTimeML_m>`. It differs from the representation scheme of ISO-TimeML (2012), `<isoTimeML>`, in two ways. First, this minimally modified version `<isoTimeML_m>` annotates event quantification by introducing the at-

[1] *SemAF-Time* refers to the document as a whole, while *ISO-TimeML* refers to the XML-based annotation language specified in that document.

tribute @quant into the element <EVENT>. Second, it marks up the scopes of temporal and event quantification explicitly by introducing the attribute @scopes into both <TIMEX3> and <EVENT>.

The second proposal annotates quantification over events in a different way. Based on the analysis of quantification in Bunt (1985), it views quantification as a a set of properties of the way a unary predicate applies to a set of arguments, or a binary predicate relates two sets of arguments. A predicate that relates a set of events to their times of occurrence, as expressed for instance by the preposition "at" in the sentence "John always calls at two o'clock" is annotated in ISO-TimeML with the relational link <TLINK>, and aspects of quantification are therefore annotated as part of this link. Since the <TLINK> tag is heavily overloaded in ISO-TimeML, as it is used for rather different purposes, hence for this specific use the tag <TIME_ANCHORING> is introduced. Annotation are marked up in a representation scheme, called the Ideal Concrete Syntax, which is designed according to the CASCADES methodology of designing annotation languages with an abstract syntax and a formal semantics (see Bunt, 2010; 2012) - this approach to the annotation of quantification is tagged <isoTimeML_ICSrep>.

While on the one hand quantification on the latter view is considered to arise when a predicate is applied to one or more sets of arguments (rather than to arguments which are single individuals), and it thus seems natural to annotate aspects of quantification as parts of relational link elements, it was noted in Bunt (1985) on the other hand that satisfactory semantic representations of sentences with quantifications can be obtained by considering aspects of quantification as parts of the compositional semantics of noun phrases. This is because NP representations can be defined in such a way that they anticipate on the use of the NP as an argument of a predicate, as already shown by Montague (1973).

The treatment of quantification proposed by Montague did not take the phenomenon of 'distributivity' into account, however, i.e. whether the members of an argument set are involved in the predication as individuals, in groups, or as a collectivity - see e.g. the example "Two men lifted the piano". Bunt (1985) showed that it is possible to construct semantic representations for noun phrases with different distributivities; interestingly, though, distributivity is often not expressed in a noun phrase, but by adverbials like "together" and "one by one", so it is not evident that this aspect of quantification would most conveniently be treated as part of NP semantics or in the semantics of combining an NP with a predicate.

The main purpose of this paper is to discuss and explore possibilities to annotate quantifications over time and events, for use in a future extended version of ISO-TimeML (2012), but also to contribute to the study of how to annotate quantification more generally, as explored in ISO project 24617-6, "Basic principles of semantic annotation", since in the end a treatment of quantification over time and events should be a special case of quantification more generally.

2 Two Annotation Schemes

Frequency is normally understood to be a number of occurrences of a repetitive event over a period of time; predicate modifiers such as "twice", "every day", and "twice a day" or "twice every day" are often treated as frequency expressions. In this section we discuss how these modifiers are annotated in two different representation schemes: <isoTimeML> and <isoTimeML_ICSrep>.

2.1 Annotation of Type 1 Modifier *twice*

ISO-TimeML (2012) annotates "twice" as a temporal entity expressing a frequency, encoding its information into the element <TIMEX3>.

(2) a. John called$_{e1}$ twice$_{t1}$.

b.
```
<isoTimeML xml:id="a1">
<EVENT xml:id="e1" pred="CALL"
tense="PAST"/>
<TIMEX3 xml:id="t1" freq="2X"/>
<TLINK eventID="#e1"
relatedToTime="#t1"
relType="DURING"/>
</isoTimeML>
```

This is interpreted as shown below:[2]

[2]The semantics of this interpretation was developed by Pratt-Hartmann as an extension of Pratt-Hartmann (2007). For the compositional process of deriving the two semantic representations, see Clause 8.4.3.3 Quantifying <TIMEX3> element in ISO-TimeML (2012), pp. 32-33.

(3) Interval_based First-Order Form:
$\exists 2 I_{e1}(R_{during}(I_{e1}, I_{t1}) \wedge p_{call}(I_{e1})$

This semantic form is understood as stating that, within a *contextually determined interval of time* I_{t1}, there were two instances of an event of type CALLING. Each interval I_{ei} is understood as an interval of time in which an event of type e is instantiated as e_i.[3]

Bunt and Pustejovsky (2010) argue that the modifier "twice" does not denote a temporal entity at all, but it is simply a counter, expressing how often a certain type of event occurred. `<isoTimeML_ICSrep>` thus annotates it as a part of the element <EVENT>.[4] This element is then specified with two attributes `@signature` with a value SET and `@cardinality` for the cardinality of the (specified) set, as shown below:.

(4) a. John$_{token1}$ called$_{token2}$ $twice_{token3}$.
b.
```
<EVENT xml:id="e1"
target="#range(token2,token3)"
type="CALL" tense="PAST"
signature="SET" cardinality="2"/>
```

This is interpreted as stating that, given a set E of two events, each event in E is of type CALL. This is also given a formal semantics in DRT (Discourse Representation Theory) of Kamp and Reyle (1993), as shown below:

(5) a. $\exists 2 e[call(e)]$
b. $\exists S[|S|{=}2 \wedge \forall e[e \in S \rightarrow call(e)]]$

These two forms are equivalent.

2.2 Type 2 Modifier *every day* as a Temporal Quantifier

Both of the annotation schemes `<isoTimeML>` and `<isoTimeML_ICSrep>` treat type 2 modifiers as temporal quantifiers, but annotate them differently. The former annotates "every day" as part of the element <TIMEX3>.

(6) a. John$_{token1}$ calls$_{token2}$ every$_{token3}$ day$_{token4}$.
b.
```
<isoTimeML>
<EVENT xml:id="e1"
target="#token2" pred="CALL" />
<TIMEX3 xml:id="t1"
target="#range(token3,token4)"
pred="EVERYDAY" type="SET"
value="DAY" quant="EVERY"/>
<TLINK eventID="#e1"
relatedToTime="#t1"
relType="DURING"/>
</isoTimeML>
```

This is interpreted as stating that, during each day, the event of John's calling occurred, as represented in two different forms:

(7) a. Interval-based:
$\forall I_{day} R_{during}(I_{call}, I_{day})$
b. Event-based:
$\forall t[day(t) \rightarrow call(e, t)]$.

The latter scheme `<ISO-TimeML_ICSrep>`, however, divides the task of annotating "every day" over two elements: one is a new element, called <PERIOD>; the other is <TIME_ANCHORING>. The temporal noun "day" in "every day", for instance, is annotated into <PERIOD>, while the quantifier "every" is annotated into <TIME_ANCHORING> as a value of its attribute `@timeQuant`.[5]

(8) a. John$_{token1}$ calls$_{token2}$ every$_{token3}$ day$_{token4}$.
b.
```
<isoTimeML_ICSrep>
<EVENT xml:id="e1"
type="CALL" target="#token2"
signature="SET"/>
<PERIOD xml:id="t1"
type="DAY" target="#token4"
signature="SET"/>
<TIME_ANCHORING
anchoredEvent="#e1"
```

[3] In TimeML, from which ISO-TimeML was developed, <EVENT> was instantiated with an element <MAKEINSTANCE> and <TLINK> related these instances to each other or to a time. The ITL-based semantics of Pratt-Hartmann (2007) followed this version of TimeML and did not treat the semantics of "twice" or any quantifier expressions: it simply fails to treat quantification over events.

[4] The way of annotating "twice" directly into the element <EVENT> is exactly the same as that approach which annotates the negative expression "n't" or the tense "did", for instance, in "didn't call" into <EVENT> by providing it with information on its polarity and tense.

[5] Lee (2012) argues that there should be some formal constraints on the assignment of attriutes to links, on the basis of which we can, for instance, justify the validity of assigning such attributes as `@timeQuant` and `@eventDistr` to <TIME_ANCHORING>.

```
anchorTime="#t1"
tempRel="INCLUDED_IN"
eventDistr="INDIVIDUAL"
timeDistr="INDIVIDUAL"
timeQuant="EVERY"/>
</isoTimeML_ICSrep>
```

The semantics defined for the abstract syntax that underlies this representation yields the desired interpretation (see Bunt 2011a,b).

Note that the quantifier "every" in "every day" has ended up in the <TIME_ANCHORING> element relating the events and their times of occurrence, rather than in the <PERIOD> element that correspond to the word "day" rather than to the NP "every day". It may be considered a drawback of this approach that NPs as such are not treated as units, which may be more convenient for human annotators. It does however seem possible to modify the <ISO-TimeML_ICSrep> scheme such that the aspect of quantification expressed by quantifier words is moved from link elements to the elements annotating the linked arguments.

2.3 Annotation of Type 3 Modifier "Twice a Day"

The type 3 modifier "twice a day" is treated in <isoTimeML> as a one structural unit and annotated in a single element <TIMEX3>, as below:

(9) a. John calls$_{e1}$ [twice a day]$_{t1}$.
b.
```
<isoTimeML>
<EVENT xml:id="e1"
target="#token2" pred="CALL"/>
<TIMEX3 xml:id="t1"
target="#token3 #token4
#token5" type="SET" value="DAY"
quant="EVERY" freq="2X"/>
<TLINK eventID="#e1"
relatedToTime="#t1"
relType="INCLUDED_IN"/>
</isoTimeML>
```

This is interpreted as follows, again based on Interval Temporal Logic:

(10) $\forall J[[p_{day}(J) \wedge R_{during}(J, I_{t1})] \rightarrow$
$\exists_{2I_{e1}}(p_{call}(I_{e1}) \wedge R_{during}(I_{e1}, J))]$

Here are two levels of restricted quantification: the range of the universal quantification $\forall J$ is restricted to the time interval I_{t1}, as expressed by $R_{during}(J, I_{t1})$, while that of the existential quantifier $\exists_{2I_{e1}}$ is restricted to the variable J for a set of days, again as expressed by $R_{during}(I_{e1}, J)$. There are at least two intervals of I_{e1} during which the event of calling holds.[6]

The <isoTimeML_ICSrep> approach, on the other hand, provides the following annotation:

(11) a. John calls$_{tok2}$ twice$_{tok3}$ a$_{tok4}$ day$_{tok5}$.
b.
```
<isoTimeML_ICSrep>
<EVENT xml:id="e1"
target="#token2" type="CALL"
signature="SET"/>
<PERIOD xml:id="t1"
target="#token5" type="DAY"
signature="SET"/>
<TIME_ANCHORING
anchoredEvent="#e1"
anchorTime="#t1"
tempRel="INCLUDED_IN"
eventDistr="INDIVIDUAL"
timeDistr="INDIVIDUAL"
eventQuant="2"
timeQuant="EVERY"/>
</isoTimeML_ICSrep>
```

This yields the interpretation which says that "a set of call events is anchored time-wise in a set of days, such that the individual events are anchored at individual days, where every day includes a time anchor for two of these events."[7]

3 Quantification, Scopes, and Distributivity

In this section we first discuss event quantification and then a way to generalize quantification over other entities than time and events. We also discuss some issues concerning distributivity and some residual issues relating to set, scopes, and binding.

3.1 Event Quantification

ISO-TimeML (2012) annotates quantified temporal expressions, but has no provisions for anno-

[6] See for details ISO-TimeML (2012), p. 35.

[7] See the end of section 5, Bunt (2010b), example (50).

tating quantified events. Both Bunt and Pustejovsky (2010) and Pustejovsky et al. (2010a,b) extend the annotation of quantification to events, but in different ways. As was discussed in the previous section, Bunt and Pustejovsky (2010) annotate event quantification by introducing the attributes `@signature="SET"` and `@eventQuant` into the element <TIME_ANACHORING>.

Pustejovsky et al. (2010a,b), on the other hand, annotate event quantification by introducing the attributes `@type="SET"`, `@scopes` and `@quant` with values such as EVERY into the element <EVENT>.

Here is an illustration:

(12) Event Quantification
a. Mary [read]$_{e1}$ during [every lecture]$_{e2}$
b.
```
<isoTimeML_m>
<EVENT xml:id="e1"
target="#token2" pred="READ"/>
<EVENT xml:id="e2"
target="#token4 #token5"
pred="LECTURE" type="SET"
quant="EVERY" scopes="#e1"/>
<TLINK eventID="#e1"
target="#token3"
relatedToEvent="#e2"/
relType="DURING"/>
</isoTimeML_mod>
```
[8]

Here, the element `<EVENT xml:id="e2">` is specified with the attributes `@type="SET"` and `@quant="EVERY"`, just as in the case of temporal quantification.

Each element in the annotation is then interpreted as below:

(13) a. `<EVENT xml:id="e1"/>`: $\exists e_1[read(e_1)]$
b. `<EVENT xml:id="e2" quant="EVERY" pred="LECTURE" scopes="e1"/>`: $\forall e_2[lecture(e_2)]$
c. `<TLINK>`: $\lambda y \lambda x[\tau(x) \subseteq y]$

Note that the attribute `@scopes` is introduced to mark up the scopes of quantifiers explicitly. Note also that, to allow the interpretation (a) above, the event of reading (`<EVENT xml:id="e1"/>`) should be understood as having undergone existential quantification; in other words, the attribute `@quant` has the default value `"SOME"`.

Given scope information, we can now combine each of the interpretations through the operation of conjunction and obtain the following overall interpretation:

(14) $\forall e_2 \exists e_1[lecture(e_2) \rightarrow [read(e_1) \wedge \tau(e_1) \subseteq \tau(e_2)]]$

As is expected, this says that, during each lecture (e_2), an event (e_1) of Mary's reading took place.

3.2 Generalizing Quantification

In natural language, almost any predication or relation can be quantified. Hence, the annotation of quantification over times and events should be viewed as a special case of quantification involving predicates about and relations between sets of any kinds of entity.

In an event-based semantics, quantification over events turns up in every sentence, not just for the relation between events and their time of occurrence, but also for the relation between events and their participants. Consider, for instance, the following example:

(15) Everybody will die.

This is an interesting example, cited in Bunt and Pustejovsky (2010), for the discussion of quantification, distributivity, and scopes, but it cannot be annotated just with temporal and event quantification only.

To extend quantification to non-temporal entities, one possibility is to introduce an element <ENTITY>, and a linking tag <SRLINK> for annotating the relations between events and their participants. For illustration, the above example can be annotated as below:

(16) Annotation
a. Everybody$_{x1}$ [will die]$_{e1}$.
b.
```
<isoSEM xml:id="asr1">
<ENTITY xml:id="x1"
target="#token1" type="HUMAN"
signatue="SET" quant="EVERY"
scopes="#e1"/>
```

[8] This example is taken from Pustejovsky et al. (2010b), (28).

```
<EVENT xml:id="e1"
target="#token2 #token3"
type="DIE" tense="FUTURE"/>
<SRLINK event="#e1"
participant="#x1"
semRole="PATIENT"/>
</isoSEM>
```

The elements of this representation may be interpreted as follows:

(17) Interpretation
a. $\sigma_{x1} := \lambda x_1[human(x_1)]$
b. $\sigma_{e1} := \lambda x[die(x, e_1)]$

Both of the elements are interpeted as denoting sets, a set of humans and a set of ones who die. Here are two notes. First, σ_{x1} is not bound by the universal quantifier, corresponding to the specification of `@quant="EVERY"`. Second, the semantic role of the first argument of the predicate $die(x_1, e_1)$ can be spelled out to be $[die(e_1) \wedge Arg_1(patient, x_1, e_1)$.[9]

3.3 Distributivity

Since the publication of Bunt (1985), the notion of distributivity has become an important issue as a property of quantification in formal semantics. Consider:

(18) The two men swallowed a beer and lifted the piano.

This sentence is interpreted as saying that each of the two men drank a beer and they together lifted the piano. To obtain such an interpretation, we need a formal mechanism of characterizing the so-called distributivity of events so that some are treated as *individual* events (e.g., "each drinking a beer") or *collective* events (e.g., "together lifting the piano").

To treat distributivity, one idea, originally proposed in Bunt (1985), is to bring in higher-order variables such as variables for sets. With these variables, we can have the following semantic form, where $\mathcal{P}_2$(MEN) denotes the set of all sets of two men.

[9]See Pustejovsky et al. (2007) for details on the annotation of event participants or argument role assignments `<ArgLink>`.

(19) $\exists M[M \in \mathcal{P}_2(\text{MEN}) \wedge \forall x[x \in M \rightarrow [$
$\exists e_1[swallow\_beer(e_1) \wedge agent(e_1, x)] \quad \wedge$
$\exists e_2[lift\_piano(e_2) \wedge agent(e_2, M)]]]]$

Now the question is how to annotate sentences like the one given above, and how to derive such an interpretation. Again, we have several alternatives. One approach could be to encode distributivity for each relevant entity in the `<ENTITY>` element and use that information to trigger an appropriate link. Another way is to mark up that information on the attribute `@eventDistr` in the ICS representation.

The first approach runs into problems because the NP "The two men" are involved in *swallow*-events with individual distributivity and in a *lift*-event with collective distributivity. Trying to annotate this as different `@distributivity` values in the two `<EVENT>` elements makes no sense from a semantic point of view: it's not the elements of the sets of events that are involved individually or collectively, but only the participants in the agent role.

The second approach would give the following annotation (leaving out the parts that are not relevant for the present discussion), where the attribute `@cardinality` is used to represent the use of quantifier words for indicating the number of elements in argument set, as opposed to other uses that these words may have:

(20) a. [The two men]$_{x1}$ [swallowed a beer]$_{e1}$ and [lifted the piano]$_{e2}$.
b.
```
<isoTimeML_ICSRep xml:id="ad1">
<ENTITY xml:id="x1"
target="#token1 #token2 #token3"
type="MAN" signature="SET"
cardinality="2" outscopes="#e1"/>
<EVENT xml:id="e1"
target="#token4" type="SWALLOW"
signature="SET"/>
<EVENT xml:id="e2"
target="#token7" type="LIFT"
signature="ELEMENT"/>
<SRLINK xml:id="r1" event="#e1"
participant="#x1" semRole="AGENT"
participantDistr="INDIVIDUAL"/>
<SRLINK xml:id="r2" event="#e2"
participant="#x1" semRole="AGENT"
participantDistr="COLLECTIVE"/>
</isoTimeML_ICSRep>
```

The DRT-based semantics for <isoTimeML_ICSRep> given in Bunt (2011a,b), which does take the distributivity of quantifications into account, produces the semantic representation (19) for this annotation when extended with the treatment of <ENTITY> elements in the same way as the non-linking elements for time and events, and with the treatment of the <SRLINK> linking element in the same way as the temporal linking elements.

4 Concluding Remarks

In this paper, we have reviewed two versions of <isoTimeML> in dealing with the annotation of temporal and event quantification.

We do not pretend to have presented a fully developed proposal for quantification over time and events that generalizes to quantification over other than temporal and 'eventual' entities, but we identified strengths and weaknesses of different proposals. We hope that this will contribute to the following tasks: (1) the revision and extension of ISO-TimeML (2012) and (2) the development of the new ISO project concerned with the annotation of spatial information ("ISO-Space"), where much the same issues relating to quantification arise as in ISO-TimeML (when the relation between a set of events and their place of occurrence is quantified, as in "Heavy thunderstorms are expected tomorrow all over the country"), and (3) the development of the ISO project concerning the basic principles of semantic annotation, ISO NP 24617-6, in which quantification has been identified as one of the burning issues to be dealt with, that cut across several attempts to define standards for semantic annotation.

Acknowledgments

We thank the five anonymous reviewers for their positive reviews and constructive comments and also Suk-Jin Chang, Jae-Woong Choe, Alex C. Fang for reading the pre-final version of this paper. The authors also wish to thank James Pustejovsky for many interesting useful discussions.

References

Bunt, H. 1985. *Mass Terms and Model-theoretic Semantics*. Cambridge University Press, Cambridge.

Bunt, H. 2010. A methodology for designing semantic annotation languages exploiting syntactic-semantic iso-morphisms. In: A. Fang, N. Ide and J. Webster (eds.) *Proceedings of ICGL 2010, the Second International Conference on Global Interoperability for Language Resources*, Hong Kong City University, pp 29–45

Bunt, H. 2011a. Introducing abstract syntax + semantics in semantic annotation, and its consequences for the annotation of time and events. In E. Lee and A. Yoon (eds.), *Recent Trends in Language and Knowledge Processing*, 157-204. Hankukmunhwasa, Seoul.

Bunt, H. 2011b. Abstract syntax and semantics in semantic annotation, applied to time and events. Revised version of Bunt (2011a). Unpublished.

Bunt, H. 2012. CASCADES: A methodology for defining semantic annotations. Forthcoming in *Language Resources and Evaluation.*

Bunt, H., and J. Pustejovsky. 2010. Annotating temporal and event quantification. In H. Bunt (ed.), *Proceedings of ISA-5, the Fifth Joint ISO-ACL/SIGSEM Workshop on Interoperabile Semantic Annotation*, 15-22.

ISO/TC 37/SC 4/WG 2. 2012. *ISO 24617-1:2012(E) Language resource management - Semantic annotation framework - Part 1: Time and events (SemAF-Time, ISO-TimeML).* The International Organization for Standardization, Geneva.

Kamp H and U. Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

Lee, K. 2012. Interoperable Spatial and Temporal Annotation Schemes. *Proceedings of The Joint ISA-7, SRSL-3 and I2MRT Workshop on Interoperable Semantic Annotation*, pp. 61-68. The Eighth International Conference on Language Resources and Evaluation (LREC 2012) Satellite Workshop, Istanbul.

Montague, R. 1973. The proper treatment of quantification in ordinary English. In R. H. Thomason (ed.), *Formal Philosophy*, pp. 247-70. Yale University Press, New Haven, CT.

Pratt-Hartmann, I. 2007. From TimeML to Interval Temporal Logic. In J. Geertzen, E. Thijsse, H. Bunt, and A. Schffrin (eds.), *Proceedings of The Seventh International Workshop on Computational Semantics*, pp. 166-180. Tilburg.

Pustejovsky, J., J. Littman, and Roser Saur[$\acute{i}$]. 2007. Arguments in TimeML: Events and Entities. In Frank Schilder, Graham Katz, and James Pustejovsky (eds.), *Annotating, Extracting and Reasoning about Time and Events*, pp. 107-127. Springer, Berlin.

Pustejovsky, J., K. Lee, H. Bunt, and L. Romary. 2010a. *ISO-TimeML: A Standard for Annotating Temporal Information in Language*. in Proceedings of LREC 2010, the Seventh International Conference on Language Resources and Evaluation, 394-397. Malta.

Pustejovsky, J., K. Lee, H. Bunt, and L. Romary. 2010b. Revised version of Pusejovsky et al. (2010a), a manuscript submitted to the journal *Language Resources and Evaluation*.

Pustejovsky, J., J.L. Moszkowicz, and M. Verhange. 2012. *Proceedings of the The Joint ISA-7, SRSL-3 and I2MRT Workshop on Interoperable Semantic Annotation*, pp. 70-77. LREC 2012 Satellite Workshop, Istanbul.

IMAGACT: Deriving an Action Ontology from Spoken Corpora

Massimo Moneglia
Gloria Gagliardi
Alessandro Panunzi
LABLITA, University of Florence
moneglia@unifi.it

Francesca Frontini
Irene Russo
Monica Monachini
ILC, CNR, Pisa
monica.monachini@ilc.cnr.it

Abstract

This paper presents the IMAGACT annotation infrastructure which uses both corpus-based and competence-based methods for the simultaneous extraction of a language independent Action ontology from English and Italian spontaneous speech corpora. The infrastructure relies on an innovative methodology based on images of prototypical scenes and will identify high frequency action concepts in everyday life, suitable for the implementation of an open set of languages.

1 Introduction

In ordinary language the most frequent action verbs are "general" i.e. they are able to extend to actions belonging to different ontological types (Moneglia & Panunzi 2007). Figure 4 below gives an example of this property. Moreover, each language categorizes action in its own way and therefore the cross-linguistic comparison of verbs denoting everyday activities presents us with a challenging task (Moneglia 2011).

Spontaneous Speech Corpora contain references both to the most frequent actions of everyday life and to their lexical encoding and can be used as a source of semantic information in the domain of an action ontology.

The term Ontology Type is used here to identify the pre-theoretical sets of objects of reference in the domain of Action. Therefore our Ontology will be identified as referring to prototypic eventualities. IMAGACT uses both corpus-based and competence-based methodologies, focusing on high frequency verbs which can provide sufficient variation in spoken corpora. Besides helping in the evaluation of data found in actual language usage, competence based judgments allow us to consider negative evidence which cannot emerge from corpora alone. These judgments are needed to set up cross-linguistic relations. IMAGACT identifies the variation of this lexicon in the BNC-Spoken and, in parallel, in a collection of Italian Spoken corpora (C-ORAL-ROM; LABLITA; LIP; CLIPS). Around 50,000 occurrences of verbs, derived from a 2 million word sampling of both corpora, are annotated.

The project started on March 2011 and involves 15 researchers participating in three main work-packages (Corpus Annotation, Supervision and Cross-linguistic mapping, Validation and Language Extension). The annotation infrastructure is produced by a software house based in Florence (Dr.Wolf srl) and will be delivered as open source.

Roughly 500 verbs per language are taken into account, this represents the basic action oriented verbal lexicon (the Italian part of the task has now been completed, while 50% of the English verbs are still pending). The corpus annotation was performed by three native Italian speaking annotators (with 30 person months devoted to the task) and two native English speaking annotators (13 person months till now).

IMAGACT will result in an Inter-linguistic Action Ontology derived from corpus annotation. Its key innovation is to provide a methodology which exploits the language independent ability to recognize similarities among scenes, distinguishing the *identification* of action types from their *definition*. This ability is exploited both at the corpus annotation level (§2), for mapping verbs of different languages onto the same cross-linguistic ontology (§3) and for validation and extension of the data set to other languages (§4). The paper presents the web infrastructure that has been

developed to this end and the annotation methodology (www.imagact.it/imagact/).

2 Corpus Annotation

The annotation procedure is structured into two main steps: "Standardization & Clustering of Occurrences" and "Types Annotation & Assessment", accomplished by annotators with the assistance of a supervisor. The first task is to examine and interpret verb occurrences in the oral context, which is frequently fragmented and may not provide enough semantic evidence for an immediate interpretation. To this end the infrastructure allows the annotator to read the larger context of the verbal occurrence in order to grasp the meaning (Figure 1 presents one of over 564 occurrences of *to turn* in the corpus). The annotator represents the referred action with a simple sentence in a standard form for easy processing. This sentence must be positively formed, in the third person, present tense, active voice and must fill the essential argument positions of the verb (possible specifiers that are useful in grasping the meaning are placed in square brackets). Basic level expressions (Rosch 1978) are preferred or otherwise a proper name, and word order in sentences must be linear, with no embedding and/or distance relationship.

Crucially, along with the standardization, the annotator assigns the occurrence to a "variation class" thus determining whether or not it conveys the verb's meaning. This is what we mean by a PRIMARY occurrence.

This task is accomplished through a synthetic judgement which exploits the annotator's semantic competence (Cresswell 1978) and is given in conjunction with Wittgenstein's hypothesis on how word extensions can be learned (Wittgenstein 1953). The occurrence is judged PRIMARY according to two main operational criteria: a) it refers to a physical action; b) it can be presented to somebody who does not know the meaning of the verb V, by asserting that "the referred action and similar events are what we intend with V". The occurrence is judged MARKED otherwise, as with "John turns the idea into a character", as shown in Figure 1 above. We have strong evidence regarding the inter-annotator agreement on this task which may require cross-verification in a few occasions of uncertainty (over 90% in our internal evaluation, based on the performance of two native English and Italian speaking expert annotators).

Only occurrences assigned to the PRIMARY variation class (216 over 564 in this case) make up the set of Action Types stored in the ontology. To this end they must be clustered into *families* which constitute the productive variation of the verb

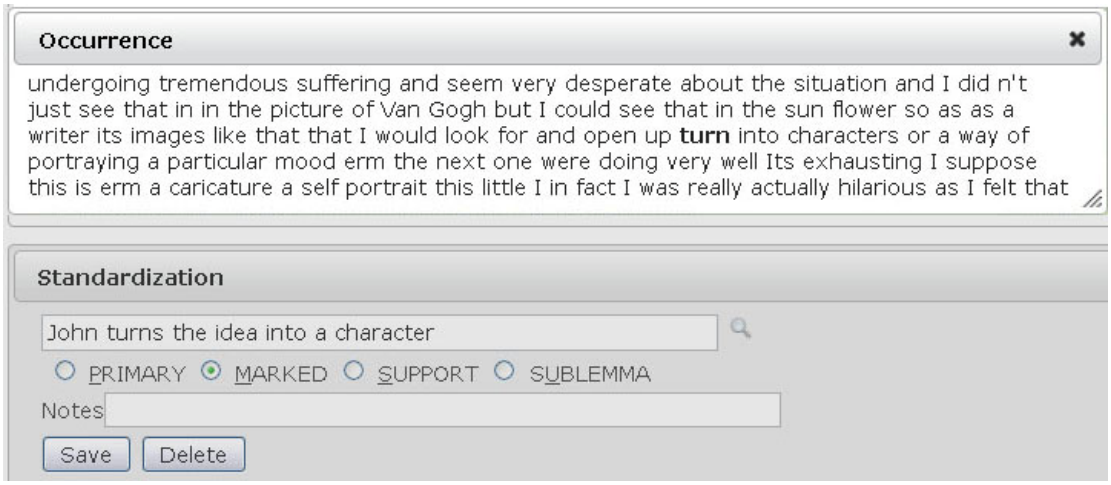

Figure 1. Verb occurrence and Standardization box

predicate. The workflow thus requires the examination of the full set of standardized primary occurrences recorded in the corpus, whose meaning is now clear.

The infrastructure is designed to allow the annotator to create types ensuring both cognitive similarity among their events and pragmatic differences between them. The overall criterion for

type creation is to keep granularity to its minimal level, assigning instances to the same type as long as they fit with one "best example". Clustered sentences should be similar as regards:

- The possibility to extend the occurrence by way of similarity with the virtual image provided by the best example (Cognitive Constraint);
- "Equivalent verbs applied in their proper meaning" i.e. the synset (Fellbaum 1998) (Linguistic Constraints);
- Involved Action schema.

Among the occurrences the annotator chooses the most representative as *best examples* of the recorded variation, creates types headed by one (or more) *best example(s)*, and assigns each individual standardization to a type by dragging and dropping. For instance, standardized occurrences of *to turn* are gathered into Type 3 and Type 5 in Figure 2 because all the occurrences can be respectively substituted by *to direct* and *to stir* and the body schema changes from movement into space to an activity on the object.

The infrastructure assists the annotator in the task by showing the types that have been created so far (on the left side) and the equivalent verbs used to differentiate them (at the bottom). The assigned instances can be shown by type and best example according to the annotator's needs (e.g. Type 3 and Type 5 in the figure). The infrastructure also provides functionality for making easy revisions to hypotheses (show instances not yet assigned, show all instances, verification of Marked variation, editing/merging/splitting types etc.).

The approach underlying the annotation strategy does not require *a priori* any inter-annotator agreement in this core task, which is strongly underdetermined, and rather relies on a supervised process of revision.

Once all occurrences have been processed, the negotiation with a supervisor leads to a consensus on the minimal granularity of the action types extended by the verb in its corpus occurrences. The verification criteria are practical: the supervisor verifies that each type cannot be referred to as an instance of another without losing internal cohesion. The operational test checks if it is understandable that the native speaker is referring to the event in *a* by pointing to the prototype in *b*. The supervisor considers the pragmatic relevance of these judgments and keeps the granularity accordingly.

The relation to images of prototypical scenes

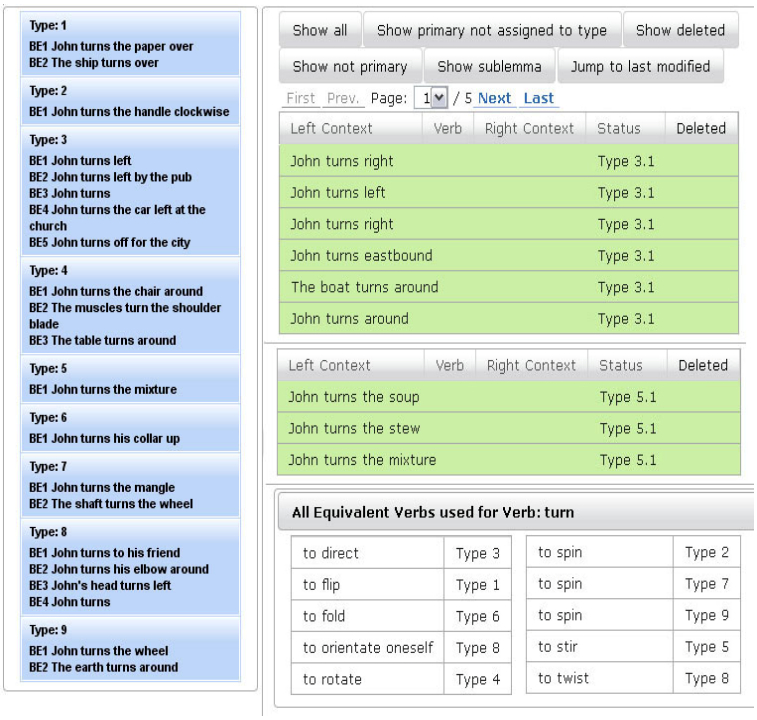

Figure 2 Clustering standardizations into types

provides a challenging question in restricting granularity to a minimal family resemblance set: "can you specify the action referred to by one type as *something like* the best example of another?" .

Granularity is kept when this is not reasonable.

Once types are verified the infrastructure presents the annotator with the "Types Annotation & Assessment" interface. Conversely, in this task the annotator assesses that all instances gathered within each type can indeed be extensions of its best example(s), thus validating its consistency. Those that aren't are assigned to other types.

The assessment runs in parallel with the annotation of the main linguistic features of a type. More best examples can be added in order to represent all thematic structures of the verb which can satisfy that interpretation. As shown in Figure 3 the thematic grid must be filled, by writing each argument in a separate cell and selecting a role-label from the adjacent combo-box. The tag-set for thematic role annotation is constituted by a restricted set of labels derived from current practices in computational lexicons. We are using Palmer's Tagset in VerbNet[1] with adaptations.

Each best example is also annotated with an aspectual class which is assigned by means of the Imperfective Paradox Test (Dowty, 1979). Aspect can assume three values: event, process or state.

Sentences that are judged peripheral instances of the type can be marked, thus identifying fuzziness in pragmatic boundaries. The annotation procedure ends when all proper occurrences of a verb have been assessed. The annotator produces a "script" for each type and delivers the verb annotation to the supervisor for cross-linguistic mapping.

3 Cross-linguistic mapping

Working with data coming from more than one language corpus, IMAGACT must produce a language independent type inventory. For instance, in the case of *to turn* Action types must be consistent with those extended by the Italian verb *girare*, which could be roughly equivalent. Therefore the supervisor will face two lists of types independently derived from corpora annotation. In this scenario, the setting of cross-linguistic relations between verbal entries relies on the identification of a strict similarity between the Types that have been identified (and not through the active writing of a definition). The task is to map similar types onto one prototypical scene that they can be an instance of.

Each prototypical scene is filmed at LABLITA and corresponds to the scripting of one of the best examples selected among all the corpus occurrences which instantiate one Type.

This procedure does not require that the verbs matching onto the same prototypical scene have the same meaning. Two words having different intensions (both within and across languages) may indeed refer to the same action type. The cross-linguistic relation is established accordingly.

Figure 4 roughly sketches the main types derived from the annotation of *to turn* and *girare* and their mapping onto scenes. The supervisor should recognize for instance, that T6 of *girare* and T1 of *to turn* are instances of the same prototype. He will produce one scene accordingly.

The cross-linguistic mapping allows us to predict relevant information which does not emerge from simple corpus annotation. For instance T2 of *girare* never occurs in the English

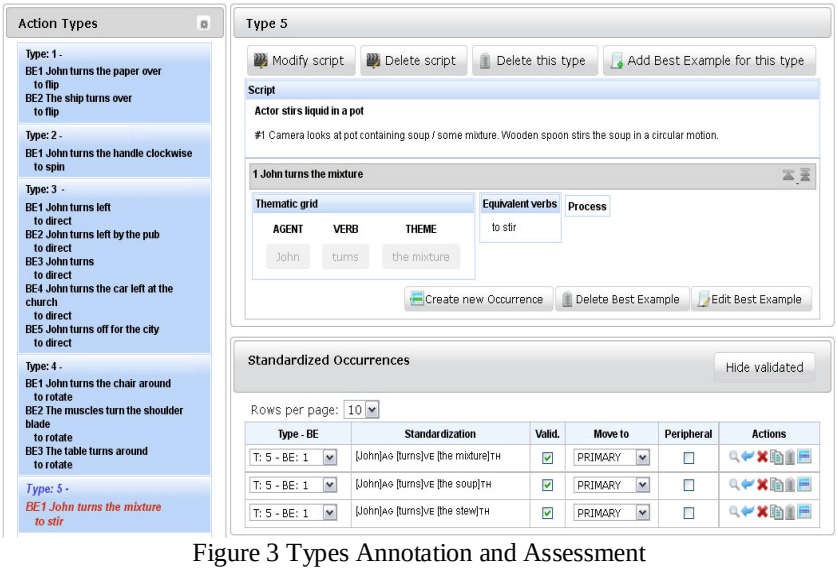

Figure 3 Types Annotation and Assessment

corpus, but native English speakers can recognize from the scene corresponding to T2 that this is also a possible extension of *to turn*. The mapping of the verb onto that type will therefore be established, providing competence based information.

On the contrary, T3 of *girare* and T6 of *to turn* never occur in the English and Italian corpora, however informants recognize that T3 of *girare* cannot be extended by *to turn* (*revolve* is applied) while T6 of *to turn* cannot be extended by *girare* (*alzare* is applied).

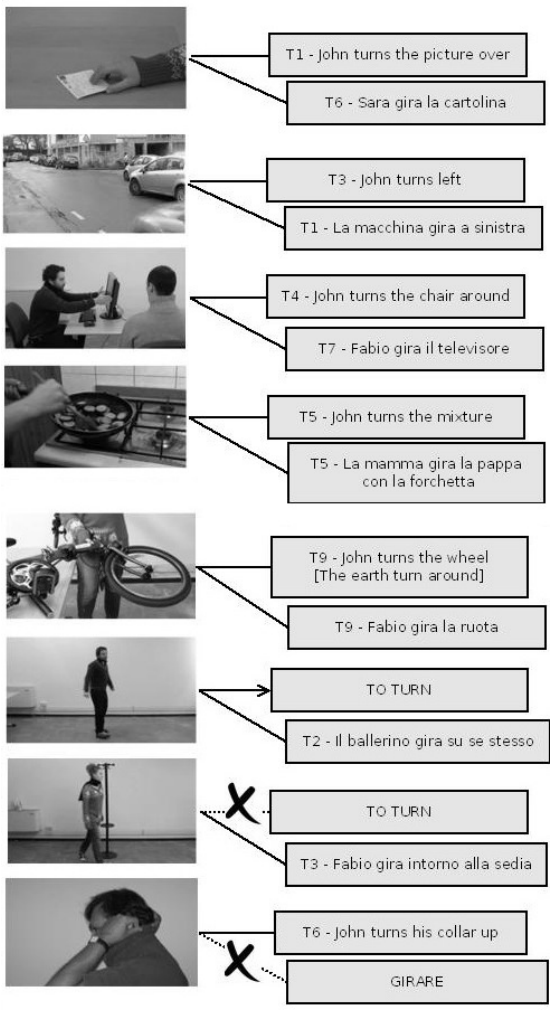

Figure 4. Mapping Action types onto Scenes

In other words the infrastructure and the methodology embodied in it allow the identification of the pragmatic universe of action and of how different languages parse it. This result is obtained in a Wittgenstein-like scenario without the comparison of definitions. The use of prototypical images bypasses this complex problem and permits the identification of the focal pragmatic variation of general verbs and their differentials in different languages.

The link of these scenes to the *sysnets* recorded in WordNet is also carried out when a proper *synset* is available (Moneglia et al. 2012).

Corpora, annotation, lexical variation and cross-linguistic equivalences recorded in each prototypical scene are stored in a database accessed via the web. No annotation format has been so far defined but several current standards in annotation could be relevant here. For the linking between an offset in the corpus and a standardized instance the ISO stand-off annotation format LAF-GrAF could be used. As for the annotation of each standardized instance with syntactic and semantic information (i.e. thematic roles) the ISO MAF and the SemAF could be applicable. Generally speaking, in the framework of the ISO working groups, the IMAGACT annotation procedure as a could be discussed as a possible new work item.

4 Validation and Extension

The direct representation of actions through scenes that can be interpreted independently of language allows the mapping of lexicons from different languages onto the same cross-linguistic ontology. On the basis of this outcome it is possible to ask informants what verb(s) should be applied in his language to each scene and to the set of English and Italian sentences headed by that scene.

Crucially, the informant will verify whether or not the choice is correct for all arguments retrieved from the corpus and assigned to that type and in doing so will verify to which extent the pragmatic concepts stored in the ontology are productive i.e. they permit generalizations at a cross-linguistic level. A concept is valid for cross-linguistic reference to action if, independently of the language, the verb that is applied to the prototypical instance can be also applied to all sentences gathered in it.

The infrastructure organizes this work into two steps: a) alignment of the English and Italian sentences gathered within each entry and generation of a data set of parallel sentences; b) competence based extension (Spanish and Chinese Mandarin). All types in the ontology are checked and all English and Italian action verbs referring to a type will find the appropriate correspondence in

the target languages for that type. The infrastructure allows for the extension to an open set of languages (Moneglia, 2011).

Figure 5 is an example of a competence based extension to Chinese for what regards the second and first scenes of Figure 4. The infrastructure: a) presents the set of sentences gathered into one scene; b) requests the user to input a verb in the target language; c) asks whether or not this verb can be applied in all gathered sentences. The Chinese informant verified that the two scenes require two different verbs (*zhuǎn* and *fān*) which were appropriate in all occurrences.

Distinguishing families of usages of general verbs from the granular variations allows us to establish productive cross-linguistic relations, so validating the Ontology entries in the real world.

Figure 5 Validation & Extension interface

References

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services URL: http://www.natcorp.ox.ac.uk/

CLIPS Corpus. URL: http://www.clips.unina.it

C-ORALROM http://catalog.elra.info/product_info.php?products_id=757

Cresswell M. F. 1978 Semantic Competence in F. Guenthner, M. Guenthner-Reutter, Meaning and translation. NY University Press: New York, 9-28

De Mauro T., Mancini F., Vedovelli M., Voghera M. 1993. Lessico di frequenza dell'italiano parlato (LIP). Milano: ETASLIBRI.

Dowty, D. 1979. Word meaning and Montague grammar. Dordrecht: Reidel.

Fellbaum, Ch. (ed.) 1998. WordNet: An Electronic Lexical Database. Cambridge: MIT Press.

Ide, N. and K, Suderman. 2007.. "GrAF: A graph-based format for linguistic annotations". In *Proceedings of the Linguistic Annotation Workshop at ACL 2007*. Prague, Czech Republic**:** 1-8.

International Organization for Standardization. 2012. ISO DIS 24612- Language Resource Management - Linguistic annotation framework (LAF). ISO/TC 37/SC4/WG 2.

International Organization for Standardization. 2008. ISO DIS 24611 Language Resource Management - Morpho-syntactic Annotation Framework (MAF). ISO/TC 37/SC4/WG 2.

International Organization for Standardization. 2008. ISO DIS 24617- Language Resource Management - Semantic annotation framework (SemAF). ISO/TC 37/SC4/WG 2.

Levin, B. 1993. English verb classes and alternations: A preliminary investigation. Chicago: University of Chicago Press.

Moneglia M. 2011. Natural Language Ontology of Action. A gap with huge consequences for Natural Language Understanding and Machine Translation, in in Z. Vetulani (ed.)Human Language Technologies as a Challenge for Computer Science and Linguistics. Poznań: Fundacja Uniwersytetu im. A. Mickiewicza 95-100.

Moneglia, M., Monachini, M., Panunzi, A., Frontini, F., Gagliardi, G., Russo I. 2012 Mapping a corpus-induced ontology of action verbs on ItalWordNet. In C. Fellbaum, P. Vossen (eds) Proceedings of the 6th International Global WordNet Conference (GWC2012) Brno. 219-226.

Moneglia, M. & Panunzi, A., 2007. Action Predicates and the Ontology of Action across Spoken Corpora. In: M. Alcántara & T. Declerck, Proceeding of SRSL7. Universidad de Salamanca, 51-58.

Rosch, E. 1978. Principles of Categorization. In E. Rosch & B.B. Lloyd (eds), Cognition and Categorization. Hillsdale: Lawrence Erlbaum, 27–48.

VerbNet http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

Wittgenstein, L. 1953. Philosophical Investigations. Oxford: Blackwell.

The Independent Encoding of Attribution Relations

Silvia Pareti
School of Informatics / The University of Edinburgh, UK
S.Pareti@sms.ed.ac.uk

Abstract

Attribution relations have been annotated as discourse relations, attributes of discourse relations, structures carrying factuality, frames for the expression of subjective language, quote–speaker relations and classes of temporal references. While this proves their relevance for different domains, whether as disruptive elements to rule out or essential carriers to treasure, it provides only a limited and marginal picture of this relation. In this paper I will overview its interconnection with other domains, in particular its strong connection with discourse relations, and motivate the need for an independent encoding. I will also highlight what the elements that constitute an attribution relation or contribute to its interpretation are and introduce the attribution corpus developed starting from the annotation in the PDTB.

1 Introduction

The annotation of attribution has been addressed by studies from different domains having however a different annotation focus, e.g. discourse relations (Prasad et al., 2006), sentiments (Wiebe, 2002), factuality (Saurí and Pustejovsky, 2009). Attribution relations (ARs) are relevant for other domains as they can be carriers or constitute themselves informative clues for other phenomena. However, the annotation of attribution has been so far tailored to suit the needs of the 'hosting domain', thus including only aspects and structures relevant to the annotation purpose. For example, the MPQA Opinion Corpus (Wiebe, 2002) has annotated the attribution of speech events, however only when these were a vehicle for private states and only intra–sententially.

All these approaches fail to fully encode attribution and are therefore not suitable to provide a solid basis for attribution studies and to train attribution extraction systems. It is therefore beneficial to separate attribution from other annotation domains and build a resource that can encompass a wider range of attribution structures and reach a more structured and deeper representation of this relation.

In this paper I will explore the interconnections between attribution and discourse and investigate what are the essential traits of an attribution and how it should be encoded. I will start with a brief presentation of the range of domains where attribution is relevant and how they have encoded this relation (Sec. 2). That will provide the framework for taking a closer look at the inclusion of attribution in the PDTB (Prasad et al., 2006) and the overlap and mutual effects of attribution and discourse (Sec. 3.1). However close these two domains are, I will show that there is no exact correspondence between attribution and discourse and that attribution should be annotated as its own relation – an "attribution relation" or AR (Sec. 3.2).

Section 4 will then present the constitutive elements of ARs (Sec. 4.1) and other elements linked to an AR that can contribute to its interpretation (Sec. 4.2). Sec. 4.3 will overview attributes of attribution, the ones that have been included in the PDTB and additional relevant ones that have been included or considered for inclusion in the annotation schema developed for attribution. This has been used to build an attribution corpus starting from the

annotation of ARs in the PDTB.

2 Background

ARs have previously been partially annotated in the context of annotating other phenomena of interest to language processing. This work has only marked the portion of attribution of interest for the main task at focus (e.g. the annotation of discourse relations or event factuality). In this section I will survey some of the most prominent annotation efforts that have included attribution and highlight how their approach has encoded attribution and the perspective they have taken at it.

A portion of ARs has been addressed and annotated by studies dealing with 'subjectivity analysis'. A subset of ARs, namely opinions and beliefs, are part of the 'private states' at focus in the MPQA Opinion Corpus (Wiebe, 2002). Despite a strong overlap in scope, the approach is considerably different. While a private state is defined as "an experiencer holding an attitude, optionally toward an object" (Wiebe, 2002, p.4), attribution goes in the opposite direction. The object is not optional, but a fundamental element of the AR, intended as "a relation of 'ownership' between abstract objects and individual or agents" (Prasad et al., 2008, p.40).

Discourse studies encode ARs annotating two elements: the *attributed span* and the **attribution span**, as in Ex.(1)[1]. When attribution itself is considered as a discourse relation (Carlson and Marcu, 2001; Wolf and Gibson, 2005), these two annotated elements correspond to discourse units. Attribution holds from the attributed span, *nucleus*, towards the attribution span, *satellite*.

(1) **Mr. Englund added** *that next month's data isn't likely to be much better, because it will be distorted by San Francisco's earthquake*. (wsj_0627)

Studies concerned with the attribution of direct quotes, e.g. the Sydney Morning Herald Corpus (O'Keefe et al., 2012), also annotate attribution as composed by two elements, i.e. *quote*–speaker pairs (Ex.(2)). The element connecting speaker and quote and expressing the type of AR (e.g. assertion or belief) is not annotated. However, the attribution of quotes implies that what is attributed is an assertion.

(2) *"The employment report is going to be difficult to interpret,"* said Michael Englund, economist with MMS International, a unit of McGraw-Hill Inc., New York. (wsj_0627)

The textual anchor establishing the relation is annotated by some studies (Glass and Bangay, 2007; Pouliquen et al., 2007), however as a device helping the identification and therefore extraction of an AR and not as integral part of the relation itself. In particular, speech verbs (e.g. say, report) are identified as their grammatical subject often expresses the source entity of the AR and their object the attributed element.

ARs also affects temporal references, and 'reporting' has been included as an event class in TimeML (Pustejovsky et al., 2003) and reporting events have been annotated in TimeBank (Pustejovsky et al., 2006). Accounting for the relation between the time the document was produced and that of the reporting event remained an issue. ARs insert an additional point in time, i.e. that of the enunciation, in case of an assertion or the temporal point where a belief or fact was factual. For example, 'John thought it was a good idea' reflects John's belief at a past point in time. This belief might have changed at the point the article was written or the present time.

Attribution has also strong implications for the factuality of the events expressed in the attributed span. This motivates its partial inclusion in FactBank (Saurí and Pustejovsky, 2009) where the attributed span itself is not marked, but events contained in it (e.g. 'left' in Ex.(3)) are linked to their source by source–introducing predicates (SIPs) in order to derive their factuality. The SIP in Ex.(3) implies that the event underlined in the example is considered by the source as just a possibility.

(3) Berven **suspects** that Freidin left the country in June. (Saurí and Pustejovsky, 2009, p.236)

3 Attribution and Discourse

Attribution is intertwined with other annotation domains. In particular, it overlaps and has implications

[1] The attribution span is highlighted in bold in the examples, while the attributed span is in italics. Examples taken from the WSJ (WSJ article reference in brackets)

relevant to discourse relations, factuality and subjectivity analysis, as briefly introduced in Sec. 2.

The PDTB is the biggest existing resource annotating ARs. However, what makes it a suitable starting point to study attribution is it has not first defined a strict set of rules that attribution should obey to be considered in the scope of the project, thereby restricting attribution to its 'pretty' and more standard structures. This, combined with the size of the corpus, means that a wide range of attribution structures can be observed. For example, attributions to unnamed or implicit entities or having no reporting verb. However, I will argue that attribution should be treated and annotated independently and motivate the effort to disjoint it from discourse annotation.

3.1 Intertwined

Attribution relations are closely tied to discourse relations, and have variously been included as a discourse relation itself (Wolf and Gibson, 2005; Carlson and Marcu, 2001) or as an attribute of discourse relations (Prasad et al., 2006). They were included in the PDTB since it was recognised that "a major source of the mismatches between syntax and discourse is the effect of attribution" (Dinesh et al., 2005, p.36).

If the arguments of a discourse connective are taken to be its syntactic arguments, attribution could lead to incorrect semantic interpretation as in Ex.(4) below (Prasad et al., 2008, p.2966). It is therefore important to recognise and exclude attribution in such cases.

(4) a. Factory orders and construction outlays were largely flat in December [Arg1.]

b. while **purchasing agents said** [Conn.]

c. *manufacturing shrank further in October* [Arg2.]. (wsj_0178)

While attribution is disruptive for discourse relations, these could be of great advantage to the identification of the *content*, i.e. the attributed span when the AR is indirect, i.e. the attributed span, is not surrounded by quote markers. While some studies (Skadhauge and Hardt, 2005; de La Clergerie et al., 2009) have taken an intra–sentential look at attribution and considered as the content of an AR the grammatical object of a reporting verb, this is not a viable solution when dealing with a wider range of ARs. Here discourse structure may play a role above the level of single sentences.

The ARs collected from the PDTB show that around 17% of ARs extend over more than one sentence (e.g. three sentences in Ex.(5)). Moreover, only half of these are attributions of direct quotes. English does not mark indirect reported speech grammatically, unlike for example German (Ruppenhofer et al., 2010), where this is associated with subjunctive mood. The issue is how to determine the content span boundaries of indirect ARs when the syntactic structure would be of no help. While not always unambiguous also for human readers, recognising a content extending over more sentences could be partly achieved with the help of discourse relations.

(5) **According to Audit Bureau of Circulations**, *Time, the largest newsweekly, had average circulation of 4,393,237, a decrease of 7.3%. Newsweek's circulation for the first six months of 1989 was 3,288,453, flat from the same period last year. U.S. News' circulation in the same time was 2,303,328, down 2.6%.* (wsj_0012)

In Ex.(5), the last two sentences are a continuation of the content but they bear no syntactic relation with the first sentence. Instead, they are two discourse relations (both entailing an implicit connective *and*, of type Comparison:Contrast:Juxtaposition) binding the first part of the content span with the second and the third sentence. Discourse alone might not provide sufficient evidence to determine the content extension. Nonetheless, in combination with other triggers, e.g. verb tense and mood, this could allow the correct identification of inter–sentential indirect ARs.

3.2 Distinct

The PDTB is rich in attribution annotation and represents a great starting point for the collection of a large resource for the study of attribution. However, what is annotated is not attribution itself but the attribution of discourse connectives and their arguments. Attribution is therefore subordinate to discourse and reconstructing a full AR can be rather complex.

The content of an AR might mot be fully corresponding to a discourse relation or one of its arguments, but be composed of several discourse connectives and their arguments. We can consider, for example, the marked AR that corresponds to the second paragraph of the excerpt below (wsj_0437):

> The reports, attributed to the Colombian minister of economic development, said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags, the analyst said.
>
> (HOWEVER) *These reports were later denied by a high Brazilian official, who said Brazil wasn't involved in any coffee discussions on quotas*, **the analyst said**. (wsj_0437_12[2])
>
> (BUT) The Colombian minister was said to have referred to a letter that he said President Bush sent to Colombian President Virgilio Barco, and in which President Bush said it was possible to overcome obstacles to a new agreement.

The content span of this AR, the text in italics, is partially included in all three discourse relations below: the two implicit ones, having *however* and *but* as connectives, and the one with discourse connective *later*. In order to reconstruct the full AR from the annotation, it was necessary to take all three discourse relations into account and merge together the text spans they were attributing to 'the analyst said'.

1. The reports said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags (Arg1)

   HOWEVER (Implicit connective)

   These reports were later denied by a high Brazilian official (Arg2)

2. The reports said Brazil would give up 500,000 bags of its quota and Colombia 200,000 bags (Arg1)

   LATER (Connective)

   These reports were denied by a high Brazilian official (Arg2)

3. *who said Brazil wasn't involved in any coffee discussions on quotas* (Arg1)

   BUT (Implicit connective)

   The Colombian minister was said to have referred to a letter that he said President Bush sent to Colombian President Virgilio Barco, and in which President Bush said it was possible to overcome obstacles to a new agreement (Arg2)

[2]Examples from the attribution corpus report the AR unique ID.

This shows that there is no exact correspondence between ARs and discourse arguments and therefore some ARs are partially or not annotated. This happens if part of their content is not corresponding to a discourse argument or when the whole AR is included in a discourse argument as in Arg1 of *But* (relation 3 above). The nested AR (i.e. '**who said** *Brazil wasn't involved in any coffee discussions on quotas*') in this attribution argument is just not annotated.

While the PDTB is a great resource for attribution, attribution cannot be handled as a mere attribute of discourse connectives and their arguments as there is no exact correspondence between ARs and discourse relations. I have therefore disjoint the annotation of discourse and attribution by collecting the ARs in the PDTB and reconstructing incomplete ARs, thus creating a separate level of annotation.

4 The Independent Encoding of Attribution

ARs are encoded in the PDTB as formed by two elements, the attributed material, i.e. abstract object or discourse units, and the attribution span. I will argue that this encoding of ARs is not sufficient and cannot suit the variety of purposes attribution could serve. It does not allow, for example, to easily identify attributions to a specific source. In the next section I will present which are the core elements of this relation, which are additional and the attributes that we can associate with ARs.

4.1 Constitutive Elements of ARs

There are three elements necessary to define the relation of attribution based on textual evidence. These elements are the two that are related, i.e. the attributed material or *content* and the entity this is attributed to, the source, which may or may not correspond to the author of the article, but also the link connecting them, i.e. the **cue**. Annotating the cue is fundamental as this represents the key to the correct identification and interpretation of the relation

it establishes. Is the AR in Ex.(6a)[3] a statement or an opinion? Is it factual or just a speculation? Does the AR in Ex.(6b) entail that the source or the author believe in the truth of the proposition in the content?

(6) a. Network officials involved in the studio talks **may hope** *the foreign influx builds more support in Washington*, but that seems unlikely. (wsj_2451.pdtb_09)

b. *"He taught me how to play like a gypsy,"* **jokes** the musician. *"I didn't learn to count until I got to Juilliard."* (wsj_1388.pdtb_02)

Although source, cue and content are constitutive elements of ARs, they can possibly be only implicitly or anaphorically expressed as in Ex.(7), where the source is implicit and the content anaphorically recalled by a pronoun.

(7) *[. . .] profound change toward free-market economics, especially in the statist countries*. **Having said** *that*, we must caution against an apparent tendency to overstate the case. (wsj_1529)

In order to encode all the constitutive elements of an AR independently, I had to further annotate the attribution corpus collected from the PDTB. The text labelled as attribution span was therefore further annotated with the source and cue elements of the AR. However, these were not the only elements constituting the attribution span.

4.2 Other Relevant Components of ARs

Beside the constitutive elements of ARs, the surrounding context can carry further information relevant to the AR, although optional. When the attribution span contains relevant elements that are neither part of the source nor of the cue, these should be marked as SUPPLEMENTAL. In particular, supplemental elements are those providing a context for interpreting an AR, including its:

- setting (time, place, audience) (Ex.(8)[4]);
- topic (Ex.(9));
- communication medium (Ex.(10));
- relevance to the author's argument (Ex.(11));
- manner (Ex.(12)).

(8) *"Ideas are going over borders, and there's no SDI ideological weapon that can shoot them down,"* he **told** [A GROUP OF AMERICANS] [AT THE U.S. EMBASSY] [ON WEDNESDAY]. (wsj_0093_07)

(9) OF SONY, Mr. Kaye **says**: "*They know there's no way for them to lose. They just keep digging me in deeper until I reach the point where I give up and go away.*" (wsj_2418_15)

(10) Trade and Supply Minister Gerhard Briksa **said** IN A LETTER PUBLISHED IN THE YOUTH DAILY JUNGE WELT *that the rise in alcohol consumption in East Germany had been halted*; (wsj_1467_05)

(11) AS AN INDICATOR OF THE TIGHT GRAIN SUPPLY SITUATION IN THE U.S., market analysts **said** *that late Tuesday the Chinese government, which often buys U.S. grains in quantity, turned instead to Britain to buy 500,000 metric tons of wheat*. (wsj_0155_16)

(12) "A very striking illusion," Mr. Hyman **says** NOW, HIS VOICE DRIPPING WITH SKEPTICISM, "but an illusion nevertheless."(wsj_0413_14)

If part of the attribution span, these elements have been included in the annotation of the attribution corpus, with the label 'supplement'. The information contained in the supplement might still not be sufficient to fully evaluate and fully understand an AR. In Ex.(12) we don't know what the source considers an 'illusion', i.e. the topic this assertion is about. Nonetheless, the supplement usually provides enough elements for the interpretation of the AR. This without having to process the whole article or resorting to external knowledge.

4.3 Features of Attribution Relations

There are several features relevant for encoding ARs. Features that can capture if an AR is factual or contribute to determine whether the attributed

[3] From now on, examples will mark the cue of an AR in bold, the source underlined and the content in italics.

[4] Supplements are represented in the examples in small capitals.

proposition is truthful, differentiate sources and attributions. These features can enable applications of attribution beyond the retrieval of ARs having a specific source or cue. The PDTB annotates four such features. One is the *type of attribution*, i.e. belief, assertion, fact or eventuality. This affects the factuality of the content since in an AR of type 'fact' this is higher, and it usually implies that the source and author believe it is truthful, while in an attributed belief the level of factuality is much lower as in Ex.(13). The source is not sure about the proposition expressed in the content being really true.

(13) Meanwhile, some U.S. officials **fear** *PLO chief Arafat is getting cold feet and may back off from his recent moderation and renunciation of terrorism*.(wsj_1682_00)

A second feature of ARs in the PDTB is the *type of source*, i.e. writer, other or arbitrary. This aspect allows to distinguish between real and 'pseudo-attributions'. In the latter the attribution is not to a third party but to the writer or author of the article, who is the default source of the whole article, and thus redundant.

There are other two attributes, *determinacy* and *scopal polarity*, accounting for the factuality of the AR (Ex.14a) and the polarity of its content respectively (Ex.14b). While in the first example the AR is just an hypothesis, therefore not factual, in the second one the AR itself is factual, the content being in the scope of the negation instead.

(14) a. [. . .] BY NEXT WEEK the network **may announce** *"Teddy Z"is moving to 8:30 p.m. from its 9:30 time slot*[. . .] (wsj_1150_00)

b. DEPOSITS **aren't expected** *to exceed withdrawals in the foreseeable future*, as the industry continues to shrink. (wsj_1293_03)

Beside the features already included in the PDTB, ARs carry other relevant ones worth annotating. As noted by (Karttunen and Zaenen, 2005), the attribution cue can indicate the *authorial stance*, i.e. the position the author takes towards the truth of the proposition expressed in the content. By choosing to use a factive (e.g. admit, regret, realise) or counter–factive cue (e.g. lie, joke (Ex.6b)), the author implies a certain degree of commitment or non–commitment towards the truth of the attributed statement. Using a non–factive cue (e.g. say, claim, suggest), the author remains instead more neutral. The authorial stance is a relevant feature of ARs as the commitment the author expresses towards the statement can be employed to uncover ideological biases or, if we assume the author to be trustworthy, to determine if the statement is truthful.

Attribution cues can also express the *source attitude*, i.e. the sentiment the source itself expresses towards the proposition, e.g. 'negative' in Ex.(13) and positive in Ex.(15). While the most frequent reporting verbs (e.g. say) tend to be neutral, other verbs normally not associated with a reporting meaning, and in particular manner verbs (e.g. smile, quip, purr), can express this feature.

(15) "*We've had the Russians and Chinese, and people from India visiting us*," Mr. Iverson **beams**. "*Everyone in the world is watching us very closely*." (wsj_2153_01)

These features have not yet been included in the annotation as a preliminary inter-annotator agreement study showed that their definition needs further investigation. In this study, two expert annotators applied the annotation schema (Pareti, 2011) to 14 articles from the WSJ corpus (380 jointly identified ARs) and assigned them values for the four features annotated in the PDTB and the two additional ones I have proposed. Cohen's Kappa values for the correct selection of the value for authorial stance (i.e. committed, not_committed, neutral) and source attitude (i.e. positive, negative, tentative, neutral, other) were .48 and .20 respectively.

Other features do not require manual annotation as they can be derived from lexical and syntactic clues of the AR elements — for example, whether a source is a group or an individual, named or unnamed. Another automatically derivable feature is whether the attribution content is completely (direct AR), partly (mixed AR) or not at all (indirect AR) surrounded by quotation markers. This feature was called "quote status" (Pareti, 2012) and included in the attribution corpus developed. It is relevant not only because direct quotes are generally used to reflect the exact words uttered by the source, and are

thus more faithful to the original statement, but also because they tend to occur with different syntactic structures and lexical choices. For example, in the attribution corpus collected, the verb cue 'suggest' never occurs in the context of a direct attribution, while 'joke' always associates with a direct quote.

5 Conclusion and Future Work

This paper overviews the importance of ARs in different domains. ARs can carry temporal events and subjective expressions, affect the factuality of events and cause a mismatch between syntactic and discourse arguments of discourse connectives. However, annotating ARs 'ad hoc', as part of other annotation projects, is rather detrimental as it prevents attribution from being encoded in an independent and more complete way.

While the PDTB represents a fundamental source of attribution annotation, I have shown the limitations of such annotation and proved the need for an independent encoding of attribution. For this reason, I have created an independent corpus of ARs starting from the annotation in the PDTB. This was done by separating the annotation of ARs from that of discourse relations and further annotating each AR according to a previously developed annotation schema. This resource could enable reaching a deeper understanding of ARs and allow the development of AR extraction systems that can be reliably employed (e.g. for information extraction or multi–perspective QA). The independent encoding would also allow projects from other domains to rely on the annotation for the portion relevant to the phenomenon at study.

The attribution corpus in its first version is in a flat CoNLL style — i.e. each line corresponds to one AR and each column to one element, feature or pointer of the AR. I am currently developing an XML format for AR annotation, which allows for the representation of nested ARs.

Acknowledgements

The author is supported by a Scottish Informatics & Computer Science Alliance (SICSA) studentship.

References

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report ISITR- 545. Technical report, ISI, University of Southern California, September.

Eric de La Clergerie, Benot Sagot, Rosa Stern, Pascal Denis, Gaelle Recource, and Victor Mignot. 2009. Extracting and visualizing quotations from news wires. In *Proceedings of L&TC 2009, Poznan, Poland*.

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (non-)alignment of syntactic and discourse arguments of connectives. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 29–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Glass and Shaun Bangay. 2007. A naive, saliencebased method for speaker identification in fiction books. In *In Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 07)*, pages 1–6, November.

Lauri Karttunen and Annie Zaenen. 2005. Veridicity. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, Extracting and Reasoning about Time and Events*, number 05151 in Dagstuhl Seminar Proceedings, Schloss Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI).

Tim O'Keefe, Silvia Pareti, James Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. Manuscript submitted for publication.

Silvia Pareti. 2011. Annotating attribution relations and their features. In Kamps J. Karlgren J. Alonso, O., editor, *ESAIR'11: Proceedings of the CIKM'11 Workshop on Exploiting Semantic Annotations in Information Retrieval*. ACM Press, October.

Silvia Pareti. 2012. A database of attribution relations. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ug(ur Dog(an, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of the International Conference Recent Advances In Natural Language Processing (RANLP 2007)*, pages 487–492.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. Annotating attribution in

the Penn Discourse TreeBank. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 31–38.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation LREC08*.

James Pustejovsky, Jos Castao, Robert Ingria, Roser Saur, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of IWCS-5, Fifth International Workshop on Computational Semantics*.

James Pustejovsky, Jessica Littman, Roser Saur, and Marc Verhagen. 2006. Timebank 1.2 documentation. Technical report.

Josef Ruppenhofer, Caroline Sporleder, and Fabian Shirokov. 2010. Speaker attribution in cabinet protocols. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*. European Language Resources Association (ELRA).

Roser Saurí and James Pustejovsky. 2009. Factbank: A corpus annotated with event factuality. *In Language Resources and Evaluation*, (43):227–268.

Peter R. Skadhauge and Daniel Hardt. 2005. Syntactic identification of attribution in the RST treebank. In *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora*.

Janyce Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Technical report, University of Pittsburgh.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Comput. Linguist.*, 31:249–288, June.

IARG-AnCora: Annotating AnCora corpus with implicit arguments

Aina Peris, Mariona Taulé
CLiC-Centre de Llenguatge i Computació
University of Barcelona
Gran Via 585, 08007 Barcelona
{aina.peris,mtaule}@ub.edu

Horacio Rodríguez
TALP Research Center
Technical University of Catalonia
Jordi Girona Salgado 1-3, 08034 Barcelona
horacio@lsi.upc.edu

Abstract

IARG-AnCora is an ongoing project whose aim is to annotate the implicit arguments of deverbal nominalizations in the AnCora corpus. This corpus will be the basis for systems of automatic Semantic Role Labeling based on Machine Learning techniques. Semantic Role Labelers are essential components of current language technology applications in which it is important to obtain a deeper understanding of the text in order to make inferences at the highest level in order and thereby obtain qualitative improvements in the results.

1 Introduction

Traditionally, the analysis of argument structure has been focused basically on verbal predicates, although it has recently been extended to nominal predicates. Most of the efforts at argument identification are restricted to those arguments that appear in the sentence, in the case of verbs, or in the Noun Phrase (NP), in the case of nouns. In a nutshell, they are focused on the identification of explicit arguments. Furthermore, Semantic Role Labeling (SRL) systems are verb-centered and reduce role labeling to explicit arguments (Márquez et al., 2008; Palmer et al., 2010). In order to move forward to the full comprehension of texts, it is necessary to take into account implicit arguments and to widen the context of analysis to the whole discourse (Gerber et al., 2009). This is especially important in the case of deverbal nominalizations since the degree of optionality of their explicit arguments is higher than for verbs.

The aim of *IARG-AnCora* is to enrich the Spanish and Catalan AnCora corpora[1] by annotating the implicit arguments of deverbal nominalizations. Currently, AnCora corpora are only annotated with arguments inside the NP of these deverbal nouns. AnCora consists of a Catalan (AnCora-Ca) and Spanish (AnCora-Es) corpora of 500,000 words each, annotated at different linguistic levels: morphology (Part of Speech, PoS, and lemmas), syntax (constituents and functions), semantics (verbal and deverbal nouns argument structure, named entities and WordNet senses), and pragmatics (coreference). The main goal is to identify implicit arguments and assign an argument position –iarg0[2], iarg1, etc.– and a thematic role (agent, patient, cause, etc.) to them. These arguments can be recovered if a wider discursive context is taken into account and their identification is therefore important to provide a deep semantic representation of sentences and texts.

2 Defining an Implicit Argument

We define an implicit argument as the argument which is not realized in the NP headed by the deverbal nominalization, but is realized instead inside (1) or outside the sentence (2) context[3]. However, the implicit argument can sometimes be inside the NP

[1] AnCora corpora are freely available at: http://clic.ub.edu/corpus/ancora.

[2] The letter 'i' at the beginning of the argument position stands for implicit argument. We note the implicit arguments as iarg<position>-<thematic role>.

[3] We focus our definition of implicit arguments on deverbal nominalizations because we deal with them in our work. However, it is worth saying that verbs can also have implicit arguments.

as long as the constituent associated to this implicit argument does not depend directly on the nominalization. For instance, constituents inside a subordinate clause complementing the deverbal noun can be implicit arguments (3) of this deverbal noun.[4]

(1) [Las escuelas de samba de Sao Paulo]$_{\text{iarg1-pat}}$ han conseguido [el **apoyo**[5] [de la empresa privada]$_{\text{arg0-agt}}$ para mejorar las fiestas de carnaval]$_{\text{NP}}$.
*[Schools of samba in Sao Paulo]$_{iarg1\text{-}pat}$ got [the **support** [of private industry]$_{arg0\text{-}agt}$ to improve Carnival celebrations]$_{NP}$.*

(2) [El carnaval de Sao Paulo es feo]$_{\text{iarg1-pat}}$, dijo hoy [el alcalde de Río de Janeiro]$_{\text{iarg0-agt}}$ en una conversación informal con periodistas cariocas, y encendió la polémica. [...] [Esa **opinión**[6]]$_{\text{NP}}$ fue respaldada por el gobernador de Río de Janeiro, quien incluso fue más allá en su crítica al comentar que el carnaval que se organiza en Sao Paulo es "más aburrido que un desfile militar".
*[The Carnival of Sao Paulo is ugly]$_{iarg1\text{-}pat}$, said [the mayor of Rio de Janeiro]$_{iarg0\text{-}agt}$ in an informal conversation with Carioca journalists, and ignited the controversy. [...] [This **opinion**]$_{NP}$ was supported by the governor of Rio de Janeiro, who went even further in his criticism when he commented that the carnival held in Sao Paulo is"more boring than a military parade".*

(3) [El **daño** [causado a [su industria aeronáutica]$_{\text{iarg1-tem}}$[7]]$_{\text{Subordinate C}}$]$_{\text{NP}}$.
*[The **damage** [caused to [its aeronautic industry]$_{iarg1\text{-}tem}$]$_{Subordinate\ C}$]$_{NP}$.*

[4]In NomBank, these cases are annotated as arguments outside the domain of locality, and are therefore not treated as implicit arguments (Meyers, 2007). We only consider explicit arguments to be those that depend directly on the nominal predicate.

[5]In AnCora corpus, 'conseguir apoyo' is not considered to be a support verb construction because the verb is not semantically bleached and it holds a predicating power (Hwang et al., 2010), so 'apoyo' is annotated as the object of 'conseguir' and they are treated as independent predicates.

[6]In Spanish, the noun 'opinión', *opinion*, is derived from the verb 'opinar', *to express an opinion*.

[7]The label 'tem' stands for theme.

Example (1) shows the deverbal nominalization 'apoyo' *support* with the agent argument ('de la empresa privada', *of private industry*) realized inside the NP, whereas the patient argument ('las escuelas de samba de Sao Paulo', *schools of samba in Sao Paulo*) is realized in the same sentence but outside the NP. In (2), the nominalization 'opinión', *opinion*, appears without any explicit argument in the NP. However, the agent argument ('el alcalde de Río de Janeiro', *the mayor of Rio de Janeiro*) as well as the patient argument ('el carnaval de Sao Paulo es feo', *the carnival of Sao Paulo is ugly*) are realized implicitly (iarg0-agt and iarg1-pat, respectively) in the previous sentence. Currently, the AnCora corpus is only annotated with arguments inside the NP, therefore 'opinión' *opinion* has no associated argument and 'apoyo' *support* only has the *agent* argument annotated. In example (3), the implicit argument of 'daño' *damage*, iarg1-tem, is the 'industria aeronáutica' (*aeronautic industry*), which is a constituent inside the subordinate clause.

3 Corpora annotated with implicit arguments

As far as we know, the only two corpora with nominal implicit arguments have been developed for English and they have been used as training data for the works presented in (Ruppenhofer et al., 2010) and (Gerber and Chai, 2010):

- The training and test corpus developed for SemEval-2010 task 10[8], *Linking events and their participants in discourse* (Ruppenhofer et al., 2010). A corpus that consists of literary texts annotated following FrameNet-style.

- A subset of the standard training, development, and testing sections of the Penn TreeBank (Marcus et al., 1993) used in (Gerber and Chai, 2010). The annotation scheme follows PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004; Meyers, 2007) proposals.

The number of occurrences annotated is 3,073 in the former, where each nominal predicate had a very small number of occurrences, and 1,253 in the latter,

[8]http://www.coli.uni-saarland.de/projects/semeval2010_FG/.

where only the ten most frequent unambiguous noun occurrences are annotated in order to avoid the problem of sparseness presented in the SemEval-2012 task 10 corpus. Both corpora are annotated only with core arguments (no adjuncts arguments).

IARG-AnCora will be the first corpus annotated with implicit arguments in Spanish and Catalan. In contrast to the English corpora, IARG-AnCora will have an extended coverage in two senses: on the one hand, all the implicit arguments of all deverbal nominalization occurrences in the corpus AnCora (approximately 19,000 for each language) will be annotated; on the other hand, we will take into account the core arguments (arg0, arg1, arg2, arg3 and arg4) as well as the adjunct arguments (argM).

4 Methodology

We will annotate the implicit arguments of AnCora in three steps combining automatic and manual processes.We have already completed the first step and now we are focused on the second.

(a) First, we have developed a manually annotated training corpus consisting of 2,953 deverbal noun occurrences in AnCora-Es. These occurrences correspond to the 883 unambiguous deverbal nominalization lemmas, that is, to those that have only one sense (with only one roleset associated) in AnCora-Nom (Peris and Taulé, 2011a). In order to ensure the quality and the consistency of the annotated data, an inter-annotator agreement test has been conducted on a subsample of 200 occurrences. The average pairwise result obtained between the three pairs of annotators was 81% of observed agreement (58.3% Fleiss kappa (Fleiss, 1981)). The features for the classification model will be inferred from this training corpus.

(b) Second, we will develop an implicit argument SRL model based on Machine Learning (ML) techniques, whose purpose is the automatic identification and classification of implicit arguments. We will use this model to automatically annotate the implicit arguments of the whole AnCora-Es. Afterwards, we will adapt this model and apply it to Catalan (AnCora-Ca) in order to analyze its transportability[9].

(c) Finally, a manual validation of the automatically annotated corpus will be carried out in order to ensure the quality of the final resource. This manual validation will allow for the evaluation of the precision and recall of the automatic system developed.

In the automatic and the manual processes, we use the verbal and nominal lexicons -AnCora-Verb (Aparicio et al., 2008) and AnCora-Nom- as lexical resources to obtain the information about the possible implicit arguments for each predicate. The candidate arguments to be localized in the local discursive context, and to be thereafter annotated, are those specified in the nominal or verbal lexical entries and not realized explicitly.

4.1 Annotation Scheme

We use the same annotation scheme as the one followed to annotate the explicit arguments of deverbal nouns (Peris and Taulé, 2011b), and the argument structure of verbs in AnCora (Taulé et al., 2008), which was in turn based on PropBank and NomBank. In this way, we ensure the consistency of the annotation of arguments of different predicates -nouns and verbs-, as well as the compatibility of Spanish and Catalan resources with English resources.

We use the iarg$_n$ tag to identify implicit arguments and to differentiate them from explicit arguments (arg$_n$) (Gerber and Chai, 2010). The list of thematic roles includes 20 different labels based on VerbNet (Kipper, 2005) proposals: agt (agent), cau (cause), exp (experiencer), scr (source), pat (patient), tem (theme), cot (cotheme), atr (attribute), ben (beneficiary), ext (extension), ins (instrument), loc (locative), tmp (time), mnr (manner), ori (origin), des (goal), fin (purpose), ein (initial state), efi (final state), and adv (adverbial).

The combination of the six argument positions labels (iarg0, iarg1, iarg2, iarg3, iarg4, iargM) with the different thematic roles results in a total of 36 possible semantic tags (iarg0-cau, iarg1-agt, iarg0-agt, iarg2-loc, etc.).

[9] Our guess is that the model learned in Spanish can be adapted directly to Catalan.

4.2 Annotation Observations

From the data annotated (2,953 deverbal noun occurrences), we can highlight that implicit arguments in Spanish are more frequent than explicit arguments in nominal predicates. The average number of implicit arguments realized among the predicates analyzed, taking into account core and adjunct arguments, is almost two implicit arguments per instance (1.9). Therefore, the annotation of implicit arguments is crucial for the semantic treatment of deverbal nominalizations and implies a gain in role coverage of 317%[10]. Specifically, the core arguments arg0-agt/cau, arg1-pat/tem and arg2-ben/atr are those more frequently realized as implicit arguments.

Another relevant conclusion is that most implicit arguments are located nearby. From the total number of implicit arguments annotated, 60% are located within the sentence containing the nominal predicate, 32% are found within the previous context and 8% in the following context. Similar observations are drawn for English in (Gerber and Chai, 2012).

5 Conclusions

This project will give rise, on the one hand, to an enriched version of AnCora corpora with the annotation of the implicit arguments of deverbal nouns and, on the other hand, to the first available model of SRL dealing with implicit arguments in Spanish and Catalan.

IARG-AnCora will be the first corpus in these languages to be annotated with explicit and implicit arguments for deverbal noun predicates, with a high coverage available to the research community. This resource follows the same annotation scheme as NomBank and PropBank for argument structure, and as (Gerber and Chai, 2010; Gerber and Chai, 2012) for implicit arguments. In this way, we ensure the compatibility of the Spanish and Catalan resources with those that are also based on this annotation scheme. In fact, we aim to create interoperable semantic resources.

IARG-AnCora will be an important resource of semantic knowledge that could be used as a learning corpus for SRL nominal systems. It will also be a useful resource for linguistics studies on the argument structure of deverbal nominalizations or on coreference chains and the entities referring to NPs.

[10]This figure is extremelly higher than the reported for English (71%) in (Gerber and Chai, 2012) due to the lower degree of instantiation of explicit arguments.

References

Juan Aparicio, Mariona Taulé, and M.Antònia Martí. 2008. AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 797–802, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Joseph L. Fleiss. 1981. *Statistical methods for rates and proportions*. John Wiley.

Matthew Gerber and Joyce Y. Chai. 2010. Beyond NomBank: a study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1583–1592, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Gerber and Joyce Y Chai. 2012. Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Computational Linguistics*. To appear.

Matthew Gerber, Joyce Chai, and Adam Meyers. 2009. The role of implicit argumentation in nominal srl. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 146–154, Boulder, Colorado, June. Association for Computational Linguistics.

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 82–90, Uppsala, Sweden, July. Association for Computational Linguistics.

K. Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, PA.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Adam Meyers, Ruth Reeves, and Catherine Macleod. 2004. NP-external arguments a study of argument sharing in English. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE*

'04), pages 96–103, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Meyers. 2007. Annotation Guidelines for NomBank Noun Argument Structure for PropBank. Technical report, University of New York.

Lluis Márquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):76–105.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling. Synthesis on Human Languages Technologies.* Morgan ang Claypool Piblishers.

Aina Peris and Mariona Taulé. 2011a. AnCora-Nom: A Spanish Lexicon of Deverbal Nominalizations. *Procesamiento del Lenguaje Natural.*, 46:11–19.

Aina Peris and Mariona Taulé. 2011b. Annotating the argument structure of deverbal nominalizations in Spanish. doi: 10.1007/s10579-011-9172-x. *Language Resources and Evaluation*.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 296–299, Uppsala, Sweden, July. Association for Computational Linguistics.

Mariona Taulé, M.Antónia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 96–101, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Acknowledgments

This work is supported by the projects IARG-AnCora (FFI2011-13737-E), KNOW2 (TIN2009-14715-C04-04) and TextMess 2.0 (TIN2009-13391-C04-03/04) from the Spanish Ministry of Science and Innovation.

Project notes of CLARIN project DiscAn: Towards a Discourse Annotation system for Dutch language corpora

Ted Sanders
University Utrecht
Utrecht Institute of Linguistics
Trans 10
NL-3512 JK Utrecht
T.J.M.Sanders@uu.nl

Kirsten Vis
University Utrecht
Utrecht Institute of Linguistics
Trans 10
NL-3512 JK Utrecht
K.Vis@uu.nl

Daan Broeder
TLA - Max-Planck
Institute for Psycholinguistics
Wundtlaan 1
NL-6525 XD Nijmegen
Daan.Broeder@mpi.nl

Abstract

Although discourse is a crucial level in language and communication, many existing corpora of Dutch language lack annotation at this level. This paper describes the recently started DiscAn project, which sets the first step to change this situation for Dutch, in line with international tendencies. The project has five main goals: 1) to standardize and open up an existing set of Dutch corpus analyses of coherence relations and discourse connectives; 2) to develop the foundations for a discourse annotation system that can be used in Dutch natural language corpora; 3) to improve the metadata within European research infrastructure project CLARIN by investigating existing CMDI profiles or adding a new CMDI profile specially suited for this type of analysis; 4) to inventorize the required discourse categories and investigate to what extent these could be included in ISOcat categories for discourse that are currently being developed; 5) to further develop an interdisciplinary discourse community of linguists, corpus and computational linguists in The Netherlands and Belgium, in order to initiate further research on cross-linguistic comparison in a European context.

1 Introduction

Over the years, the notion of "discourse" has become increasingly important in linguistics - a remarkable development, considering that linguistics used to deal almost exclusively with sentences in isolation. Nowadays, the discipline includes the study of form and meaning of utterances in context, and formal, functional, and cognitive approaches exist that consider the discourse level as the core object of study. There seems to be a consensus that what makes a set of utterances into genuine discourse is (primarily) their meaning rather than their form. More specifically, there is a shared belief that "discoursehood" is based on the possibility to relate discourse segments to form a coherent message (Kehler, 2002; Sanders, Spooren & Noordman, 1992; Taboada & Mann, 2006; Wolf & Gibson, 2005).

Language users establish coherence by relating the different information units in the text. The notion of coherence has a prominent place in both (text-)linguistic and psycholinguistic theories of text and discourse. When confronted with a stretch of discourse, language users make a coherent representation of it. At the same time, discourse itself contains (more or less) overt signals that direct this interpretation process. In general, two types of coherence and their textual signals are distinghuished: (i) Referential coherence: how does reference to individuals create continuity and (as a result) coherence? The linguistic signals considered involve reference to persons (*Beatrix* , *she*, *the professor*), objects and concepts; (ii) Relational coherence: how do coherence relations like causals and contrastives constitute connectedness? The linguistic signals considered are connectives and lexical cue phrases. This project focuses on the second type of coherence.

Existing corpora of natural language use often lack systematic information on the discourse level. For Dutch corpora like the Corpus of Spoken Dutch ('Corpus Gesproken Nederlands', CGN), for in-

stance, lexical, syntactic and even semantic annotations are available, but typical discourse phenomena like referential and relational coherence are not addressed. Still, the discourse level is a crucial level of description for language and communication

Internationally, the last decennium has shown a tendency to change this situation. Initiatives like the Penn Discourse Treebank (Prasad et al., 2008) and the RST treebank (Carlson & Marcu, 2001) aim at creating a level of corpus annotation focusing on discourse structure information. The DiscAn project aims at developing the first step in this direction for the Dutch language community, with the explicit ambition of taking it to a cross-linguistic level. The project, that runs from April 1, 2012 until April 1, 2013, is part of and funded by CLARIN, a large-scale European research infrastructure project designed to establish an integrated and interoperable infrastructure of language resources and technologies, cf. www.clarin.nl.

2 Research data

The first aim of the DiscAn project is to integrate existing corpora of Dutch discourse phenomena in the CLARIN infrastructure, in order to standardize a valuable amount of corpus work on coherence relations and discourse connectives, and to make it available and more easily accessible for a much wider range of researchers in the humanities in general and in linguistics in particular.

The data in the existing corpora take various forms. They typically exist as fragments in doc files from scanned or copied files from newspaper, chat, spoken or child language corpora, which are analyzed on discourse variables using a systematic annotation scheme or code book. The analysis is usually available in the form of excel- or SPSS-files. Table 1 below presents a global overview of corpora, the discourse phenomena analyzed, the type of corpus, as well as the amount of analyzed cases.

3 Annotation Scheme

The various corpora have not been analyzed in identical ways, but large similarities exist with respect to the basic categories that are present in every analysis. An important part of the DiscAn project is the conceptual and text-analytical work that needs to be done, in order to identify overlapping of relevant categories, to make the analyses comparable. Earlier international work (Mann & Thompson, 1988; Sanders et al., 1992; Sanders, 1997; Sweetser, 1990; Taboada & Mann, 2006; Wolf & Gibson, 2005) will be inspiring and leading here. The Penn Discourse Treebank (Prasad et al., 2008) provides a classification, as Bunt et al. (2012) do. We expect to see similarities, but also deviations from these proposals, for both theoretical and empirical reasons. The results from our first applications to corpora will shine a light on the validity of our classification. In sum, based on existing theoretical and analytical work, the basic categories include:

- polarity: positive / negative relation (*because/omdat* and *and/en* versus *but/maar* and *although/hoewel*);
- nature: causal / temporal / additive (*because/omdat*, *then/toen*, *and/en*)
- order: antecedens-consequens or vice versa (*therefore/dus*, *because/omdat*)
- subjectivity: objective / content (*as a result/daardoor*) vs. subjective / epistemic (*therefore/dus*) vs. speech act (*so/dus*)
- perspective: subject of consciousness; first, second person, etc.
- adjacency: how far are the related segments apart?
- linguistic marking of relations: connectives / lexical cue phrase / implicit
- semantic-pragmatic characteristics of segments: modality, tense and aspect.

The discourse analytical data is available in various formats: excel tables, doc files, SPSS files etc. The data in the DiscAn project will be made available in a uniform and acceptable format, both in terms of metadata and discourse annotation categories.

Discourse phenomena	Author	Cases
Causal connectives	Bekker (2006)	500 explicit (*doordat, want, dus, daarom, nadat, voordat*) / 200 implicit
Causal connectives	Degand (2001)	150 (*want, aangezien, omdat*) from newspapers
Coherence relations	Den Ouden (2004)	70 (causal implicit, non-causal)
Connectives	Evers-Vermeul (2005)	600 historical data / 4400 from Childes
Causal connectives	Pander Maat & Degand (2001)	150 (*dus, daarom*) from newspaper corpora
Coherence relations	Pander Maat & Den Ouden (2011)	795 implicit and explicit relations from a self-assembled corpus of 40 press releases
Causal connectives	Pander Maat & Sanders (2000)	150 (*dus, daarom, daardoor*) from a newspaper-corpus (Volkskrant)
Causal connectives	Persoon (2010)	105 (*omdat, want*) from CGN
Causal connectives	Pit (2003)	200 (*aangezien, omdat, doordat, want*) newspaper / 100 (*omdat, doordat, want*) narrative; from newspaper (Volkskrant) and fictional books
Causal connectives	Sanders & Spooren (2009)	100 newspaper (Volkskrant) / 275 from CGN / 80 from Chat (*want, omdat*)
Coherence relations	Sanders & van Wijk (1996)	100 childrens explanatory texts; ca. 1500 coherence relations
Coherence relations	Spooren & Sanders (2008)	1100 coherence relations (children elicit responses)
Causal connectives	Spooren et al. (2010)	275 (*want, omdat*) spoken, from CGN; 100 (*want, omdat*) written
Causal connectives	Stukker (2005)	300 (*daardoor, daarom, dus*) newspaper / 300 historical data (*daarom, dus*)
Coherence relations	Vis (2011)	135 texts; 643 subjective relations
Connectives	Van Veen (2011)	1951 *waarom- (why-)* questions and their answers (Childes)

Table 1: Overview of DiscAn corpora.

3.1 Importance of DiscAn

The availability of this corpus, with its possibility to search on discourse terms, will be of great importance to many linguists, especially those interested in discourse structure in language use. In addition to the particularly large group of discourse analysts, text linguists and applied linguists working on text and discourse, we can think of theoretical linguists working on the syntax-semantics-discourse interface, language acquisition researchers, sociolinguists interested in language variation, as well as researchers in the field of (language and) communication. However, the merits of the DiscAn project are not limited to the availability of these corpora. The standardized annotation scheme that was used for the subcorpora will be used to further to develop the foundations for a discourse annotation system that can be used to apply in existing Dutch natural language corpora. The standardized discourse category coding scheme developed in the first phase, will be the basis for this second phase. Finally, we expect to be able to contribute to the ISOcat categories for discourse that are currently being developed. The end product of DiscAn will be a set of annotated subcorpora with discourse coherence phenomena which will allow researchers to search for

connectives and the way they are used, but also, for instance for a certain type of causal relation in spoken discourse. Researchers interested can be found in linguistics and language use (syntax, semantics, child language) and communication studies (subjectivity, variance across genres and media).

References

Birgit Bekker. 2006. *De feiten verdraaid. over tekstvolgorde, talige markering en sprekerbetrokkenheid*. Doctoral dissertation, Tilburg University, Tilburg, The Netherlands.

Harry Bunt, Rashmi Prasad and Aravind Joshi. 2012. *First steps towards an ISO standard for annotating discourse relations*. Proceedings of ISA-7 workshop (Interoperable Semantic Annotation) at LREC 2012, Istanbul.

Lynn Carlson and Daniel Marcu. 2001. *Discourse Tagging Reference Manual*. http://www.isi.edu/ marcu/discourse/

Liesbeth Degand. 2001. *Form and function of causation: A theoretical and empirical investigation of causal constructions in Dutch*. Leuven: Peeters.

Liesbeth Degand and Henk Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. In: Arie Verhagen and Jeroen van de Weijer (eds.), *Usage based approaches to Dutch*, 175-199. Utrecht: LOT.

Hanny den Ouden. 2004. *Prosodic realizations of text structure*. Doctoral dissertation, Tilburg University, Tilburg, The Netherlands.

Jacqueline Evers-Vermeul. 2005. *The development of Dutch connectives: Change and acquisition as windows on form-function relations*. Doctoral dissertation, Utrecht University, Utrecht, The Netherlands

Jacqueline Evers-Vermeul and Ted Sanders. 2009. The emergence of Dutch connectives; how cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language 36* (4), 829-854.

Andy Kehler. 2002. *Coherence, reference and the theory of grammar*. Chicago: The University of Chicago Press.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes 18*: 3562.

Alistair Knott and Ted Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics 30*: 135175.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text 8* (3), 243-281.

Henk Pander Maat and Liesbeth Degand. 2001. Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics 12*, 211245.

Henk Pander Maat and Ted Sanders. 2000. Domains of use or subjectivity: The distribution of three Dutch causal connectives explained. In: Elizabeth Couper-Kuhlen and Bernd Kortmann (eds.), *Cause, condition, concession, and contrast: Cognitive and discourse perspectives*, 5782. Berlin et al.: Mouton de Gruyter.

Ingrid Persoon, Ted Sanders, Hugo Quené and Arie Verhagen. 2010. Een coördinerende omdat-constructie in gesproken Nederlands? Tekstlinguïstische en prosodische aspecten. *Nederlandse Taalkunde*, 15, 259-282.

Mirna Pit. 2003. *How to express yourself with a causal connective? Subjectivity and causal connectives in Dutch, German and French*. Amsterdam: Editions Rodopi B.V.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. *The Penn Discourse Treebank 2.0*. In Proceedings of LREC08.

Ted Sanders. 1997. Semantic and pragmatic sources of coherence: on the categorization of coherence relations in context. *Discourse Processes, 24*, 119-147.

Ted Sanders and Wilbert Spooren. 2009. Causal categories in discourse - Converging evidence from language use. In Ted Sanders and Eve Sweetser (eds.). *Causal categories in discourse and cognition*. (pp. 205-246). Berlin: Mouton de Gruyter.

Ted Sanders, Wilbert Spooren and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes, 15*, 1-35.

Ted Sanders and Wilbert Spooren. 2008. The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics, 40* (12), 2003-2026.

Wilbert Spooren, Ted Sanders, Mike Huiskes and Liesbeth Degand. 2010. Subjectivity and Causality: A Corpus Study of Spoken Language. In: Sally Rice and John Newman (eds.) *Empirical and Experimental Methods in Cognitive/Functional Research*. (pp.241-255). Chicago: CSLI publications.

Manfred Stede. 2004. *The Potsdam commentary corpus*. Proceedings of the ACL-04 workshop on discourse annotation. Barcelona, July 2004.

Ninke Stukker. 2005. *Causality marking across levels of language structure: a cognitive semantic analysis of causal verbs and causal connectives in Dutch*. Doctoral dissertation, Utrecht University, Utrecht, The Netherlands.

Ninke Stukker, Ted Sanders and Arie Verhagen. 2008. Causality in verbs and in discourse connectives. Con-

verging evidence of cross-level parallels in Dutch linguistic categorization. *Journal of Pragmatics 40* (7), 1296-1322.

Ninke Stukker and Ted Sanders. 2011. Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics, 44* (2), 169-190.

Eve Sweetser. 1990. *From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.

Maite Taboada and William Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies, 8* (4), 567-588.

Rosie Van Veen. 2011. *The acquisition of causal connectives: the role of parental input and cognitive complexity*. PhD dissertation. Utrecht University, Utrecht, The Netherlands.

Kirsten Vis. 2011. *Subjectivity in news discourse: A corpus-linguistic analysis of informalization*. Amsterdam: VU University Amsterdam dissertation.

Florian Wolf and Edward Gibson. 2005. Representing Discourse Coherence: A corpus-based study. *Computational Linguistics 31* (2), 249-287.

Florian Wolf and Edward Gibson. 2006. *Coherence in natural language. Data structures and applications.* MIT Press, Cambridge Mass.

Studying the Distribution of Fragments of English Using Deep Semantic Annotation

Camilo Thorne
KRDB Research Centre for Knowledge and Data
3, Piazza Domenicani, 39100 (Italy)
thorne@inf.unibz.it

Abstract

We present a preliminary study on how to use deep semantic annotation, namely the Boxer statistical semantic parser, that is capable of producing FO semantic representations for English sentences, to understand the distribution of families of so-called fragments of English. In particular, we try to answer the questions, relevant for the field of natural logic, of whether and how the semantic complexity of those fragments (viz., the computational complexity of the satisfiability problem of their FO semantic representations) correlates with their frequency.

1 Introduction

Natural logic (Moss, 2010; MacCartney and Manning, 2007; van Eijck, 2005) is a relatively recent area of cognitive science, logic and computational linguistics which has as its main goal to understand which logical formalisms best model common-sense deductive reasoning as "embedded" in spoken and written language.

More recently (Muskens, 2010; Szymanik, 2009), interest has arised regarding the relationship of such formalisms to, on the one hand, *formal semantics*, the Montagovian HO and FO modelling of natural language semantics and compositionality via logic *meaning representations* (MRs) and, on the other hand, *semantic complexity*, the computational complexity of satisfiability for such MRs. This with two goals in mind: *(i)* Measuring the complexity of natural reasonings. *(ii)* Inferring correlations between complexity and frequency (viz., how often the formal models occur in natural language data) and/or accuracy (viz., what proportion of such formal reasonings are correctly inferred by speakers).

This study purports to contribute to this debate by considering the following two approaches: *(i)* Generating FO MRs from natural language text via *semantic annotation* in the form of deep (compositional and Montagovian-based) semantic parsing. *(ii)* Focusing on so-called *fragments of English*, viz., controlled subsets of English wherein ambiguity has been removed, semantics is compositional and deterministic and that give rise, modulo formal semantics, to fragments of FO (Pratt-Hartmann and Third, 2006).

By studying the semantic complexity of the fragments, via computational complexity analysis and their approximate distribution in corpora, via semantic annotation (compositional semantic parsing), we can, we believe, understand better how complexity correlates with use. For instance, complex, recursive, syntactic structures (e.g., center embedding) are less frequent in English than simpler, non-recursive structures. To see if this also holds for semantic complexity, we try to test the following hypothesis:

Semantic complexity is inversely proportional to frequency. (H)

2 Semantic Complexity and The Fragments of English

A (controlled) fragment of English (Pratt-Hartmann and Third, 2006) is a linguistically salient, ambiguity free subset of English constructed using context-free semantically enriched grammars which generate and recognize, alongside the grammatical utterances of the fragment, their logical (HO and FO) MRs, modulo Montagovian compositional translations $\tau(\cdot)$ (defined via semantic actions attached to the grammar rules). Figure 1 recalls the definition of the base fragment, whose coverage is subsequently expanded to larger subsets of English.

A *positive* fragment is any such fragment *without negation*, and a *negative* fragment is a fragment *with negation*. Each fragment of English

Fragment	Coverage	FO Operators and Relations
COP($\neg$)	Copula ("is a"), nouns ("man"), intransitive verbs ("runs"), "every", "some" names ("Joe"), adjectives ("thin") (+"not"))	$\{\forall, \exists, (\neg)\}$ $\cup$ $\{P_i^1 \mid i \in \mathbb{N}\}$
COP($\neg$)+TV	COP($\neg$) +transitive verbs ("loves")	$\{\forall, \exists, (\neg)\}$ $\cup \{P_i^1, P_j^2 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+DTV	COP($\neg$) +ditransitive verbs ("gives")	$\{\forall, \exists, (\neg)\}$ $\cup \{P_i^1, P_j^3 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+TV+DTV	COP($\neg$)+TV + ditransitive verbs	$\{\forall, \exists, (\neg)\}$ $\cup \{P_i^1, P_j^2, P_k^3 \mid i, j, k \in \mathbb{N}\}$
COP($\neg$)+Rel	COP($\neg$)+relative pronouns ("who", "that", "which") "and", intersective adjectives (+"or")	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup$ $\{P_i^1 \mid i \in \mathbb{N}\}$
COP($\neg$)+Rel+TV	COP($\neg$)+Rel +transitive verbs	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup \{P_i^1, P_j^2 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+Rel+DTV	COP($\neg$)+Rel +ditransitive verbs	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup \{P_i^1, P_j^3 \mid i, j \in \mathbb{N}\}$
COP($\neg$)+Rel+TV+DTV	COP($\neg$)+Rel+TV +ditransitive verbs	$\{\forall, \exists, \wedge, (\neg, \vee)\}$ $\cup \{P_i^1, P_j^2, P_k^3 \mid i, j, k \in \mathbb{N}\}$

Table 1: The ("positive" and "negative") fragments of English. See (Pratt-Hartmann and Third, 2006; Thorne, 2010) for more detailed definitions. Please note that we have modified slightly the former notation of the fragments, for readability reasons.

gives rise to *(i)* a distinct combination of FO operators (i.e., $\forall, \exists, \vee\neg$ and $\wedge$) *(ii)* a distinct combination of unary, binary and ternary relation symbols (together with individual constants). See Table 1. More in general, it generates a unique FO fragment, whose computational complexity for satisfiability constitutes the semantic complexity of the fragment (Pratt-Hartmann and Third, 2006), which can be studied in general, viz., *combined complexity*, or relatively to the number of constants occurring in the MRs, viz., *data complexity* (Thorne, 2010).

The fragment's content lexicon (nouns, common nouns, verbs, adjectives, names) will thus convey the signature (constants and relations) of the engendered FO fragment, while the function lexicon will convey the logical operators. Semantic complexity will be, in general, correlated to the fragment's function lexicon. Two big classes of fragments can be observed:

- "Non-Boolean-closed" fragments: are fragments that cannot express Boolean functions, viz., the positive fragments, together with COP$^\neg$, COP$^\neg$+TV, COP$^\neg$+DTV and COP$^\neg$+TV+DTV.
- "Boolean-closed" fragments: fragments expressive enough to encode Boolean satisfiability, viz., COP$^\neg$+Rel, COP$^\neg$+Rel+TV, COP$^\neg$+Rel+TV and COP$^\neg$+Rel+TV+DTV.

Table 2 summarizes the computational properties (data and combined) that arise for the fragments (for the proofs, we refer the reader to (Pratt-Hartmann and Third, 2006) and (Thorne, 2010). As the reader can see, "Boolean-closedness" gives rise, in general, to an exponential blowup in complexity. Fragments that are "non-Boolean-closed", such as the positive fragments and the fragments without relatives and transitive verbs, have *tractable* (at most PTIME) combined or data complexity, whereas "Boolean-closed" fragments have *intractable* (at least NPTIME-hard) combined or data complexity.

3 Empirical Analysis

3.1 Corpus Analysis

In this section we summarize our analysis regarding the co-occurence of negations, conjunctions, disjunctions, and universal and existential quantification in English question and sentence corpora via semantic annotation.

More precisely, we consider the frequency of sentences expressing, modulo formal semantics, *positive* (not containing $\neg$) and *negative* (contain-

Phrase Structure Rules	
$\mathbf{S} \rightarrow \mathbf{NP\ VP}$	$\tau(\mathbf{S}) = \tau(\mathbf{NP})(\tau(\mathbf{VP}))$
$\mathbf{VP} \rightarrow$ is a $\mathbf{N}$	$\tau(\mathbf{VP}) = \tau(\mathbf{N})$
$\mathbf{VP} \rightarrow$ is $\mathbf{Adj}$	$\tau(\mathbf{VP}) = \tau(\mathbf{Adj})$
$\mathbf{VP} \rightarrow \mathbf{IV}$	$\tau(\mathbf{VP}) = \tau(\mathbf{IV})$
$\mathbf{NP} \rightarrow \mathbf{Pn}$	$\tau(\mathbf{NP}) = \tau(\mathbf{Pn})$
$\mathbf{NP} \rightarrow \mathbf{Det\ N}$	$\tau(\mathbf{NP}) = \tau(\mathbf{Det})(\tau(\mathbf{N}))$
$(\mathbf{VP} \rightarrow$ is $\mathbf{Ng}$ a $\mathbf{N}$	$\tau(\mathbf{VP}) = \tau(\mathbf{Ng})(\tau(\mathbf{N})))$
$(\mathbf{VP} \rightarrow$ does $\mathbf{Ng\ IV}$	$\tau(\mathbf{VP}) = \tau(\mathbf{Ng})(\tau(\mathbf{IV})))$

Function Lexicon	
$\mathbf{Det} \rightarrow$ every	$\tau(\mathbf{Det}) = \lambda P.Q.\forall x(P(x) \rightarrow Q(x))$
$\mathbf{Det} \rightarrow$ some	$\tau(\mathbf{Det}) = \lambda P.Q.\exists x(P(x) \wedge Q(x))$
$(\mathbf{Ng} \rightarrow$ not	$\tau(\mathbf{Ng}) = \lambda P.\lambda x.\neg P(x))$

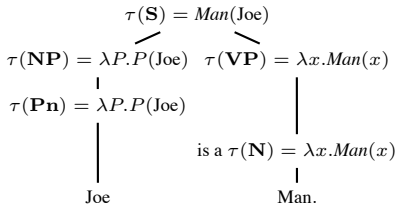

Figure 1: Top: COP(¬). Bottom: COP(¬) parse tree for "Joe is a man.". We omit the content lexicon. Notice how to each grammar rule a semantic action is attached, defining $\tau(\cdot)$.

ing ¬) classes $c \subseteq \{\forall, \exists, \neg, \wedge, \vee\}$ of FO operators. Each such class approximates MRs belonging, modulo logical equivalence, to a distinct fragment of FO and expressible by a distinct fragment of English. For instance the class $\{\forall, \exists, \wedge, \vee\}$ identifies MRs from the positive fragment of FO. But it also identifies MRs belonging to English fragments such as, e.g., COP(+Rel)+TV+DTV. Specifically, after semantically annotating the corpora we observed the frequency of

- 4 "Boolean-closed" classes viz.: $\{\exists, \wedge, \neg\}$, $\{\exists, \wedge, \neg, \forall\}$, $\{\exists, \wedge, \neg, \forall, \vee\}$ and $\{\neg, \forall\}$, and of
- 4 "non-Boolean-closed" classes viz.: $\{\exists, \wedge\}$, $\{\exists, \wedge, \forall\}$, $\{\exists, \wedge, \vee\}$ and $\{\exists, \wedge, \forall, \vee\}$,

where by "Boolean-closed" and "non-Boolean-closed", we mean, by abuse, classes, resp., expressive or not expressive enough to encode Boolean satisfiability.

To obtain a representative sample, we considered corpora of multiple domains and with sentences of arbitrary type (declarative and interrogative). We considered: *(i)* a subset (A: press articles) of the Brown corpus[1]; *(ii)* a subset (Geoquery880) of the Geoquery corpus[2]; *(iii)* a corpus of clinical questions[3]; and *(iv)* a sample from the TREC 2008 corpus[4]. Table 3 summarizes their main features.

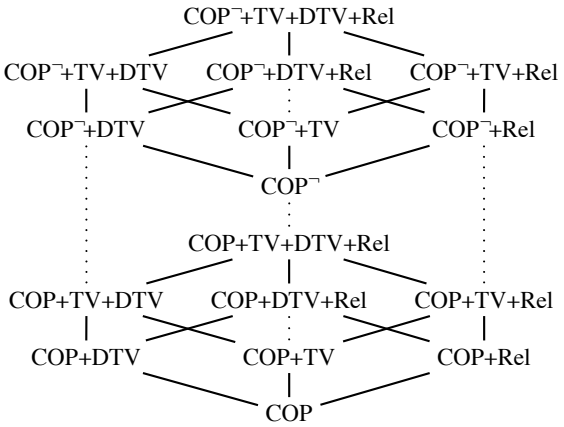

Figure 2: Relative expressive power of the fragments.

We used two methods, that we describe below. The left column of Figure 3 provides plots of the statistics collected with both methods. They also plot the mean class frequencies across the corpora, and the mean cumulative frequency.

Semantic Parsing with Boxer. We exploited the availability of wide-coverage (statistical) deep semantic parsers and annotators. In particular, the Boxer and Clark & Curran tools (Bos, 2008), based on combinatorial categorial grammar and discourse representation theory (DRT), that output first-order MRs. The pipeline of this system consists in the following three basic steps: *(i)* each part of speech in a sentence is annotated with its most likely (categorial grammar) syntactic category; *(ii)* the most likely of the resulting possible combinatorial categorial grammar derivations (or proofs) is computed and returned; and *(iii)* a neo-Davidsonian semantically weakened[5] FO meaning representation is computed using DRT.

For instance, when parsing Wh-questions from

[1] `http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml`

[2] `http://www.cs.utexas.edu/users/ml/nldata/geoquery.html`

[3] `http://clinques.nlm.nih.gov`

[4] `http://trec.nist.gov`

[5] In this setting, the semantics of verbs is represented in terms of events connected via thematic roles to verb arguments (agents, themes, etc.). In addition, the semantics of non-FO constructs such as "most" is weakened to some FO representation.

	Combined	Data
COP	LSPACE	LSPACE
COP+TV	PTIME	LSPACE
COP+DTV	PTIME	LSPACE
COP+TV+DTV	PTIME	LSPACE
COP+Rel	PTIME-c	LSPACE
COP+Rel+TV	PTIME-c	PTIME
COP+Rel+DTV	PTIME-c	PTIME
COP+Rel+DTV+TV	PTIME-c	PTIME

	Combined	Data
$\text{COP}^{\neg}$	LSPACE	LSPACE
$\text{COP}^{\neg}$+TV	NLSPACE-c	LSPACE
$\text{COP}^{\neg}$+DTV	PTIME	LSPACE
$\text{COP}^{\neg}$+TV+DTV	PTIME	LSPACE
$\text{COP}^{\neg}$+Rel	NPTIME-c	LSPACE
$\text{COP}^{\neg}$+Rel+TV	EXPTIME-c	NPTIME-c
$\text{COP}^{\neg}$+Rel+DTV	NEXPTIME-c	NPTIME-c
$\text{COP}^{\neg}$+Rel+DTV+TV	NEXPTIME-c	NPTIME-c

Table 2: Semantic complexity of the fragments of English, positive and otherwise (Pratt-Hartmann and Third, 2006; Thorne, 2010).

Corpus	Size	Domain	Type
Brown	19,741 sent.	Open (news)	Decl.
Geoquery	364 ques.	Geographical	Int.
Clinical ques.	12,189 ques.	Clinical	Int.
TREC 2008	436 ques.	Open	Int.

Table 3: Corpora used in this study.

the TREC 2008 corpus such as "What is one common element of major religions?", Boxer outputs

$$\exists y \exists z \exists e \exists u(\texttt{card}(y,u) \wedge \texttt{c1num}(u) \\ \wedge \texttt{nnumeral1}(u) \wedge \texttt{acommon1}(y) \\ \wedge \texttt{nelement1}(y) \wedge \texttt{amajor1}(z) \\ \wedge \texttt{nreligions1}(z) \wedge \texttt{nevent1}(e) \\ \wedge \texttt{rof1}(y,z))$$

where $\wedge$ and $\exists$ co-occur, but not $\vee$, $\neg$, or $\rightarrow$.

After semantically annotating each corpus with Boxer, we checked for each MR produced, whether it belongs to a "Boolean-closed" or a "non-Boolean-closed" class.

Pattern-based. Boxer is considered to have a reasonably good performance (covering over 95% of English, with approx. 75% accuracy), when parsing and annotating declarative sentences and corpora, but not necessarily so over interrogative sentences and corpora. It also commits us to a neo-Davidsonian semantics, whose event-based verb semantics biases interpretation towards positive existential MRs, making its output somewhat noisy.

To understand how useful Boxer (or similar deep semantic annotators) can be to discover statistical trends of the kind stated in our hypothesis (H), we decided compare to its results to those that one may obtain using a simple methodology based on patterns. Indeed, modulo formal semantics, English function words convey or express FO operators. As such we considered the following patterns, *(i)* for $\neg$: "not", "no" *(ii)* for $\exists$: "some", "a" *(iii)* for $\forall$: "all", "every, "each" *(iv)* for $\wedge$: "who", "what", "which", "and" *(v)* for $\vee$: "or", and their combinations/co-occurences *within sentences* to approximate the"Boolean-" and "non-Boolean-closed" classes that interest us.

3.2 Basic statistical tests

The mean (cumulative) frequency plots obtained in Figure 3 show a distribution where class frequency is skewed towards positive existential classes: $\{\exists, \wedge\}$, $\{\exists, \forall, \wedge\}$ and positive existential $\{\exists, \forall, \wedge, \vee\}$ MRs occur quite frequently, whereas the opposite holds for negation (low frequency overall). The question is whether this substantiates our hypothesis (H). We ran some basic statistical tests to understand how random or significant this phenomenon is. Table 4 summarizes the test results, which we explain below.

Power Law Behavior. A *power law distribution* is a kind of exponential, non-normal and skewed distribution where the topmost (i.e., most frequent) 20% outcomes of a variable concentrate 80% of the probability mass.

Power law distributions are widespread in natural language data (Baroni, 2009; Newman, 2005). It makes sense to understand whether the relationship stated by (H) can be restated and described as a power law relation between *class frequency* $fr(c)$ and *class rank* $rk(c)$, viz.,

$$fr(c) = \frac{a}{rk(c)^b}. \quad (1)$$

To approximate the parameters a and b it is customary to run a least squares linear regression, since (1) is equivalent to a linear model on the log-log scale:

$$\log_{10}(fr(c)) = \log_{10}(a) - b \cdot \log_{10}(rk(c)). \quad (2)$$

	$H(C)$	$H_{rel}(C)$	Skewness	χ^2	p-value	df.	Power law	R^2
Boxer	1.53	0.51	1.93	293731473.0	0.0	7	$fr(c) = \frac{47.86}{rk(c)^{1.94}}$	0.92
Patterns	1.50	0.50	1.21	906727332.0	0.0	7	$fr(c) = \frac{5128.61}{rk(c)^{4.09}}$	0.73

Table 4: Summary of test results.

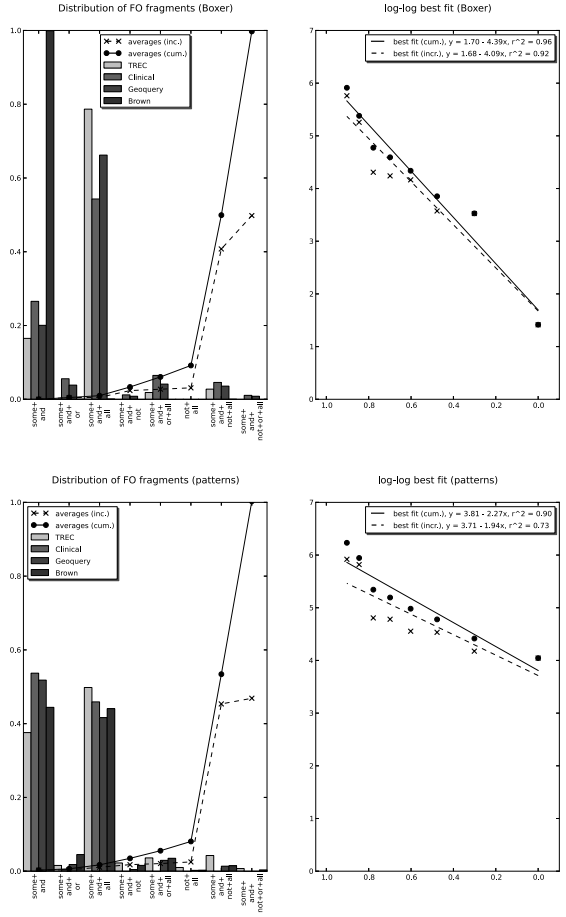

Figure 3: Fragment distribution with Boxer and the pattern-based method, together with their log-log regression plot. We plot class cumulative and mean frequencies (the latter in increasing order).

Entropy, Skewness and χ^2 Tests. Following mainly (Gries, 2010), we conducted the following tests. We computed class *entropy* $H(C)$, where C denotes $\{\forall, \exists, \vee, \wedge, \neg\}$. This number tries to measure the degree of randomness of a distribution:

$$H(C) = -\sum_{c \in C} fr(c) \cdot \log_2(fr(c)). \qquad (3)$$

A low number indicates a low degree of randomness. Entropy can be complemented with its *relative entropy* $H_{rel}(C)$:

$$H_{rel}(C) = \frac{H(C)}{\log_2(\#(C))}. \qquad (4)$$

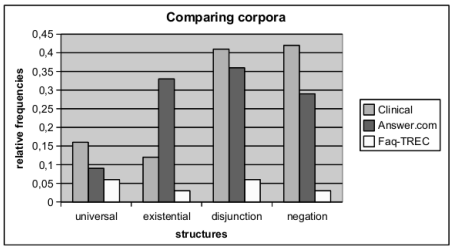

Figure 4: Relative frequency of FO operators in question corpora (Bernardi et al., 2007).

In addition to measuring $H(C)$ and $H_{rel}(C)$, we also run a χ^2 test (since distributions in natural language data are often non-parametric) and measured the overall skewness of the distribution.

3.3 Discussion and Related Work

Table 4 shows that the distributions observed under both methods possess a relatively low entropy (relatively to a peak entropy of 3.0), and thus appear not to be so random. The χ^2 statistic, moreover, entails that such distributions differ from uniform or random distributions (the null hypothesis rejected by the test), since $p < 0.01$. They also show a high measure of skewness. Lest, but not least, the cumulative and non-cumulative distributions seem to follow, to some extent a power-law model. The reader will find on the second (right) column of Figure 3 the plots of the log-log regressions, which show a strong positive correlation (the R^2 index), stronger for Boxer (0.92) than for the patterns (0.73)[6].

This analysis can be compared to the more linguistics-based methodology followed in (Bernardi et al., 2007), in which was analyzed the distribution, in (solely) interrogative corpora, of classes of logical word patterns (but not of their co-occurence), e.g., "all", "both", "each", "every", "everybody", "everyone", "any", "none", "nothing". See Figure 4.

This may suggest that, while users use negation or disjunction words as frequently as conjunction

[6]Quite strong for both cumulative distributions: 0.96 and 0.90 resp., in the plot.

and existential words, and all these more than universal words, when combining them *within* sentences, "non-Boolean-closed" combinations are preferred.

Moreover (Szymanik, 2009) reports results that may seem to imply that natural reasoning *accuracy* (and not only their distribution in corpora) may be inversely correlated to expressiveness and semantic complexity, where by accuracy is meant the ability of speakers to correctly infer logical consequences from texts describing logical arguments (in the experiments, arguments regarding FO and HO generalized quantifiers). Users make more mistakes when the underlying logic (or logical MRs) are NPTIME-hard than when they are PTIME, and take more time to understand and infer such consequences.

This said, the corpora considered in our study were small, and the two methods (Boxer and the patterns) inaccurate (indeed, the pattern-based method remains quite simple). The results reported here, while encouraging, cannot be regarded yet as fully conclusive for (H).

4 Conclusions

We have presented a preliminary study on how to apply deep semantic annotation techniques to understand the distribution of specific fragments of English in English corpora, and specifically to understand if is is possible to infer relationships between their distribution and their semantic complexity (i.e., the computational complexity of their logic MRs).

We have experimented with a methodology based on the Boxer semantic parser, and applied some basic statistical tests on the distribution obtained that may seem to indicate that "non-Boolean-closed" (tractable) fragments might occur more often than "Boolean-closed" (intractable) fragments, although the results obtained thus far remain still inconclusive.

To counter these shortcomings we would like in the future to *(i)* run the experiment with other deep semantic annotation methods and parsers (such as, e.g., those based on minimal recursion semantics (Copestake, 2007)), *(ii)* consider larger corpora, in particular declarative corpora (over which the performance of Boxer is higher) *(iii)* consider more involved statistical tests, to try to understand how the fragments are distributed. We believe however that the methodology proposed is interesting and promising, and all the more due to the current advances in semantic annotation, which may yield better results once points *(i)–(iii)* are addressed.

References

Marco Baroni. 2009. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics: An International Handbook*, volume 2, pages 803–821.

Raffaella Bernardi, Francesca Bonin, Domenico Carbotta, Diego Calvanese, and Camilo Thorne. 2007. English querying over ontologies: E-QuOnto. In *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence (AI*IA 2007)*.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing (STEP 2008)*.

Ann Copestake. 2007. Semantic composition with (robust) minimal recursion semantics. In *Proceedings of the ACL-07 workshop on Deep Linguistic Processing*.

Stefan Th. Gries. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez and Moisés Almela, editors, *A mosaic of corpus linguistics: selected approaches*, pages 269–291. Peter Lang.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing (RTE 2007)*.

Lawrence S. Moss. 2010. Natural logic and semantics. In *Proceedings of the 2009 Amsterdam Colloquium (AC 2009)*.

Reinhard Muskens. 2010. An analytic tableau system for natural logic. In *Proceedings of the 2009 Amsterdam Colloquium (AC 2009)*.

M. E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Ian Pratt-Hartmann and Allan Third. 2006. More fragments of language. *Notre Dame Journal of Formal Logic*, 47(2):151–177.

Jakub Szymanik. 2009. *Quantifiers in Time and Space*. Institute for Logic, Language and Computation.

Camilo Thorne. 2010. *Query Answering over Ontologies Using Controlled Natural Languages*. KRDB Centre for Knowldge and Data.

Jan van Eijck. 2005. Natural logic for natural language. In *Proceedings of the 6th International Tbilisi Symposium on Logic, Language, and Computation TbiLLC 2005*.

Semantic Annotation of Metaphorical Verbs with VerbNet: A Case Study of 'Climb' and 'Poison'

Susan Windisch Brown
University of Florence
Piazza Savonarola 1
50132 Firenze, Italy
susanwbrown@att.net

Martha Palmer
University of Colorado
295 UCB
Boulder, CO 80309, U.S.A.
martha.palmer@colorado.edu

Abstract

Metaphor is commonplace in language, regardless of genre, register or tone. As natural language processing moves beyond surface-level analyses into deeper semantic analysis, accurate identification and representation of metaphoric meaning becomes more important. In this paper, we look at several issues that arise when metaphorical language is semantically annotated, including the identification of appropriate thematic role labels and semantic representations. We look at the applicability of VerbNet's classification for verbs that are commonly used both literally and metaphorically, using the verbs *climb* and *poison* as illustrations. We found that the full complexity of metaphor and other figurative language is not captured by annotation with VerbNet at this time, but that the current VerbNet structure provides accurate labelling of general thematic roles for metaphors and often accurate semantic representations as well.

1 Introduction

Metaphor is commonplace in language, regardless of genre, register or tone. As natural language processing moves beyond surface-level analyses into deeper semantic analysis, accurate identification and representation of metaphoric meaning becomes more important. In this paper, we look at several issues that arise when metaphorical language is semantically annotated, including the identification of appropriate thematic role labels and semantic representations. We look at the applicability of VerbNet's classification for verbs that are commonly used both literally and metaphorically, using the verbs *climb* and *poison* as illustrations.

VerbNet, inspired by Beth Levin's (1993) classification of English verbs, is a verb lexicon that groups verbs into classes based on similarities in their syntactic and semantic behavior (Schuler, 2005). It was not created with a division between literal and metaphoric uses of language as an organizing principle. Therefore, this feature of language manifests itself in several different ways: (1) separate class assignments for literal and metaphoric uses of a verb; (2) one class encompassing both literal and metaphoric uses; or (3) only literal uses specifically accounted for in a class, with metaphorical uses unattributable to any class. These different outcomes result from various factors, such as how conventionalized a metaphoric usage is and the kinds of semantic restrictions that are specified by VerbNet for the arguments of a class.

The choice to focus on the verbs *climb* and *poison* stems from their representativeness in two different areas. First, both verbs are members of broad VerbNet classes that include many other verbs that are also used literally and metaphorically. We look more closely at this representativeness at the beginnings of section 2 and section 3. Second, taken together, these two verbs give examples of the various ways VerbNet deals with metaphor, as described above.

In discussing metaphor, we will refer to Lakoff and Johnson's (1980) definition, in which

one conceptual domain (the source) maps to another domain (the target). Generally the source domain is a concrete, everyday domain that is used to elucidate a more abstract domain, the target.

2 *Climb*

This verb is currently a member of four VerbNet classes: *calibratable_cos-45.6; escape-51.1; run-51.3.2* and *meander-47.7*. These classes contain 257 verbs, many of which are used metaphorically in similar ways to *climb*. In fact, many of these verbs join *climb* in multiple classes, largely because of these metaphorical patterns. For example, *plunge* and *rise* are both members of the *calibratable_cos-45.6; escape-51.1;* and *meander-47.7* classes, and many other verbs belong to some combination of two or three of the classes that include *climb*.

These four classes belong to three broad categories of classes in VerbNet: Verbs of Change of State, Verbs of Existence, and Verbs of Motion. These encompass 24 VerbNet classes with hundreds of verbs. While *climb* seems representative of many verbs in the four classes, further study is needed to determine if its metaphorical patterns are representative of these broad categories of verbs.

The *calibratable_cos-45.6* class includes verbs that describe an entity's change along a scale, such as *increase, decrease,* and *multiply*. Most verbs in this class, however, fit it only when they are used metaphorically, such as *rise, fall, plunge,* and *climb*. The same is true of the *meander-47.7* class. The verbs in this class are primarily motion verbs that are being used metaphorically to describe the spatial configuration of something, such as *"The path climbs through the woods"*.

Conversely, *climb's* other two classes, *escape-51.1* and *run-51.3.*2, seem to accommodate both literal and metaphoric uses of their member verbs, at least when considering the alternations, general thematic roles and the semantic representations of the class. However, the semantic restrictions on the classes' thematic roles often result in excluding metaphoric extensions of the verb, as we show in sections 2.1 and 2.2.

2.1 *Escape-51.1*

Escape-51.1 includes verbs of motion along a path, such as *come, go, return,* and *climb*. The syntactic frames that characterize this class include

- NP V PP.initial_loc (*He came from France)*
- NP V PP.Destination *(He came to Chicago)*
- NP V PP.Trajectory *(He came through the door)*

One literal sense of *climb* fits all these alternations (e.g., *"He climbed from the bottom of the hill to the top").* A corresponding metaphorical sense also fits all these alternations *(e.g., "He climbed from the gutter to the board room").*

In addition, a member of the *escape-51.1* class should work with the following thematic roles[1]:

- **THEME [+CONCRETE]**
- **INITIAL_LOCATION [+CONCRETE]**
- **DESTINATION [+CONCRETE]**
- **TRAJECTORY [+CONCRETE]**

Figurative sentences using *climb* can fit the thematic roles and the syntactic patterns of this class without satisfying the semantic restrictions, such as "*John* [Theme [+concrete] *climbed from poverty* [Initial_Location [-concrete] *to wealth* [Destination [-concrete]". *"Her feelings for him climbed from indifference to genuine love"* in which even the Theme is not concrete. For an application interested only in annotations of basic thematic roles, annotating these instances of *climb* as belonging to this class would not be a problem.

The issue of applying the semantic representation is a bit more complicated.

MOTION(DURING(E), THEME**)**
PATH(DURING(E), THEME, INITIAL_LOCATION, ?TRAJECTORY, DESTINATION**)**

The Theme is not actually in motion and does not change locations, but this is rather a metaphorical reference to changing circumstances. Without an indication that the instance is metaphoric, incorrect inferences would be drawn from the application of

[1] These roles and predicates represent a new treatment of motion in VerbNet which is still being developed, (Hwang et al., 2012)

this semantic representation. With an indication that the instance is metaphoric and a key to map the predicates from the source domain to the target domain, the semantic representation for this sentence could be appropriate. Annotation using this class for metaphoric as well as literal language would require an additional layer identifying metaphors and referring to mappings from the source domain to the target domain.

Although most metaphoric sentences would be excluded by strictly followed semantic restrictions, some will not be. *"John climbed from the slums of the South Side to the trading floor of the Chicago Stock Exchange"*, where the arguments, at least on the surface, satisfy the semantic restrictions. The semantic representation would not be incorrect: John probably was at one time at the location of the slums and moved to the literal location of the Chicago Stock Exchange. However, the representation misses the important implication that John's circumstances in life have changed dramatically, from poverty to wealth. And his route was much more complex than traversing the physical distance between the two locations. A literal interpretation would lead to incorrect assumptions.

2.2 *Run-51.3.2*

The run class has a similar literal focus and requires agents and themes that are +animate and locations that are +concrete. The semantic representation for a sentence like "John climbed the hill" is

MOTION(DURING(E), THEME**)**
VIA(DURING(E), THEME, LOCATION**)**[2]

Figurative sentences like *"John is climbing the social ladder"* would fit this class's syntactic alternations and would receive a semantic representation with similar accuracies and inaccuracies to the figurative sentences in the *escape-51.1* class.

2.3 *Calibratable_cos-45.6*

Certain figurative uses of climb would be annotated with the *calibratable_change_of_state-45.6* class, for example, *"The stock's price climbed $15 in one day of trading."* This sense of *climb* is expressed in all the alternations of this class, which also captures the intended meaning of this conventionalized metaphor for *climb*. The roles of this class, Patient, Attribute and Extent, fit well with this usage, with the *stock* as the Patient that undergoes a change, the *price* as its Attribute and the *$15* change as the Extent. The semantic representation fits as well, with no need to map from a source domain to a target domain:

CHANGE_VALUE(DURING(E), DIRECTION, ATTRIBUTE, PATIENT**)**
AMOUNT_CHANGED(DURING(E), ATTRIBUTE, PATIENT, EXTENT**)**

2.4 *Meander-47.7*

The final class that includes *climb* as a member is the *meander-47.7* class, which describes fictive motion verbs. Certain motion verbs can be used metaphorically to describe the stative configuration of a path (Ramscar, Boroditsky & Matlock 2010; Talmy 2000). A typical fictive motion use of *climb* would be: *"The trail climbs through the trees."* The *meander-47.7* class uses the roles Theme [+elongated] and Location [+concrete]. These roles, excepting the semantic restriction on the Theme, are the same as those in the motion-oriented classes *run-51.3.2* and *escape-51.1*. The semantic representation is vastly different, however, and accurately describes the metaphoric meaning intended:

PREP(DURING(E), THEME, LOCATION**)**
EXIST(DURING(E), THEME**)**

Rather than an event describing the change of location of a Theme, the semantic representation describes a state (i.e., EXIST) of a Theme being in a particular Location.

3 *Poison*

The verb *poison* is a member of two VerbNet classes*: butter-9.9* and *poison-42.2*. The *butter-9.9* class comprises 140 verbs that express putting a Theme in a Destination, such as *cloak, glaze, salt,* and *poison.* It belongs to a larger category of 10

[2] As part of ongoing revisions to VN, the VIA predicate here may change to TRAJECTORY.

classes called Verbs of Putting. The *poison-42.2* class comprises 22 verbs, such as *shoot, stab* and *poison,* and belongs to the larger category of classes, Verbs of Killing. Although these classes include fewer verbs than the Verbs of Putting, they are more frequently used in conventional metaphors.

3.1 *Butter-9.9*

The *butter-9.9* class has several selectional restrictions on its thematic roles:

- **AGENT [+ANIMATE]**
- **THEME [+CONCRETE]**
- **DESTINATION [+LOCATION & -REGION]**

The semantic representation for a sentence like "Lora poisoned the stew" seems likewise concrete:

MOTION(DURING(E), THEME**)**
NOT(LOCATION(START(E), THEME, DESTINATION**))**
LOCATION(END(E), THEME, DESTINATION**)**
CAUSE(AGENT, E**)**

Poison is not usually used figuratively with these roles, except with the phrase "poison the well", as in *"By the time I joined the board, John had already poisoned the well and no one would even listen to my plans."* As explained in the next section, the sentence, *"He poisoned her mind with lies*" fits better with the *poison-42.2* class, where "her mind" would be interpreted as the Patient that gets harmed by the poison rather than the destination of the poison.

3.2 *Poison-42.2*

The *poison-42.2* class accommodates both physical, concrete events of poisoning and at least some figurative events of poisoning. The class is characterized by only four syntactic frames:

- NP V NP *(The witch poisoned Mary);*
- NP V NP Adj *(The witch poisoned Mary dead);*
- NP V NP PP.Result *(The witch poisoned Mary to death);*
- NP V NP PP.Instrument *(The witch poisoned Mary with arsenic).*

The metaphoric uses of the verb *poison* also follow these frames, as we show below.

The thematic roles in the *poison-42.2* class are

- **AGENT [+ANIMATE]**
- **PATIENT [+ANIMATE]**
- **INSTRUMENT**
- **RESULT**

The semantic predicate for a sentence like *"The queen poisoned Snow White with the apple"* is

CAUSE(AGENT, E**)**
HARMED(DURING(E), PATIENT**)**

Clearly, physical events of poisoning fit perfectly well with the semantics of this class. Figurative poisoning seems to work as well. For example, the sentence *"John poisoned Mary with his lies"* has an animate Agent and an animate Patient and, because there are no selectional restrictions on the Instrument role, "lies" fits with that role. The semantic representation does not specify what kind of harm the patient undergoes, physical or otherwise, so it seems equally appropriate for this metaphorical sentence.

Although the class accommodates some metaphorical usages, it is not wholly free in its applicability. Sentences like *"John poisoned the process with his double-dealing"* and *"Max poisoned his Mercedes with low-grade gasoline"* would violate the restriction that the Patient should be animate. The consequences of ignoring this selectional restriction do not seem grave, as the semantic representation still seems perfectly adequate in describing these events.

4 Semlink annotation of metaphor

The SemLink project (Palmer, 2009) has implemented semantic annotation with VerbNet classes, labeling a portion of the Wall Street Journal corpus with VerbNet class labels on the verbs and thematic role labels from those classes on the arguments. A lenient approach to class assignments was used, often applying thematic role criteria without their semantic restrictions when determining a token's class assignment (CLEAR, 2012). This approach resulted in many metaphoric verb tokens being annotated with classes that, under stricter criteria, would only apply to literal verb usages. These tokens would have otherwise

remained unannotated, as no class represented the purely metaphorical interpretation.

Annotation for the verb *climb* provides a good example of the variety of ways metaphoric tokens were annotated. Tokens of the *type "Share prices of many of these funds have climbed much more sharply than the foreign stocks they hold* [wsj_0034.mrg 2613]," where *climb* is used metaphorically to map from the source domain of motion and change of location to the target domain of change in value, were annotated with the *calibratable_cos* class. For these tokens, the thematic roles and semantic representation suit the target domain (the metaphoric meaning). Several other metaphoric tokens of the verb *climb* were assigned to the *escape-51.1* class, including *"Japan has climbed up from the ashes of World War II* [wsj_1120.mrg 0 2]" and *"It has always been the case that those outside the club want to climb in* [wsj_1986 49 14]." In these cases, the thematic roles and semantic representation follow the source domain (the literal meaning).

5 Conclusion

Although the full complexity of metaphor and other figurative language is not captured by annotation with VerbNet at this time, the current VerbNet structure provides a more accurate labelling of semantic features than one would first suppose from a resource not designed with literal-figurative distinctions in mind. Often conventionalized metaphoric uses of a verb are separated from literal uses and placed in classes where the metaphoric meaning is the primary one, such as the *calibratable_change_of_state* class. In those cases, the semantic representation captures the actual metaphoric meaning, rather than that of the source of the metaphor. VerbNet has many such classes, such as the "psychological" classes, where figurative uses of verbs like *cut*, *wound*, and *shake* are annotated with Experiencer and Stimulus roles and the semantic representation indicates a change in the Experiencer's emotional state.

Where metaphoric uses have no appropriate separate class, VerbNet affords a very accurate shallow semantic annotation in the form of thematic roles. These are applicable to both literal and figurative uses of verb members of a class, especially when selection restrictions are disregarded. The semantic representation is sometimes equally applicable, such as with the *poison-42.2* class. More often, though, the semantic representation would need some sort of indication that it is to be interpreted metaphorically to avoid inaccurate inferences from being drawn, such with the *run* and *escape* classes.

References

CLEAR (Computational Language and Education Research). 2012. VerbNet Annotation Guidelines. University of Colorado, Boulder.

Jena D. Hwang, Martha Palmer, and Annie Zaenen, (2012), From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen, Ed: Tracy Holloway King and Valeria de Paiva, CSLI On-LinePublications, to appear, 2012.

Beth Levin. 1993. English Verb Classes and Alternations. University of Chicago Press, Chicago, IL.

George Lakoff and Mark Johnson. 1980. Metaphors We Live By. University of Chicago Press, Chicago, IL.

Martha Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. Proceeding of the Generative Lexicon Conference. Pisa, Italy.

M. Ramscar, L. Boroditsky, and T. Matlock. 2010. Time, motion and meaning: The experiential basis of abstract thought. In K.S.,Mix, L.B. Smith, and M. Gasser (eds.), The Spatial Foundations of Language and Cognition. Oxford University Press, Oxford, U.K.

Karin Kipper Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA.

Leonard Talmy. 2000. Toward a Cognitive Semantics. Vol.1. MIT Press, Cambridge, Mass.

Empirical validations of multilingual annotation schemes for discourse relations

Sandrine Zufferey
U.C. Louvain
Institut langage et communication
Place B. Pascal, 1
B-1348 Louvain
sandrine.zufferey@uclouvain.be

Liesbeth Degand
U.C. Louvain
Institut langage et communication
Place B. Pascal, 1
B-1348 Louvain
liesbeth.degand@uclouvain.be

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny
andrei.popescu-belis@idiap.ch

Ted Sanders
Universiteit Utrecht
Utrecht Institute of Linguistics
Trans 10
NL-3512 JK Utrecht
T.J.M.Sanders@uu.nl

Abstract

This paper discusses the empirical validation of annotation schemes proposed for discourse relations, when signaled explicitly by discourse connectives, through their application to texts in several languages. Considering a monolingual annotation scheme as a starting point, the paper explains the reasons for either specifying or generalizing some labels, illustrating them with a review of experiments in translation spotting of connectives. Then, an experiment with the PDTB scheme applied to five languages (EN, FR, DE, NL, and IT) shows how specification and generalization are put to work in order to build a scheme which has an improved empirical validity for several languages.

1 Introduction

Several corpora with annotated discourse relations have become available in the past years, inspired by the first lexicalized discourse structure annotation performed for the Penn Discourse Treebank (PDTB, Prasad, Dinesh, Lee et al., 2008), which has become a landmark in the field – see Webber and Joshi (2012) for a review. These annotation efforts have reused and sometimes redefined the annotation instructions and the classification of discourse relations proposed by the PDTB. This taxonomy holds for discourse relations that can be lexicalized through the use of discourse connectives, but also for implicit relations that are not lexicalized.

In this paper, we focus on lexicalized discourse relations, made explicit by discourse connectives, in a parallel corpus with translations from English into four other languages. Through a series of experiments with the PDTB taxonomy of discourse relations, we show how this taxonomy should be adapted to suit the needs of several languages and to make the annotation process more accurate (Sections 4 to 6). However, we initially reflect from a more general perspective on the benefits of multilingual annotation for designing a standardized taxonomy of discourse relations applicable across languages. After stating the problem theoretically (Section 2), we review monolingual and multilingual annotations, including translation spotting of discourse connectives in parallel corpora (Section 3).

2 Impact of multilingual annotations on taxonomies of discourse relations

The attempt to define a universally acceptable list of discourse relations (Bunt, Prasad and Joshi, 2012) raises several theoretical questions about the principles governing such a list. In our view, some of the most important ones are:

- What counts as a discourse relation and what theory should be used to list possible relations?
- Are discourse relations truly language-independent, i.e. can all of them be encountered in texts from any language?
- Are all discourse relations equally achievable by implicit and explicit means? In particular, are there, in a given language, connectives to express all relation types?
- What is the relation between a language-independent taxonomy of discourse relations and the range of discourse connectives available in a given language? How can such a taxonomy be used to map discourse connectives from one language to another?
- Do all discourse relations that can be expressed by a given connective count as possible meanings of that connective?
- Given that one connective is almost never fully substitutable with another one, are there more meanings than connectives? And what accounts for the diversity of connectives in European languages?

These questions are, of course, far beyond the scope of this paper. In this section, we will first state two principles that govern the relations between a taxonomy of discourse relations and the vocabularies of discourse connectives in several European languages. We will also briefly discuss the relation between semantic meaning and meaning in context for discourse connectives.

2.1 Specification vs. generalization in a taxonomy of discourse senses

Let us consider first an existing taxonomy such as the PDTB, used for the annotation of a large English corpus, and let us suppose a translation of the annotated corpus is available in French. Then, when examining all occurrences of an English discourse connective C_i annotated with a sense R_n from the taxonomy, it might happen that several different translations of C_i are observed (with significant frequencies), and that these different translations correspond to a previously uncategorized distinction of the discourse relation R_n. Hence, in this case, R_n must be subdivided into two more specific relations, say R_{n1} and R_{n2}. We call this the *specification process* (or *refinement*) of the taxonomy.

Consider now a different case: after application to annotation over large corpora in several languages, it is found that two senses of a taxonomy, say R_{p1} and R_{p2} exhibit low inter-annotator agreement, and are often dispreferred in favor of their supersense (in the taxonomy), say R_p. In this case, it makes sense to prune the two senses from the taxonomy and keep only their supersense. Of course, this does not rule out the possibility that when a new language is annotated, the supersense must be again specified. However, until such additional evidence is found, a more compact taxonomy ensures higher inter-coder agreement. We call *abstraction* (or *generalization*) the process described above.

The main stance of this paper is that, in order to obtain a normalized scheme, one can: (1) start with a theoretically-grounded taxonomy (e.g. the PDTB, or an RST-based one), and (2) submit it to empirical testing, which means using specification and generalization to make it evolve into a truly universal, empirically-grounded multilingual taxonomy.

2.2 Semantic vs. contextual meanings of discourse connectives

A difficulty for the annotation of the rhetorical relations conveyed by connectives is that connectives can be used to convey a different relation than the one(s) that they semantically encode. The best-known case of this type of semantic under-determination is the connective *and*, which often conveys in context a more specific relation than its semantic meaning of addition, notably a temporal or a causal meaning (e.g. Spooren, 1997; Carston, 2002). These relations are then called its pragmatic meanings. Most analyses treat these pragmatic meanings as inferable in context but not as part of the semantic meaning of *and*. This phenomenon is also observed with other connectives; for example, temporal connectives may at times convey a causal or a contrastive relation as their pragmatic meaning, without having these relations as part of their semantic core meaning. This phenomenon is distinct from the semantic

ambiguity of connectives (such as *since*) that can alternatively convey distinct semantic meanings (for *since*, temporal or causal).

Therefore, an important question is to define what level of meaning (semantic or pragmatic) has to be annotated. Obviously, the pragmatic relation conveyed in context is more helpful for understanding the contribution of a connective in a given utterance than its core semantic meaning. However, relations that differ in context from the semantic meaning of a connective give rise to an important number of disagreements between annotators, probably because in such cases the interpretation rests on inference, a process that varies across speakers (cf. Spooren and Degand 2010).

In our view, a way to deal with the under-determinacy question is to make annotators aware of this phenomenon and encourage the annotation of the meaning perceived in context, even when it departs from the connective's core semantic meaning. However, the latter meaning must be taken into account if the annotation is used to establish the range of possible semantic meanings of discourse connectives, and in particular if frequency information is desired. This is especially the case for lexicographic analyses which look for statistics regarding semantic meanings only.

3 Previous work and results

Evidence for the applicability of the PDTB to several languages comes from recent experiments with monolingual annotations. The PDTB has indeed set the example for a number of other lexicalized, monolingual taxonomies of discourse relations (reviewed by Webber and Joshi, 2012), namely in Czech (Zikánová et al., 2010), Arabic (Al-Saif and Markert, 2010), Chinese (Huang and Chen, 2011; Zhou and Xue, 2012), Hindi (Kolachina et al., 2012) and Turkish (Zeyrek et al., 2010). An annotation project aiming at a French Discourse Treebank is also in progress (Danlos et al., 2012). Most of these taxonomies have used the PDTB top-level classification and brought a number of adjustments to its sub-levels in order to account for all the specificities of their language. For example, in the Arabic version (Al-Saif and Markert, 2010), a *background* relation has been added as a variety of expansion. This is therefore a case of specification with respect to the PDTB taxonomy. Conversely, the subtypes of *contrast* (opposition vs. juxtaposition) and *condition* (hypothetical, etc.) were removed from the Arabic taxonomy. This goes in the direction of a generalization of the taxonomy for these labels.

Another potential source of evidence for validating multilingual taxonomies comes from recent experiments with "translation spotting" of discourse connectives in parallel corpora (typically, Hansard or Europarl). Rather than annotate connectives in each monolingual part with PDTB-style labels, this approach aims at identifying (manually or automatically) the actual translation of each connective (Danlos and Roze, 2011; Popescu-Belis et al. 2012). This deals therefore only with explicit relations, not implicit ones. By clustering afterwards the observed translations according to their meaning and frequency, it is possible to derive labels which are less precise than the PDTB ones, but are still useful for applications such as machine translation (Meyer et al. 2011) or for translation studies (Cartoni et al., 2011).

Information from translation spotting can give a lower bound on the number of different meanings a connective can convey, which can be compared to the number of labels for that connective from a PDTB-style annotation, checking for any serious mismatch. For instance, if a connective is mainly annotated with one label, but is rendered in translation by two equally frequent target connectives, it is worth examining whether the sense label should not be *specified* any further.

Manual translation spotting has been performed on a large English/French section of the Europarl corpus with about 2,500 occurrences of nine connectives (Popescu-Belis et al. 2012). It is also currently being performed on English/German/Italian parallel fragments of Europarl within the same project. An experiment with automatic English/Arabic translation spotting, using word alignment software, is also ongoing for seven English connectives, illustrating ambiguity patterns (one vs. several preferred translations).

In what follows, we present two multilingual annotation experiments with explicit discourse relations in five European languages, with an

adaptation of the PDTB in between, using the two processes of specification and generalization introduced above.

4 Applying the PDTB taxonomy to a parallel corpus of five languages

4.1 Data and procedure

In order to compare and annotate connectives in five languages, a small parallel corpus made of four journalistic texts was gathered from the www.PressEurop.eu website. The size of the corpus was around 2,500 words per language. All four texts came from different European newspapers, and the source language was different in all of them. In the English version of the corpus, used as a pivot language, 54 tokens of connectives were identified, corresponding to 23 different connective types. Connectives were defined as lexical items encoding a coherence relation between two abstract objects, following Asher (1993). The criteria used to select tokens of connectives were similar to those applied in the PDTB project. However, only connectives that had been translated by a connective in a given language were annotated. This means that a slightly different subset of all the occurrences of English connectives was annotated in each case. The list of English connectives is given in Table 1.

after (1)	despite (1)	then (1)
after all (1)	for instance (1)	therefore (2)
and (7)	however (4)	though (2)
as (1)	if (2)	thus (2)
as long as (1)	in as much as (1)	when (4)
because (2)	meanwhile (1)	whereas (1)
before (1)	nevertheless (3)	while (1)
but (11)	so (1)	

Table 1. List of connective types from the English corpus with their token frequency.

Table 2 summarizes the number of connectives that have been inserted or removed in the target languages, with respect to the English texts. All these occurrences have therefore not been annotated.

In every language, the annotation task was performed independently by two annotators. The tokens of discourse connectives to be annotated were spotted on the English version of the corpus by the authors. For every other language of the study, one annotator was asked to spot the translation equivalents. All tokens of connectives that had been translated in the target text by a connective were annotated with a discourse relation from the PDTB hierarchy by two annotators.

	French	German	Dutch	Italian
Nb. of additions	6	12	19	15
Nb. of removals	10	10	7	18
Total	16	22	26	33

Table 2. Differences in number of connectives between source and target texts.

All annotators were asked to use the definition of discourse relations provided in the PDTB annotation manual (The PDTB Research Group, 2007). As it was the case in the PDTB project, annotators were instructed to use tags from the most precise level from the hierarchy (third level) if they were confident about the relation or more generic relations in case of doubt. Annotators were also allowed to use two labels in two different cases: when they felt that the relation was ambiguous and that both tags applied; or when they felt that two tags had to be used in order to fully describe the meaning of the relation. In the first case, the two tags had to be linked with OR and in the second with AND.

4.2 Results

The inter-annotator agreement was computed from a monolingual and from a cross-linguistic perspective. The percentage of agreement for the two annotators working on the same language is reported in Table 3.

level	English	French	German	Dutch	Italian
1	98%	95%	95%	90%	94%
2	67%	69%	71%	60%	63%
3	44%	48%	51%	38%	42%

Table 3. Monolingual inter-annotator agreement.

Results from Table 3 indicate that the level of agreement is similar across languages. In every case, the agreement is very good at the first level (94% on average), medium at level 2 (66% on average) but poor at level 3 (44% on average).

By comparison, in the PDTB, the inter-annotator agreement was 92% at the top-most level and 77% at the third level of the hierarchy (Mitsalkaki *et al*., 2008).

An analysis of cases of disagreement between the monolingual annotations reveals that similar problems occur in all languages. The problematic cases mostly concern the distinction between concession and contrast, for which the annotators agree in only 50% of the relations, when the 'comparison' tag is used. This agreement even drops to 40% on average at the third level (distinctions between opposition and juxtaposition and between expectation and contra-expectation). Moreover, for the relations tagged as 'conditional', the agreement for the third level tags is also only 40%. Taken together, these cases represent on average 87% of the disagreements at the third level of the hierarchy. Finally, the use of the so-called 'pragmatic' tags from the PDTB scheme was very problematic. An agreement on the use of this tag was reached only in 16% on the cases on average, and some annotators didn't use it at all.

Cross-linguistic inter-annotator agreement is reported in Table 4.

level	English/ French	English/ German	English/ Dutch	English/ Italian
1	91%	90%	88%	85%
2	67%	66%	64%	58%
3	42%	51%	35%	35%

Table 4. Cross-linguistic inter-annotator agreement.

An analysis of cross-linguistic disagreements reveals two distinct phenomena. At the top level of the hierarchy, disagreements are always more numerous cross-linguistically than mono-lingually. These additional disagreements always correspond to meaning shifts due to translation. For example, the connective *when*, annotated with a temporal tag in English, was once translated by *alors que*, a connective annotated with a contrast tag by French-speaking annotators. Disagreements at the first level were systematically checked and discussed with annotators, with the conclusion that such cases of meaning shift occur on average in 10% of the cases in every language. This problem shows the limitations of using parallel corpora, under the assumption that connectives are translation equivalents across languages. An annotation of comparable corpora, where equivalences are established based on the similarity of rhetorical relations, does not run into similar problems.

For lower levels of the hierarchy, differences in the annotation could not be related to changes in translation but rather to genuine disagreements between annotators regarding the interpretation of a given relation. For this reason, at these levels, disagreements are on average not significantly higher cross-linguistically than monolingually.

The first annotation experiment described above clearly indicated that the areas of disagreements were recurrent across annotators and languages. In order to reach a reliable annotation that could be applied cross-linguistically, some adjustments were made to the PDTB taxonomy.

5 Proposals for revisions to the PDTB taxonomy

First, through a generalization process, the sub-categories of conditional relations were removed because in all the languages of our study, all these uses were conveyed by a single connective (*if* in English, *si* in French, *als* in Dutch, etc.). For our objective to provide an accurate representation of the meaning of connectives enabling the definition of cross-linguistic equivalences in European languages, the second level *condition* tag is fine-grained enough.

Second, the categories labeled with the PDTB 'pragmatic' tag were redefined. In the PDTB taxonomy, the kind of examples grouped under this category was not always clearly defined and therefore was rather inconsistently applied by the annotators. For example, while a reference to epistemic uses is clearly made in the case of pragmatic causes, pragmatic conditions are simply defined as "used for instances of conditional constructions whose interpretation deviates from that of the semantics of 'Condition'" (The PDTB Research Group, 2007: 31). In the revised version, the 'pragmatic' tag consistently includes all occurrences corresponding to speech-act and epistemic uses of connectives, as defined by Sweetser (1990).

Again, the rationale for this specification comes from differences in connectives. In many languages, content (non-pragmatic) and speech act and epistemic (pragmatic) relations are expressed by specific connectives (see Sanders and Stukker, 2012 for a cross-linguistic illustration in the causal domain). The pragmatic uses of connectives thus defined can occur for causal, conditional and concessive connectives. Therefore, for these tags, an additional annotation level has been specified to account for the pragmatic/non-pragmatic distinction. In the case of causals, this change involved the addition of a fourth level in the hierarchy. The addition of this level shows how certain semantic characteristics of relations occur across several categories, which leads to a systematic proposal (cf. Sanders et al., 1992).

1. Temporal
- synchronous
- asynchronous
  - precedence
  - succession

2. Contingency
- cause
  - reason
    - pragmatic
    - non-pragmatic
  - result
    - pragmatic
    - non-pragmatic
- condition
    - pragmatic
    - non-pragmatic

3. Comparison
- contrast
- concession
  - pragmatic
  - non-pragmatic
- parallel

4. Expansion
- conjunction
- instantiation
- restatement
  - specification
  - equivalence
  - generalization
- alternative
- exception
- list

Figure 1. Revised taxonomy based on the results of multilingual annotation.

Third, the *comparison* category was reorganized through a process of generalization. More specifically, the third level from the PDTB was removed, because it did not contribute to make additional distinctions between connectives. Furthermore, a 'parallel' tag was added, in order to account for the meaning of connectives such as *similarly*, which did not have a suitable tag in the PDTB taxonomy. All these changes lead to the revised taxonomy described in Figure 1. Similar adjustments were already proposed in some monolingual adaptations of the PDTB, notably in Arabic by Al-Saif and Markert (2010).

6 Annotation experiment with the revised taxonomy

A second corpus was gathered from the PressEurop website, including the same five languages used in the first experiment. This corpus, of about 8,500 words per language, contained in English 203 tokens of connectives corresponding to 36 different types (Table 5).

after (1)	given (that) (2)	since (1)
although (6)	however (7)	so (2)
and (50)	if (11)	that is why (1)
as (3)	in fact (1)	then (3)
as well as (1)	in order to (1)	therefore (3)
because (5)	in other words (1)	though (5)
before (4)	in short (1)	thus (2)
but (41)	in spite of (1)	well (1)
despite (6)	indeed (1)	when (7)
even if (4)	meanwhile (1)	whether (2)
for example (3)	now (2)	while (9)
for instance (1)	or (5)	yet (8)

Table 5. Connective types with token frequency.

In every language, the translation equivalents were spotted. The number of explicitly translated connectives ranged from 136 to 155. The important number of non-translated connectives provides further indication of the important volatility of these lexical items in translation. The rhetorical relations conveyed by explicit connectives were annotated with the revised taxonomy described in Figure 1. Results from the annotation task are reported in Table 6.

	English/ French	**English/ German**	**English/ Dutch**	**English /Italian**
level 1	94%	93%	88%	93%
level 2	85%	74%	75%	78%
level 3	75%	66%	69%	66%
level 4	66%	93%	62.5%	70%

Table 6. Cross-linguistic inter-annotator agreement.

These results confirm the validity of our second monolingual annotation experiment, with cross-linguistic data. The improvement of agreement scores with respect to the first experiment are significant, and the additional coverage of connective types did not reveal the need for additional relations or the existence of important differences between languages. This

experiment also confirmed that most disagreements at the first level of the taxonomy were due to meaning shifts in translation, as confirmed through manual checking and discussion with the annotators.

7 Conclusion

This paper is a first attempt towards a unified framework designed to relate connectives to one another over the languages. This existence of such a framework is a sorely needed resource for many domains such as applied linguistics, translation and language engineering. Such a resource is all the more necessary because existing multilingual resources such as bilingual dictionaries and contrastive grammars are insufficient to correctly describe them.

Yet, much work remains to be done to achieve this goal. Importantly, larger scale annotation experiments involving more languages and tokens for the annotation should be carried out. Another important step will be to test the granularity of the taxonomy by systematically comparing all tokens annotated with the same label, both monolingually and cross-linguistically, in order to ensure that they provide genuine semantic equivalences. In other words, the need for additional specifications should be systematically checked. Finally, another important step will be to include the implicit dimension in the cross-linguistic comparison of connectives. In some cases, the absence of connectives seems to be the preferred translation choice. A case in point is the French connective *en effet*, very frequently used to mark an elaboration, and most of the time not translated into English. Similar cases should be detected, and zero translations taken into account as possible translation equivalents.

Acknowledgments

This study was partially funded by the Swiss National Science Foundation through a grant awarded to the first author (SNSF Grant PA00P1_139613/1). The article was also supported by the COMTIS Swiss NSF Sinergia project (www.idiap.ch/comtis). The second author is Senior Research Associate with the Belgian National Science Foundation (FSR-FNRS). The authors thank all the annotators who took part in the experiments.

References

Al-Saif, A. and Markert K., 2010. "The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic". *Proceedings of the 7th Int. Conf. on Language Resources and Evaluation (LREC 2010)*, Marrakech, p.2046-2053.

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer.

Bunt H., Prasad R. and Joshi A. 2012. "First steps towards an ISO standard for annotating discourse relations". *Proceedings of ISA-7 workshop (Interoperable Semantic Annotation) at LREC 2012*, Istanbul.

Carston R. 2002. *Thoughts and Utterances. The Pragmatics of Explicit Communication*. Oxford: Blackwell.

Cartoni, B. Zufferey, S. Meyer, T. Popescu-Belis, A. 2011. "How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives". *Proceedings of 4th Workshop on Building and Using Comparable Corpora*, Portland, OR.

Danlos L. and Roze C. 2011. "Traduction (automatique) des connecteurs de discours". *Proceedings of TALN 2011*, 6 p., Montpellier.

Danlos L., Antolinos-Basso D., Braud C., and Roze C. 2012. "Vers le FDTB : French Discourse Tree Bank". *Proceedings of JEP-TALN-RECITAL 2012*, vol. 2, p. 471–478, Grenoble.

Dixon, R.M. and Aikhenvald, A. (eds). 2009. *The Semantics of Clause Linking. A Cross-Linguistic Typology*. Oxford: Oxford University Press.

Granger, S. and Tyson, S. 1996. "Connector Usage in English Essay Writing of Native and Non-Native EFL Speakers of English". *World Englishes* 15(1):19-29.

Halverson, S. 2004. "Connectives as a Translation Problem". In *An International Encyclopedia of Translation Studies*, H.

Kittel *et al.* (eds), 562-572. Berlin/New York: Walter de Gruyter.

Huang, H.-H. and Chen, H.-H. 2011. "Chinese discourse relation recognition". *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 1442-1446.

Kolachina, S., Prasad, R., Sharma, D. and Joshi, A. 2012. "Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank". Proceedings of *of LREC 2012 (8th International Conference on Language Resources and Evaluation)*, Istanbul.

Meyer T., Popescu-Belis A., Zufferey S. and Cartoni B. 2011. "Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation". *Proceedings of SIGDIAL 2011 (12th annual SIGdial Meeting on Discourse and Dialogue)*, Portland, OR, p.194-203.

Miltsakaki, E. Robaldo, L. Lee, A. and Joshi, A. 2008. "Sense Annotation in the Penn Discourse Treebank". *Lecture Notes in Computer Science* 4919: 275-286.

Popescu-Belis A., Meyer T., Liyanapathirana J., Cartoni B. & Zufferey S. 2012. "Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns". *Proceedings of LREC 2012 (8th International Conference on Language Resources and Evaluation)*, Istanbul.

Prasad, R. Dinesh, N. Lee, A. *et al.* (2008). "The Penn Discourse TreeBank 2.0". *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2961-2968.

Sanders, T., Spooren, W. and Noordman, L. 1992. Towards a taxonomy of coherence relations. *Discourse Processes* 15: 1-35.

Sanders, T. & Stukker, N. 2012. "Causal Connectives in Discourse: A Cross-Linguistic Perspective". *Journal of Pragmatics* 34(2): 131-137.

Spooren, W. 1997. "The Processing of Underspecified Coherence Relations". *Discourse Processes* 24(1): 149-168.

Spooren, W. and Degand, L. 2010. "Coding Coherence Relations: Reliability and Validity". *Corpus Linguistics and Linguistic Theory* 6 (2): 241-266.

Sweetser, E. 1990. *From Etymology to Pragmatics*. Cambridge: Cambridge University Press.

The PDTB Research Group. 2007. "The Penn Discourse Treebank 2.0 Annotation Manual". *IRCS Technical Reports Series*, 99 p.

Webber B. and Joshi A. 2012. "Discourse Structure and Computation: Past, Present and Future". *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, p. 42–54, Jeju, Republic of Korea.

Zeyrek D., Demirsahin I., Sevdik-Calli A., Balaban H.O., Yalcinkaya I., and Turan U.D. 2010. "The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations". *Proceedings of the Fourth Linguistic Annotation Workshop at ACL 2010*, p. 282–289, Uppsala, Sweden.

Zhou, Y. and Xue, N. 2012. "PDTB-style Discourse Annotation of Chinese Text". *Proceedings of ACL 2012 (The 50th Annual Meeting of the Association for Computational Linguistics)*, p.69-77, Jeju, Republic of Korea.

Zikánová, S., Mladová, L. Mírovský, J. and Jínová, P. 2010. "Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank". *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, p.2002-2006, Valetta.