The Workshop Programme

Monday, May 26

14:30 -15:00 Opening
Nancy Ide and Adam Meyers

15:00 -15:30 SIGANN Shared Corpus Working Group Report
Adam Meyers

15:30 -16:00 Discussion: SIGANN Shared Corpus Task

16:00 -16:30 Coffee break

16:30 -17:00 Towards Best Practices for Linguistic Annotation
Nancy Ide, Sameer Pradhan, Keith Suderman

17:00 -18:00 Discussion: Annotation Best Practices

18:00 -19:00 Open Discussion and SIGANN Planning

Tuesday, May 27

09:00 -09:10 Opening
Nancy Ide and Adam Meyers

09:10 -09:50 From structure to interpretation: A double-layered annotation for event factuality
Roser Saurí and James Pustejovsky

09:50 -10:30 An Extensible Compositional Semantics for Temporal Annotation
Harry Bunt, Chwhynny Overbeeke

10:30 -11:00 Coffee break

11:00 -11:40 Using Treebank, Dictionaries and GLARF to Improve NomBank Annotation
Adam Meyers

11:40 -12:20 A Dictionary-based Model for Morpho-Syntactic Annotation
Cvetana Krstev, Svetla Koeva, Duško Vitas

12:20 -12:40 Multiple Purpose Annotation using SLAT - Segment and Link-based Annotation Tool (DEMO)
Masaki Noguchi, Kenta Miyoshi, Takenobu Tokunaga, Ryu Iida, Mamoru Komachi, Kentaro Inui

12:40 -14:30 Lunch

14:30 -15:10 Using inheritance and coreness sets to improve a verb lexicon harvested from FrameNet
Mark McConville and Myroslava O. Dzikovska

15:10 -15:50 An Entailment-based Approach to Semantic Role Annotation
Voula Gotsoulia

16:00 -16:30 Coffee break

16:30 -16:50 A French Corpus Annotated for Multiword Expressions with Adverbial Function
Eric Laporte, Takuya Nakamura, Stavroula Voyatzi

16:50 -17:20 On Construction of Polish Spoken Dialogs Corpus
Agnieszka Mykowiecka, Krzysztof Marasek, Małgorzata Marciniak, Joanna Rabiega-Wisniewska, Ryszard Gubrynowicz

17:20 -17:40 A RESTful interface to Annotations on the Web
Steve Cassidy

17:40 -18:30 Panel : Next Challenges in Annotation Theory and Practice

18:30 -19:00 Open Discussion

Workshop Organisers

Nancy Ide, Vassar College (co-chair)
Adam Meyers, New York University (co-chair)

Inderjeet Mani, MITRE Corporation
Antonio Pareja-Lora, SIC, UCM / OEG, UPM
Sameer Pradhan, BBN Technologies
Manfred Stede, Universitat Potsdam
Nianwen Xue, University of Colorado

Workshop Programme Committee

David Ahn, Powerset
Lars Ahrenberg, Linkoping University
Timothy Baldwin, University of Melbourne
Francis Bond, NICT
Kalina Bontcheva, University of Sheffield
Matthias Buch-Kromann, Copenhagen Business School
Paul Buitelaar, DFKI
Jean Carletta, University of Edinburgh
Christopher Cieri, Linguistic Data Consortium/University of Pennsylvania
Hamish Cunningham, University of Sheffield
David Day, MITRE Corporation
Thierry Declerck, DFKI
Ludovic Denoyer, University of Paris 6
Richard Eckart, Darmstadt University of Technology
Tomaz Erjavec, Jozef Stefan Institute
David Farwell, New Mexico State University
Alex Chengyu Fang, City University Hong Kong
Chuck Fillmore, International Computer Science Institute
John Fry, San Jose State University
Claire Grover, University of Edinburgh
Eduard Hovy, Information Sciences Institute
Baden Hughes, University of Melbourne
Emi Izumi, NICT
Aravind Joshi, University of Pennsylvania
Ewan Klein, University of Edinburgh
Mike Maxwell, University of Maryland
Stephan Oepen, University of Oslo
Martha Palmer, University of Colorado
Manfred Pinkal, Saarland University
James Pustejovsky, Brandeis University
Owen Rambow, Columbia University
Laurent Romary, Max-Planck Digital Library
Erik Tjong Kim Sang, University of Amsterdam
Graham Wilcock, University of Helsinki
Theresa Wilson, University of Edinburgh

Table of Contents

Long papers

From structure to interpretation: A double-layered annotation for event factuality
Roser Saurí, James Pustejovsky 1

An Extensible Compositional Semantics for Temporal Annotation
Harry Bunt, Chwhynny Overbeeke 9

Using Treebank, Dictionaries and GLARF to Improve NomBank Annotation
Adam Meyers 17

A Dictionary-based Model for Morpho-Syntactic Annotation
Cvetana Krstev, Svetla Koeva, Duško Vitas 25

Using inheritance and coreness sets to improve a verb lexicon harvested from FrameNet
Mark McConville, Myroslava O. Dzikovska 33

An Entailment-based Approach to Semantic Role Annotation
Voula Gotsoulia 41

Short papers

A French Corpus Annotated for Multiword Expressions with Adverbial Function
Eric Laporte, Takuya Nakamura, Stavroula Voyatzi 48

On Construction of Polish Spoken Dialogs Corpus
Agnieszka Mykowiecka, Krzysztof Marasek, Małgorzata Marciniak, Joanna Rabiega-Wisniewska, Ryszard Gubrynowicz 52

A RESTful interface to Annotations on the Web
Steve Cassidy 56

Demonstration

Multiple Purpose Annotation using SLAT - Segment and Link-based Annotation Tool
Masaki Noguchi, Kenta Miyoshi, Takenobu Tokunaga, Ryu Iida, Mamoru Komachi, Kentaro Inui 61

Author Index

Bunt, Harry 9
Cassidy, Steve 56
Dzikovska, Myroslava O. 33
Gotsoulia, Voula 41
Gubrynowicz, Ryszard 52
Iida, Ryu 61
Inui, Kentaro 61
Koeva, Svetla 25
Komachi, Mamoru 61
Krstev, Cvetana 25
Laporte, Eric 48
Marasek, Krzysztof 52
Marciniak, Małgorzata 52
McConville, Mark 33
Meyers, Adam 17
Miyoshi, Kenta 61
Mykowiecka, Agnieszka 52
Nakamura, Takuya 48
Noguchi, Masaki 61
Overbeeke, Chwhynny 9
Pustejovsky, James 1
Rabiega-Wisniewska, Joanna 52
Saurí, Roser 1
Tokunaga, Takenobu 61
Vitas, Duško 25
Voyatzi, Stavroula 48

From structure to interpretation: A double-layered annotation for event factuality

Roser Saurí and James Pustejovsky

Lab for Linguistics and Computation
Computer Science Department
Brandeis University
{roser,jamesp}@cs.brandeis.edu

Abstract

Current work from different areas in the field points out the need for systems to be sensitive to the factuality nature of events mentioned in text; that is, to recognize whether events are presented as corresponding to real situations in the world, situations that have not happened, or situations of uncertain status. Event factuality is a necessary component for representing events in discourse, but for annotation purposes it poses a representational challenge because it is expressed through the interaction of a varied set of structural markers. Part of these factuality markers is already encoded in some of the existing corpora but always in a partial way; that is, missing an underlying model that is capable of representing the factuality value resulting from their interaction. In this paper, we present FactBank, a corpus of events annotated with factuality information which has been built on top of TimeBank. Together, TimeBank and FactBank offer a double-layered annotation of event factuality: where TimeBank encodes most of the basic structural elements expressing factuality information, FactBank adds a representation of the resulting factuality interpretation.

1. Introduction

In the past decade, most efforts towards corpus construction have been devoted to encoding a variety of semantic information structures. For example, much work has gone to annotating the basic units that configure propositions (PropBank, FrameNet) and the relations these hold at the discourse level (RST Corpus, Penn Discourse TreeBank, GraphBank), as well as specific knowledge that has proved fundamental in tasks requiring some degree of text understanding, such as temporal information (TimeBank) and opinion expressions (MPQA Opinion Corpus).[1]

The field is moving now towards finding platforms for unifying them in an optimal way –e.g., Pradhan et al. (2007); Verhagen et al. (2007). It therefore seems we are at a point where the first elements for text understanding can be brought together.

Nonetheless, current work from different areas in the field points out the need for systems to be sensitive to an additional level of information; namely, that conveying whether events in text are presented as corresponding to real situations in the world, situations that have not happened, or situations of uncertain status. We refer to this level as *event factuality*.

The need for this further type of information is demonstrated in highly domain-oriented disciplines such as bioinformatics (Light et al., 2004), as well as more genre-oriented tasks. For example, Karttunen & Zaenen (2005) discusses the relevance of veridicity for IE. Factuality is critical also in the area of opinion detection (Wiebe et al., 2005), given that the same situation can be presented as a fact in the world, a mere possibility, or a counterfact according to different sources. And in the scope of textual entailment, it has been taken as a basic feature in some of the systems participating in (or using the data from) previous PASCAL RTE challenges.

For example, Tatu & Moldovan (2005) treat intensional contexts, de Marneffe et al. (2006) look at features accounting for the presence of polarity, modality, and factivity markers in the textual fragments, while Snow & Vanderwende (2006) check for polarity and modality scoping over matching nodes in a graph. Most significantly, the system that obtained the best absolute result in the three RTE challenges, scoring an 80% accuracy (Hickl & Bensley, 2007), is based on identifying the set of publicly-expressed beliefs of the author; that is, on the author's commitments of how things are in the world according to what is expressed in text –either asserted, presupposed, or implicated.

Event factuality is a necessary component for representing events in discourse, together with other levels of information such as argument structure or temporal information. Inferences derived from events that have not happened, or that are only possible, are different from those derived from events judged as factual in nature. For instance, it is basic for temporally ordering the events in a given text.

For annotation purposes, however, it poses a representational challenge. The factuality of events is expressed through the interaction of elements from different linguistic categories. It involves, for instance, polarity (events can be presented as positive or negative) as well as modality – epistemic modality, for instance, expresses the degree of certainty of a source about what is asserted, and events qualified with other types of modality are generally presented as mere possibilities. Other information at play is evidentiality (e.g., a seen event is presented with a factuality degree stronger than that of an event reported by somebody else) or mood (e.g., indicative vs. subjunctive). Factuality is also a component in the semantics of specific syntactic structures with presuppositional effects (e.g., appositions and relative clauses), as well as certain types of

[1] The main references for these corpora are: PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998), RST Corpus (Carlson et al., 2002), Penn Discourse TreeBank (Miltsakaki et al., 2004), GraphBank (Wolf & Gibson, 2005), TimeBank (Pustejovsky et al., 2003), MPQA Opinion Corpus (Wiebe et al., 2005).

predicates –most notoriously, the so-called factive and implicative predicates, but also others; compare, for instance, the effect that *decision* in (1a) and *refusal* in (1b) have on the factuality status of the underlined event.

(1) a. A senior Russian politician has hailed a **decision** by Uzbekistan to shut down a United States military base.

b. A senior Russian politician has hailed the **refusal** by Uzbekistan to shut down a United States military base.

Part of these factuality markers are already encoded in some of the existing corpora (for example, TimeBank annotates polarity particles, modality operators, as well as the aforementioned predicates), but always in a partial way; that is, missing an underlying model capable of representing the factuality value that results from their interaction.

In this paper, we introduce FactBank, a corpus of events annotated with factuality information which has been built on top of TimeBank. Together, TimeBank and FactBank offer a double-layered annotation of event factuality: the former encodes most of the basic structural elements expressing factuality information, whereas the latter represents the resulting factuality interpretation.

In the next section, we set the linguistic grounding of our work by defining event factuality as a semantic property of events, establishing its possible values, and identifying its structural markers. Then, section 3 presents the main challenges for automatically recognizing it, which motivate the double-layered corpus annotation. We review some of the existing corpora where this information has already been annotated in section 4. Finally, section 5 focuses on FactBank, which is evaluated in section 6.

2. Linguistic foundations

2.1. What is event factuality

Eventualities in discourse can be couched in terms of a veridicality axis that ranges from truly factual to counterfactual, passing through a whole spectrum of degrees of modality. In some contexts, the factual status of events is presented with absolute certainty. Events are then characterized as *facts* (2) or *counterfacts* (5). Other contexts introduce different shades of uncertainty. Depending on the polarity, events are then qualified as *possibly factual* (3) or *possibly counterfactual* (4).

(2) Five U.N. inspection teams visited a total of nine other sites.

(3) United States may extend its naval quarantine to Jordan's Red Sea port of Aqaba.

(4) They may not have enthused him for their particular brand of political idealism.

(5) The size of the contingent was not disclosed.

Factuality can therefore be characterized as involving polarity and modality (more precisely, epistemic modality). Polarity is a discrete category with two values, positive and negative. Epistemic modality expresses the speaker's degree of commitment to the truth of the proposition (Palmer, 1986), which ranges from uncertain (or possible) to absolutely certain (or necessary). For methodological reasons, however, we need a discrete categorization of that system.

2.2. Factuality values

Within modal logic, two operators are typically used to express a modal context: necessity ($\Box$) and possibility ($\Diamond$); e.g., Lewis (1968). On the other hand, most of the work in linguistics points towards a three-fold distinction: *certain*, *probable*, and *possible*; e.g., (Lyons, 1977; Halliday & Matthiessen, 2004). Interestingly, Horn (1989) analyzes modality and its interaction with polarity based on both linguistic tests and logical relations at the basis of the Aristotelian Square of Opposition. He presents modality as a continuous category. Yet, he provides a good grounding for differentiating the three major modality degrees just mentioned. Based on that, we represent factuality by means of the features in Table 1:

Table 1: Factuality values

	Positive	**Negative**	**Underspecified**
Certain	Fact: <CT,+>	Counterfact: <CT,−>	Certain but unknown output: <CT, u>
Probable	Probable: <PR,+>	Not probable: <PR,−>	(NA)
Possible	Possible: <PS,+>	Not certain: <PS,−>	(NA)
Underspecif.	(NA)	(NA)	Unknown or uncommitted: <U,u>

The factual value of events is then presented as a tuple $<mod, pol>$, containing a modality and a polarity value.[2] The polarity axis divides into positive, negative, and unknown, while the modality axis distinguishes among certain (CT), probable (PR), possible (PS), and unknown (UN). The *unknown* values are added to account for cases of uncommitment.

The table includes six fully committed (or specified) values (<CT,+>, <CT,−>, <PR,+>, <PR,−>, <PS,+>, <PS,−>), and two underspecified ones: the partially underspecified <CT,u>, and the fully underspecified <U,u>.

The partially underspecified value, <CT,u>, is for cases where there is total certainty about the factual nature of the event but it is not clear, however, what the output is –e.g., (6). The fully underspecified <U,u>, on the other hand, is used when any of the following situations applies: (*i*) The source does not know what is the factual status of the event, as in (7a); (*ii*) the source is not aware of the possibility of the event –e.g., (7b); or (*iii*) the source does not overtly commit to it –e.g., (7c). The following examples illustrate each of these preceding situations for the underlined event when evaluated by source *John*:

(6) **John** knows whether Mary came.

(7) a. **John** does not know whether Mary came.
b. **John** does not know that Mary came.
c. **John** knows that Paul said that Mary came.

For simplicity, in what follows the factuality values will be represented in the abbreviated form of CT+, PR−, Uu, etc.

[2] Semantically, this can be interpreted as: $Val(mod)(Val(pol)(e))$ –i.e., the modal value scopes over the polarity value.

2.3. Discriminatory tests

In characterizing the factuality of events, the polarity parameter offers no problem, but distinguishing between the modality values (e.g., between *possible* and *probable*) is not always evident. In order to determine the modality parameter, we designed a battery of tests based on the logical relations considered in Horn (1989) to pinpoint the basic categories of epistemic modality; i.e., Law of Contradiction and Law of Excluded Middle. They are copredication tests.

Underspecification (U) versus different degrees of certainty (CT, PR, PS): Events with an underspecified value can be copredicated with both: a context in which they are characterized as certainly happening (CT+), and a context in which they are presented as certainly not happening (CT−). For example, sentence (8) can be continued by either fragment in (10), the first of which maintains the original underlined event as certainly happening (CT+), and the second as certainly not happening (CT−). This is not the case, however, for sentence (9), where the underlined event is explicitly characterized as probable.

(8) Iraq has agreed to allow Soviets in Kuwait to leave.

(9) Soviets in Kuwait will most probably leave.

(10) a. ... They will take the plane tomorrow early in the morning. (CT+)
b. ... However, most of them decided to remain there. (CT−)

Absolute certainty (CT) versus degrees of uncertainty (PR, PS): Eventualities presented as certain (CT) cannot at the same time be assessed as *possible* (PS) in a context of *opposite polarity*. In the examples below, the symbol # is used to express that there is some sort of semantic anomaly.

(11) a. Hotels are only thirty (CT+) percent full.
b. #... but it is possible that they aren't (PS−).

(12) a. Nobody believes (CT−) this anymore.
b. #... but it is possible that somebody does (PS+).

On the other hand, eventualities characterized with some degree of uncertainty (PS or PR) allow for it:

(13) a. I *think* it's not going to change (PR−) for a couple of years.
b. ... but it *could* happen otherwise. (PS+)

(14) a. He *probably* died (PR+) within weeks or months of his capture.
b. ... but it is also possible that the kidnappers kept him alive for a while. (PS-)

In (13), the source expressed by the pronoun *I* characterizes the underlined event as PR− by presenting it under the scope of the predicate *think* used in 1st person. The fragment in (13b) can be added without creating any semantic anomaly. A similar situation is presented in (14): the adverb *probably* is characterizing the event as PR+, and the additional fragment presents the possibility of things being otherwise.

Probable (PR) versus possible (PS):

As seen, both degrees of uncertainty (PR and PS) accept copredication with PS in a context of opposite polarity. However, only the lowest degree of uncertainty (PS) accepts copredication with PR in a context of opposite polarity.

(15) a. I *think* it's not going to change (PR−) for a couple of years.
b. #... but it *probably* will. (PR+)

(16) a. It *may* not change (PS−) for a couple of years.
b. ... but it most *probably* will. (PR+)

Table 2 summarizes the different copredication tests just introduced. The resulting epistemic modality values assigned to events are listed in the rows, while the tests are presented in the columns, abbreviated as $\text{EM}_{subindex}$. EM expresses the epistemic modality value of the context to be copredicated to the original sentence, whereas $subindex$ indicates its polarity: $=$ means context of the same polarity, and op, context of opposite polarity.

Table 2: Tests for discriminating among modality degrees.

	CT$_=$	**CT**op	**PR**op	**PS**op
U	ok	ok	ok	ok
PS	ok	#	ok	ok
PR	ok	#	#	ok
CT	ok	#	#	#

For example, given an event e presented under a context of negative polarity in its original sentence, test PR_{op} requires the creation of a new fragment in which e is used in a context where the modality degree is *probable* and the polarity is positive: PR+.[3]

(17) Original: I *think* it's not going to change. (PR−)
Testing e_2 with PR_{op}: #... but it probably will. (PR+)

2.4. Factuality markers

Event factuality in natural language is marked by both lexical items and syntactic constructions.

2.4.1. Lexical Markers

Event Selecting Predicates (ESPs). These are predicates (verbs, nouns, or adjectives) that select for an argument denoting an eventuality of any sort. Syntactically, they subcategorize for *that*-, gerundive-, and *to*- clauses, or NPs headed by event-denoting nouns. The ESPs in (18) are in bold face; their embedded events, underlined.

(18) a. Uri Lubrani also **suggested** Israel was **willing** to withdraw from southern Lebanon.
b. Kidnappers **kept** their **promise** to kill a store owner they took hostage.

[3] As appreciated, test CT= is non-discriminative. It is added there because, when combined with CPop, it allows to identify U values from the rest.

ESPs contribute to characterizing the factuality of the event denoted by its complement. For example, complements to weak assertive predicates (Hooper, 1975) (*think*, *suppose*) are depicted as not totally certain; complements of reporting predicates (Bergler, 1992) are presented as certain according to a particular source; factive (*regret, know*) and implicative predicates (*manage, prevent*) characterize their embedded complements as either factual or counterfactual (Kiparsky & Kiparsky, 1970; Karttunen, 1970, 1971); and arguments of volition and commitment predicates (*wish; offer*) are presented as possible in a future point in time.

Modal Particles. These include modal auxiliaries (*could, may, must*), but also clausal and sentential adverbial modifiers (*maybe, likely, possibly*).

Polarity Particles. These include elements of a varied nature: adverbs (*not, until*), quantifiers (*no, none*), pronouns (*nobody*), etc. They switch the polarity of its context. When scoping over a modal particle, they also affect its modal interpretation.

2.4.2. Syntactic Contexts

Syntactic structures conveying factuality information involve two clauses, one embedded under the other. In some cases, the embedded event is presupposed as holding; e.g., relative clauses (19), cleft sentences (20), and subordinated temporal clauses.

(19) *Rice, who became secretary of state two months ago today, took stock of a period of tumultuous change.*

(20) *It was Mr. Bryant who, on July 19, 2001, asked Rep. Bartlett to pen and deliver a letter to him.*

In others, the event denoted by the embedded clause is intensional in nature; e.g., purpose clauses (21) and conditional constructions (22).

(21) *The environmental commission must adopt regulations to ensure people are not exposed to radioactive waste.*

(22) *EZLN will return to the negotiating table if the conflict zone is* **demilitarized**.

3. Challenges in identifying event factuality

Annotating event factuality poses challenges at two levels. First, factuality is in many cases the result of different factuality markers interacting among them. They can all be in the local context of the event, but it is also common for them to be at different levels. Second, the factuality of an event is always relative to one or more sources. Hence, they must be included as part of the annotation scheme as well. The following subsections elaborate on these two issues. Refer to Saurí (2008) for a more comprehensive view on event factuality and its identification.

3.1. Interpreting the factuality of events

Event factuality involves local but also non-local information. Consider the following examples:[4]

(23) a. The Royal Family will **continue** to **allow** detailed fire brigade **inspections**$_e$ of their private quarters.

b. The Royal Family will **continue** to **refuse** to **allow** detailed fire brigade **inspections**$_e$ of their private quarters.

c. The Royal Family **may refuse** to **allow** detailed fire brigade **inspections**$_e$ of their private quarters.

The event *inspections* in (23a), where *allow* is embedded under the factive predicate *continue*, is characterized as a fact in the world –i.e., there have been such inspections. Example (23b), on the other hand, depicts *inspections* as a counterfact because of the effect of the predicate *refuse* scoping over *allow*. Now contrast the two previous sentences with that in (23c), where the factual status of the event *inspections* is uncertain due to the modal auxiliary *may* scoping over *refuse*.

Hence, the factuality status of a given event cannot be obtained from the strict local modality and polarity operators scoping over that event but, if present, appealing to their interaction with other non-local markers as well. Consequently, annotating factuality from a surface-based approach, accounting for the structural elements and without considering their interaction, will miss an important piece of information.

3.2. Relevant sources

The second challenge to encoding event factuality involves the notion of perspective. Different discourse participants may present divergent views about the factuality nature of the very same event. Recognizing these sources is crucial for any task involving text entailment, such as question answering or narrative understanding. For example, event e in (24) (i.e., Slobodan Milosevic having been murdered in The Hague) will be inferred as a fact in the world if it cannot be qualified as the assertion of a specific source; namely, Milosevic's son.

(24) Slobodan Milosevic's son said Tuesday that the former Yugoslav president had been **murdered**$_e$ at the detention center of the UN war crimes tribunal in The Hague.

By default, events mentioned in discourse always have an implicit source, viz., the author of the text. Additional sources are introduced in discourse by means of predicates of reporting (*say, tell*), knowledge and opinion (e.g., *believe, know*), psychological reaction (*regret*), etc. Because of their role in introducing a new source, we call them Source Introducing Predicates (SIPs).

The status of the additional sources is, however, different from that of the author of the text. For instance, in (25) the reader learns Izvestiya's position only according to what the author asserts –in other words, the reader does not have direct access to the factual assessment of Izvestiya about event e_2 –or, for that matter, to the assessment of G-7 leaders about e_3.

(25) Izvestiya **said**$_{e_1}$ that the G-7 leaders **pretended**$_{e_2}$ everything **was OK**$_{e_3}$ in Russia's economy.

Thus, we need to appeal to the notion of *nested source* as presented in Wiebe et al. (2005). *Izvestiya* is not a licit source of the factuality of event e_2, but *Izvestiya* according to the author instead, represented here as *izvestiya_author*.[5]

[4] As startling as it may result, the original sentence in this set is (23b), from the BNC.

[5] Equivalent to the notation $<author,izvestiya>$ in Wiebe's work.

Similarly, the source referred to by the G-7 leaders corresponds to the chain: *g7leaders_izvestiya_author*.
As it happens, the same event can have more than one relevant source relative to which its factuality is assessed. In some cases, they coincide in the factual status of the event but in others there is disagreement. In (25), for example, event e_3 is assessed as being a fact (CT+) according to the G-7 leaders (corresponding to source *g7leaders_izvestiya_author*), but as being false (CT−) according to Izvestiya (i.e., *izvestiya_author*). The text author, on the other hand, remains uncommitted (Uu).
The factuality value assigned to events in text must be relative to the relevant sources at play, which may be one or more. Only under this assumption it is possible to account for the potential contradictions between factual values assigned to the same event, and the different opinions commonly found in news reports.

4. Factuality information in existing corpora

To our knowledge, factuality-related information is annotated in three corpora: the MPQA Opinion Corpus (Wiebe et al., 2005), the Penn Discourse TreeBank (Miltsakaki et al., 2004), and TimeBank (Pustejovsky et al., 2003). Currently, it is also being annotated in the ACE 2008 program.[6]
The factuality-relevant expressions annotated in the MPQA Opinion Corpus are private states (opinions, beliefs, thoughts) and speech events. They both convey the stance of a source with regard to what is believed or said. Nevertheless, event factuality is not the focus of the annotation, and hence these events and states are not characterized in terms of the factual degree they convey but in terms of perspective (i.e., objective vs. subjective).
Another common feature between the MPQA Opinion Corpus scheme and our model of event factuality is the encoding of sources. Both approaches structure them as chains of nested sources. From our perspective, however, the MPQA Opinion Corpus is limited in that it only acknowledges one relevant source for each event.
Another limitation in the MPQA annotation scheme is that it is not grammatically grounded. That is, the annotation of text spans is not guided according to the grammatical structure of the sentence, and this can pose an obstacle for tasks of automatic recognition.
The Penn Discourse TreeBank (PDTB) seems closer to our perspective in that it contemplates the attribution of abstract objects (corresponding here to what we refer to as events), and encodes both their sources and the degree of factuality associated to them (Prasad et al., 2007). The task is approached from a compositional approach, contrary to the MPQA Opinion Corpus.
In spite of these similarities, there are two significant differences. With regard to sources, PDTB does not encode the nesting relation that can hold among them, neither accounts for the possibility of more than one source for a given abstract object (or event).
The second difference concerns the factuality degree associated to the attributed event, which is assigned based on the type of action described by the predicate embedding it. In particular, events embedded under communication predicates are characterized as asserted; events embedded by propositional attitude predicates, as beliefs; and events embedded under factive predicates, as facts. As it happens, however, each of these types of predicates is not uniform in terms of the factuality they project to the embedded event. *Suggest*, for instance, is a communication verb which nevertheless conveys a nuance of belief. Similarly, *forget* is a factive predicate which, contrary to others in its class, expresses an uncommitted (or ignorant) stance of the source (i.e., the participant expressed by its subject) with regards to the factual status of its embedded complement. The classification misses therefore important factuality distinctions. Finally, PDTB annotation is not concerned with the effect of other markers of modality (modal auxiliaries and adverbials) on the factuality of abstract objects.
The last corpus to evaluate is TimeBank, a corpus annotated with TimeML (Pustejovsky et al., 2005), a specification language representing temporal and event information in text. Given the surface-based approach of TimeML, TimeBank is the corpus that takes the most compositional approach to annotation among the three reviewed corpora.
The factuality-relevant information encoded in TimeBank is mainly lexical: grammatical particles expressing event modality and polarity, as well as event selecting predicates (*cf.* section 2.4.1.), which project a factual value to their embedded event by means of subordination links (or slinks). Thus, TimeBank provides us with the basic components expressing factuality information in text –a consequence of the explicit surface-based approach of TimeML. And whereas there is some characterization of event factuality (through slinks), it does not deal with the interaction among the different markers scoping over the same event.

5. Creating a corpus of event factuality

5.1. FactBank

FactBank is a corpus annotated with factuality information. It consists of 208 documents and contains a total of 8837 events manually annotated. FactBank includes all the documents in TimeBank and a subset of those in the AQUAINT TimeML Corpus (A-TimeML Corpus)[7]. The contribution of each of these corpora to FactBank is shown in Table 3.

Table 3: FactBank sources

	# Documents		# Events	
TimeBank	183	(88%)	7935	(90%)
A-TimeML Corpus	25	(12%)	902	(10%)
Total	**208**		**8837**	

Because both TimeBank and AQUAINT TimeML Corpus are annotated with the TimeML spec, FactBank incorporates a second layer of factuality information on top of that in the original corpora. Thus, while the former two encode the structural elements expressing factuality information in language, the latter represents the resulting interpretation. The new annotation is kept in separate documents

[6] http://projects.ldc.upenn.edu/ace/annotation/. Because it still is an ongoing project, we will not comment on that corpus here.

[7] http://www.timeml.org/site/timebank/timebank.html

and is linked to the original data by means of the events IDs, which are the same in both annotation layers.[8]

5.2. Corpus annotation

We argued earlier that identifying event factuality requires linguistic processing at different layers. First, it involves the interaction of local and non-local context. Second, it puts into play at least one, but generally more, relevant sources for each event, which bear a nesting relation among them. Hence, if not structured adequately, the annotation task could become too complex and would inevitable result in a questionable outcome. Annotating event factuality needs to be addressed by steps that could both help annotators to mentally structure and comprehend the different information layers involved, as well as allow us to partially automate certain parts of the annotation process. We divide the annotation effort into three consecutive tasks.

5.2.1. Task 1: Identifying Source-Introducing Predicates (SIPs)

Given a text with the events already recognized and marked as such, the annotators identified those that correspond to Source-Introducing Predicates. SIPs were briefly described in section 3.2. as including predicates of reporting, knowledge and opinion, among others. They are the linguistic elements that contribute a new source to the discourse. Such new sources, which must be nested relative to any previous relevant source, will have a role in assessing the factuality of the SIP event complement –recall example (25).

This initial task allowed annotators to get familiarized with both the notion of source and the notion of SIP as marker of factuality information. Moreover, for processing purposes Saurí & Pustejovsky (2007) show that identifying SIPs is fundamental for the automatic computation of relevant sources. The manual annotation resulting from this task was then used to prepare the final task.

5.2.2. Task 2: Identifying sources

The annotator was provided with a text with the following information already annotated: (a) all the SIPs in the text –obtained from the previous task; and (b) for each of these SIPs, a set of elements that can potentially express the new source it introduces; that is, a set of new source candidates. New source candidates had been automatically identified by selecting NP heads holding any of the syntactic functions listed here:[9]

1. Subject of any verbal predicate in the sentence.
2. Agent of a SIP in a passive construction (e.g., *The crime was reported by the* ***neighbor***.)[10]
3. Direct object of a SIP that has, as one of its arguments, a control clause headed by another SIP (e.g., *He criticized* ***Ed*** *for saying...*).
4. Complement of preposition *to* at the beginning of a sentence (e.g., *To* ***me****, she...*).
5. Complement of preposition *to* that is in a dependency relation with a SIP (e.g., *according to* ***me***, *it seems to* ***me***.)
6. Complement of preposition *of* that is in a dependency relation with a noun SIP (*the announcement of* ***Unisys*** *Corp.*).
7. Possessor in a genitive construction whose noun head is a SIP (e.g., ***Unisys*** *Corp.'s announcement*).

For every SIP, the annotator selected the new source it introduces among those in the candidate set. Two exceptional situations were also accounted for: (i) The new source did not correspond to any of the candidates in the list. The annotator would in these cases select option OTHER, and a posterior adjudication process would pick the adequate text item. (ii) There was no explicit segment in the text referring to the new source –for instance, in the case of generic sources (e.g., *it was expected/assumed that...*). The annotator would then select for option NONE. The new source is then interpreted as generic –i.e., it can be paraphrased as *everybody*. They will be represented as GEN in the resulting chain expressing the relevant source (e.g., GEN_*author*).

5.2.3. Task 3: Assigning factuality values

This final task was devoted to selecting the factuality value assigned to events by each of their relevant sources. The annotators were provided with a text where every event expression was paired with its relevant sources. Hence, sentences containing events with more than one relevant source were repeated several times, each presenting a different event-relevant source pair.

The set of relevant sources for each event had been automatically computed given the new sources manually identified in the previous task, and based on the algorithm for finding them presented in Saurí & Pustejovsky (2007).

The annotators had to choose among the set of factuality values presented in Table 4, which corresponds *grosso modo* to Table 1 with the addition of values PRu and PSu. In establishing the former table, these two values were estimated as non relevant, but we wanted to confirm they were also considered unnecessary by the annotators when looking at real data.

Two further values were allowed as well in order to pinpoint potential limitations in our value set: OTHER, covering situations where a different value would be required (e.g., the combinations U+ and U−), or when the annotator did not know what value to select; and NA (non-applicable), for events whose factuality cannot be evaluated.

To discern among the different factuality values, the annotators were asked to apply the discriminatory tests presented in section 2.3.

6. Evaluation

FactBank has been annotated by a pair of annotators. Overall, three annotators participated in the effort: annotators A and B participated in the first task, and annotators B and C carried out tasks 2 and 3. All of them are competent undergraduate Linguistics Majors. In addition, there were two

[8] FactBank annotation can be expressed by means of XML tags representing the factuality value assigned by a source to a given event. Because each event can be assigned more than one factuality value (as many as relevant sources it has), these must be non-consuming tags. Alternatively, given the correspondence between events IDs in both layers, the mapping can be established by means of stand-off markup as well.

[9] These syntactic functions were obtained from parsing the corpus with the Stanford Parser (de Marneffe et al., 2006).

[10] In this and coming examples, the new source candidate is marked in bold face and the SIP, underlined.

Table 4: Factuality values

VAL	USE
	Committed Values
CT+	According to the source, it is **certainly** the case that X.
PR+	According to the source, it is **probably** the case that X.
PS+	According to the source, it is **possibly** the case that X.
CT-	According to the source, it is **certainly not** the case that X.
PR-	According to the source it is **probably not** the case that X.
PS-	According to the source it is **possibly not** the case that X.
	(Partially) Uncommitted Values
CTu	The source knows whether it is the case that X or that not X.
PRu	The source knows whether it is probably the case that X or that not X.
PSu	The source knows whether it is possibly the case that X or that not X.
Uu	The source does not know what is the factual status of the event, or does not commit to it.
	Other Values
Other	Covering the following two situations - A different value is required here (e.g., U+, U-). - The annotator does not know what value to assign.
NA	The factuality nature of the eventuality cannot be evaluated.

adjudicators handling cases of disagreement in each task before annotators would continue with the next one.

Task 1. The interannotation ratio achieved is k=0.88 over 40% of the corpus (on the number of events).[11] Some of the most common cases of disagreement concern:

- SIP candidates with implicit sources –e.g., generic, as in: *He's **expected** to meet with Iraqi deputy prime minister Tariq Aziz later this afternoon.*
- SIP candidates lacking an explicit event complement (e.g., *The executives didn't **disclose** the size of the expected gain.*).
- Negated SIP candidates (e.g., *didn't* **disclose***, did not* **tell**, in the examples above).

Task 2. The interannotation agreement achieved for this task is k=0.95 over 40% of the corpus (on the number of events). Such good results come as no surprise since it is a very well-defined task, both in syntactic and semantic terms –essentially, it requires identifying SIP logical subjects. The most common cases of disagreements are those in which:

- There is a second expression in the text correfering with the new source. For example, the first person pronoun in a quoted fragment (e.g., *"We are going to maintain our forces in the region for the foreseeable future,"* **said** *spokesman Kenneth* **Bacon**.)[12]

  Another common situation was given with relative clauses (e.g., *British police* **officers** *who had been searching for Howes* **concluded** *that ...*).

- The new source introduced by the SIP referred to a non-human entity (e.g., **Reports** *attributed to the Japanese foreign ministry **said** ...*). One of the annotators would choose a different option.

Task 3. Interannotion agreement for this last task scores at k=0.82 over the 30% of the corpus (in terms of number of events). We consider this a very acceptable result, given the complexity of the task. In a comparable work devoted to classify certainty in text according to a five-fold categorization (*absolute, high, moderate, low*, and *uncertain*) (Rubin, 2007), the interannotation score obtained was k=0.15, which improved to k=0.41 when stricter annotation instructions were provided.

Furthermore, an analysis of disagreement cases on the 10% of our corpus shows that around two thirds of them are cases of true ambiguity, originated from different constructions. Some of the most common concerned the scope of a reporting predicate –or, in other words, the span of the attributed fragment. In (26), for example, the reporting predicate (in bold face) can be interpreted as scoping over both events *want* and *traveled*, or only only over *traveled*.

(26) Authorities want to question the unidentified woman who allegedly traveled with Kopp, **according** to an investigator quoted by the newspaper.

A second common case of ambiguity is caused by syntactic constructions typically triggering a presupposition (e.g., relative clauses, temporal clauses, appositions) when embedded under a reporting predicate (27). Annotators would disagree on whether the presupposition would be projected to the main clause –in our terms, the disagreement concerns whether the author of the text commits to the embedded event (underlined below) as a fact.

(27) The killing of Dr. Barnett Slepian, a gynecologist in Buffalo who performed abortions, has become a factor in at least two campaigns in New York, **say** political consultants and some campaign advisers.

7. Conclusions

Event factuality is an important component for representing events in discourse, but identifying it poses a two-fold challenge. First, factuality is in many cases the result of different factuality markers interacting among them. They can all be in the local context of the event, but it is also common for them to be at different levels. Second, the factuality value assigned to events in text must be relative to the relevant sources at play, which may be one or more.

In this paper, we introduced FactBank, a corpus of events annotated with factuality. FactBank contributes a semantic layer of factuality information on top of the grammar-based layer provided in TimeBank.

The interannotation agreement scores obtained for the three annotation tasks we designed are encouraging. Specifically, for the task of selecting the factuality value assigned to events by each of their relevant sources, we achieved k=0.82 over 30% of the corpus. That suggests that event factuality as modeled in our work is well-grounded in linguistic data, and that its identification is achievable using an approach along the lines of that proposed here. FactBank will be made available to the community in a near future.

[11] We apply Cohen *Kappa* (Cohen, 1960), hence assuming any potential distortion in the resulting figures due to the skewed distribution of categories (the so-called prevalence problem) as well as the degree to which the annotators disagree (the bias problem). Refer to Di Eugenio & Glass (2004).

[12] In this and the following examples, the SIP is presented in bold face and the new source to be selected in bold face and underlined. If an additional expression enters in consideration as new source candidate as well, it will only be underlined.

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*: 86–90.

Bergler, S. (1992). *Evidential Analysis or Reported Speech*. PhD thesis, Brandeis University.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). Building a discourse-tagged corpus in the framework of rhetorical structure theory.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *10*, 37–46.

de Marneffe, M.-C., MacCartney, B., Grenager, T., Cer, D., Rafferty, A., & Manning, C. D. (2006). Learning to distinguish valid textual entailments. In *Second PASCAL RTE Challenge (RTE-2)*.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Di Eugenio, B. & Glass, M. (2004). The kappa statistic: a second look. *Computational Linguistics*, *30*.

Halliday, M. A. K. & Matthiessen, C. M. (2004). *An introduction to functional grammar*. London: Hodder Arnold.

Hickl, A. & Bensley, J. (2007). A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*: 171–176.

Hooper, J. B. (1975). On assertive predicates. In J. Kimball (Ed.), *Syntax and semantics, IV*. New York: Academic Press: 91–124.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

Karttunen, L. (1970). Implicative verbs. *Language*, *47*, 340–358.

Karttunen, L. (1971). Some observations on factivity. *Papers in Linguistics*, *4*, 55–69.

Karttunen, L. & Zaenen, A. (2005). Veridicity. In Katz, G., Pustejovsky, J., & Schilder, F. (Eds.), *Dagstuhl Seminar Proceedings*, Schloss Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum (IBFI).

Kiparsky, P. & Kiparsky, C. (1970). Fact. In M. Bierwisch & K. E. Heidolph (Eds.), *Progress in Linguistics. A Collection of Papers*. The Hague: Mouton, 143–173.

Lewis, D. (1968). Counterpart theory and quantified modal logic. *Journal of Philosophy*, *65*, 113–126.

Light, M., Qiu, X. Y., & Srinivasan, P. (2004). The language of Bioscience: Facts, speculations, and statements in between. In *BioLINK 2004: Linking Biological Literature, Ontologies, and Databases*: 17–24.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004). The Penn Discourse Treebank. In *Proceedings of LREC 2004*.

Palmer, F. R. (1986). *Mood and Modality*. Cambridge, England: Cambridge University Press.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1).

Pradhan, S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007). OntoNotes: A unified relational semantic representation. In *Proceedings of IEEE International Conference on Semantic Computing*.

Prasad, R., Dinesh, N., Lee, A., Joshi, A., & Webber, B. (2007). Attribution and its annotation in the Penn Discourse Treebank. *Traitement Automatique des Langues*, *47*(2).

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., & Lazo, M. (2003). The TimeBank corpus. In *Proceedings of Corpus Linguistics 2003*, (pp. 647–656).

Pustejovsky, J., Knippen, B., Littman, J., & Saurí, R. (2005). Temporal and event information in natural language text. *Language Resources and Evaluation*, *39*(2), 123–164.

Rubin, V. L. (2007). Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Proceedings of the NAACL-HLT 2007*.

Saurí, R. (2008). *A Factuality Profiler for Eventualities in Text*. PhD thesis, Brandeis University.

Saurí, R. & Pustejovsky, J. (2007). Determining modality and factuality for text entailment. In *Proceedings of 1st IEEE International Conference on Semantic Computing*.

Snow, R. & Vanderwende, L. (2006). Effectively using syntax for recognizing false entailment. In *HLT-NAACL 2006*.

Tatu, M. & Moldovan, D. (2005). A semantic approach to recognizing textual entailment. In *Proceedings of HLT/EMNLP*: 371–378.

Verhagen, M., Stubbs, A., & Pustejovsky, J. (2007). Combining independent syntactic and semantic annotation schemes. In *Proceedings of the Linguistic Annotation Workshop*.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2), 165–210.

Wolf, F. & Gibson, E. (2005). Representing discourse coherence: A corpus-based analysis. *Computational Linguistics*, *31*(2), 249–287.

An Extensible Compositional Semantics for Temporal Annotation

Harry Bunt, Chwhynny Overbeeke

Tilburg University, Department of Communication and Information Sciences
P.O. Box 90153, 5000 LE Tilburg, Netherlands,
harry.bunt@uvt.nl., info@chwhynny.nl

Abstract

In this paper we present an event-based formal semantics for temporal annotation, in particular for the ISO-TimeML annotation language under development in the International Organization for Standardization. This semantics has the form of a *compositional translation* into First-Order Logic (FOL) using terms that denote concepts in an extended OWL-Time. Given the fact that FOL has a compositional semantics, our ISO-TimeML semantics is compositional because its translation into FOL is compositional in the sense that the translation of the annotation of a text is a function of the translations of its subexpressions (where any well-formed subexpression can be translated independently of other subexpressions) and the structure of the annotation, as encoded in its linking tags. The approach presented here has been designed to be extensible to the semantic annotation of other than temporal information.

1. Introduction

Linguistic annotation, according to Ide & Romary (2004), is the process of adding linguistic information to language data, or that information itself. The primary aim of annotation is usually the identification of certain linguistic patterns, in order to support the investigation of linguistic phenomena illustrated by such patterns, in particular for applying machine learning algorithms. As such, syntactic annotation as well as morphosyntactic, prosodic and pragmatic annotation have been useful in the development of data-driven linguistic models and theories.

Semantic annotations are meant to capture some of the meaning in the annotated text. This is not only potentially useful for identifying certain linguistic semantic patterns, but the meaning that is captured by the annotation should also support the exploitation of that semantic information in language processing tasks. For instance, Pustejovsky et al. (2003) argue that their annotation language TimeML, designed to support the automatic recognition of temporal and event expressions in natural language text, should also support *"temporal and event-based reasoning in language and text, particularly as applied to information extraction and reasoning tasks"*. (See also Han & Lavie (2004).) Bunt & Romary (2002) argue that any adequate semantic annotation formalism should have a well-defined semantics. Existing approaches to semantic annotation, by contrast, tend to take the semantics of the annotations for granted.

A current development in the area of semantic annotation is the design of an international standard for the annotation of temporal information, undertaken in the project "Semantic Annotation Framework, Part 1: Time and Events", which is carried out by an expert group within the International Organisation for Standardisation ISO. The annotation language that is defined in this project is based on TimeML and is therefore called ISO-TimeML. This project includes an effort to provide a formal semantics for the annotation language based on Pratt-Hartman's proposal of a formal semantics for TimeML (Pratt-Hartman, 2007) using Interval Temporal Logic, a first-order logic for reasoning about time. In this framework, the annotations are interpreted as statements about time intervals associated with events; events are not represented explicitly. While representing a substantial step forward, this semantics, described in the ISO (2007) document, has certain important limitations:

1. it applies only to a rather limited fragment of the annotation language, not including for instance tense, aspect, and durations;
2. it is not compositional, in the sense that it involves a translation from ISO-TImeML to ITL in such a way that the translation of a subexpression of an annotation structure is dependent on that of other subexpressions;
3. it is applicable to temporal information only, and not extensible to other kinds of semantic information, such as the identification of the participants in the events of which the temporal properties are considered.

In this paper we present an alternative, event-based formal semantics for ISO-TImeML, which applies to a substantially greater part of the annotation language, which is fully compositional, and which is not limited to dealing with temporal information. This approach follows the familiar 'interpretation-by-translation' paradigm, translating ISO-TimeML annotations, as represented in XML, into First-Order Logic (FOL). The compositionality of the approach rests on making this translation compositional.

In discussing this approach we will follow the TimeML terminology and speak of *'events'* in the generalized sense for which Bach (1981) introduced the term *eventualities*, as covering both states and events, where events may be subcategorized in various ways, for instance in processes and transitions.

This paper is organized as follows. In section 2 we briefly look at temporal information from a (onto-)logical and a linguistic point of view, and at the role that temporal annotation has to play. In section 3 we describe the translation of ISO-TimeML tags into formal representations. In section 4

we discuss the problem of making a formal semantics for XML-based annotations compositional, and present our solution to the problem. We end with concluding remarks in section 5.

2. Temporal Information

From a (onto-)logical point of view, the fundamental concepts relating to time are *time point*; the *ordering* relation between time points (*'before'*); *temporal interval*; the *begin* and *end* points of an interval; the relation *'inside'* between points of time and temporal intervals; and the *length* of a temporal interval, which requires the notion of a *temporal unit* of measurement. The general framework of Allen (1984), which has been very influential in the computational modelling of time, distinguishes 7 relations (and their inverses) between temporal intervals: *equals, before/after, meets, overlaps, starts, finishes, during/ contains*. These relations can all be defined in terms of the *before* relation among time points and the begin- and end points of intervals. In our FOL translations of ISO-TimeML annotations we will use polymorphic versions of Allen's relations, applying them both to time points and temporal intervals where appropriate. (For instance, we will use a predicate 'Before' which can apply to two temporal intervals, to two time points, to a time point and a temporal interval, or to a temporal interval and a time point, with the obvious interpretations.)

From a linguistic point of view, the issue is in what way these temporal objects and relations are described by linguistic expressions, and how language relates temporal objects to other concepts; in particular to states and events.

Temporal annotation, when endowed with a formal semantics, can be viewed as a bridge between the linguistic encoding of temporal information and the logical modeling of temporal structures and relations. For the formal semantics of ISO-TimeML (ISO, 2007), we will use an extension of the OWL-Time ontology (Hobbs & Pan, 2004). To the basic concepts of OWL (*interval, instant, beginning, end, inside, time zone*) we add the concepts of *temporal unit* and *duration*; and concepts needed for interpreting tense: *event time, speech time, and reference time*.[1]

2.1. Dates, Times and Periods

To represent dates, ISO-TimeML follows ISO standard 8601 and uses the format `yyyy-mm-dd` to encode year, month and day. This representation is unsatisfactory from a logical point of view, as it does not make the components of this information available for reasoning. For specifying a point in time we will use functions like *calYear, calMonth, calDay*, and *clockTime* (which specifies a time as shown on the clock in a given time zone):

(1) *March 16th 2007 at 10:15 a.m. CET*

λt : INSTANT(t) $\wedge$ calYear(t,cet) = 2007 $\wedge$ calMonth(t,cet) = march $\wedge$ calDay(t,cet) = 16 $\wedge$ clockTime(t,cet) = 10:15

It is rather unusual to be as explicit about a time zone as in (1); the time zone in which a clock time is considered is usually assumed to be obvious from the context in which the text fragment occurs that mentions the time. We will use the constant z_c to indicate the contextually relevant time zone in which a clock time is intended.

We use predicates like DAY and MONTH to represent intervals such as days, weeks, months, and years. The predicate DAY, for instance, is true of an interval starting at twelve midnight in some time zone, ending 24 hours later.

Again using ISO standard 8601, ISO-TimeML represents weekdays according to the format `xxxx-wxx-d`, where *d* is the number of the weekday. Thus, Monday would be `xxxx-wxx-1`, and Friday would be `xxxx-wxx-5`. We will use predicates of the weekdays and Allen's relations between temporal intervals to interpret the ISO-TimeML annotation of such expressions:

(2) (a) *Friday*
λt . $\exists$T : FRIDAY(T) $\wedge$ Inside(t,T)

(b) *every Friday*
λP . $\forall$T : FRIDAY(T) $\rightarrow$ P(T)

(c) *each year in March*
λP . $\forall T_1$: (YEAR(T_1) $\wedge$ $\exists T_2$: calMonth(T_2,z_c) = march $\wedge$ Before(Start(T_1),Start(T_2)) $\wedge$ Before(EndT_2),End(T_1)) $\rightarrow$ P(T_1)

We will use the constant *today* to refer to an interval that is a day inside which lies the speech time: *today* $\Leftrightarrow$ DAY(T) $\wedge$ Inside(T_0,T):

(3) (a) *yesterday*
DAY(T) $\wedge$ END(T) = START(today)

(b) *the day before yesterday*
DAY(T_1) $\wedge$ START(today) = END(T_1) $\wedge$ DAY(T_2) $\wedge$ END(T_2) = START(T_1)

2.2. Durations

To define durations we introduce the function *TimeAmount*, which constructs an amount of time from a numerical specification and a temporal unit, as illustrated in (4a).

(4) *for 2 hours*
λT : DURATION(T) = TimeAmount(2,hour)

A conversion function which specifies a numerical relation between temporal units, such as Conversion(hour, minute) = 60 explain equivalences like TimeAmount(1,day) = TimeAmount(24,hour) (see further Bunt (1985) for a calculus of amounts).

2.3. Tense and Aspect

Following Reichenbach (1947), we analyse tenses in terms of *event time, speech time*, and *reference time* (ET, T_0, and RT in the formal representations). ISO-TimeML uses PAST, PRESENT, and FUTURE as values of the `tense` attribute. If

[1] Hobbs & Pan (2004) use the term 'duration' to indicate a time span during which an event or state occurs. This is to be distinguished from our use of the term as indicating the length of a time span.

an utterance applies to an event in the past, the event time lies before the speech time; if it applies to an event in the present, the speech time is contained in the event time; if it applies to an event in the future, its event time is after the speech time. We can therefore conclude that:

PAST(e)	$\Leftrightarrow$	Before(ET(e),T_0)
PRESENT(e)	$\Leftrightarrow$	Inside(T_0,ET(e))
FUTURE(e)	$\Leftrightarrow$	Before(T_0,ET(e))

Some examples::

(5) (a) *Igor coughed.*
$\exists e\ \exists x$: SLEEP(e) $\wedge$ AGENT(x,e) $\wedge$ IGOR(x) $\wedge$ Before(ET(e),T_0)

(b) *Igor coughs.*
$\exists e\ \exists x$: SLEEP(e) $\wedge$ AGENT(x,e) $\wedge$ IGOR(x) $\wedge$ Inside(T_0,ET(e))

(c) *Igor will cough.*
$\exists e\ \exists x$: SLEEP(e) $\wedge$ AGENT(x,e) $\wedge$ IGOR(x) $\wedge$ Before(T_0,ET(e))

Note that in these examples we consider a literal interpretation of tenses, treating tense as an indicator of the temporal ordering relation between event time, speech time and reference time. Tense information should not always be taken literally, however. For instance, in (6) the event time lies after the speech time, in spite of the present tense of the verb:

(6) *I am at the office tomorrow.*

The temporal adverb *tomorrow* determines this, even though the present tense of the verb would suggest that the event time includes the speech time. This is a complication for any semantic interpretation of temporal annotation. One way to handle this problem could perhaps be to assign a different value to the `tense` attribute in such cases when annotating the text (e.g., Lee (2008) uses *'future present'*), but this has the drawback altering the linguistic concept of tense. Similar problems may arise with the interpretation of other syntactic attributes like `gender` and `number`.

The *progressive aspect* indicates that an event is occurring over a certain period of time and has not yet ended. That is, the speech time lies between the starting point and the end point of the event time.
Similarly, the *perfective aspect* indicates that an event has been ended, or refers to a state resulting from an event that has occurred:

(7) *Igor had already slept.*
$\exists e\ \exists x$: SLEEP(e) $\wedge$ AGENT(x,e) $\wedge$ IGOR(x) $\wedge$ Before(ET(e),RT) $\wedge$ Before(RT,T_0)

2.4. Temporal Anchoring

The Reichenbach (1947) notion of 'event time', originally introduced to interpret tenses, can obviously also be used for describing the temporal anchoring of an event to a time point or a temporal interval:

(8) *Igor died between 10 and 11 AM.*
$\exists e\ \exists x\ \exists T\ \exists t_1\ \exists t_2$: DIE(e) $\wedge$ PIVOT(x,e) $\wedge$ IGOR(x) $\wedge$ Interval(t_1,t_2) = T $\wedge$ clockTime(t_1,z_c) = 10:00 $\wedge$ clockTime(t_2,z_c) = 11:00 $\wedge$ Inside(ET(e),T) $\wedge$ Before(T,T_0)

ISO-TimeML also supports the temporal anchoring of an event with a specification of frequency, which may involve several temporal elements, such as *two hours a day* and *three days every month*. The ISO-TimeML annotation of such cases and our formal representations are as follows:

(9) `<TIMEX3 tid="t1" type="SET" value="P1M" quant="EVERY" freq="3D">`
three days every month `</TIMEX3>`
λ P. $\forall T_1$: MONTH(T_1) $\wedge$ $\exists^{!3} T_2$: DAY(T_2) $\wedge$ Inside(T_2,T_1) $\wedge$ P(T_1)

2.5. Relations between events

ISO-TimeML distinguishes three types of relation linking events to temporal elements or other events.

First, TLINK relates two temporal elements to one another, temporal elements to events, or eventualities to events, like for instance *20 minutes* to *every Friday* and *every Friday* to RUN in *Igor runs 20 minutes every Friday*, and LEAVE to ARRIVE in *Amy left before Igor arived.*

Second, SLINK is a subordination link between events for cases like *Igor wants to run* and *Amy believes that Igor loves her*. There are six types of SLINK relations: modal (e.g. PROMISE, WANT), evidential (e.g. SEE), negative evidential (e.g. DENY), factive (e.g. REGRET), counter-factive (e.g. PREVENT), and conditional (e.g. *if*). SLINK is not a temporal relation, and its interpretation is thus outside the scope of this paper (but see Bunt, 2007).

Third, ALINK indicates an aspectual relation between two eventualities: initiation, culmination, termination, continuation, or re-initiation, as exemplified by *Igor started to run*. These relations are more than just temporal relations. They can be viewed as a thematic relation (notably a THEME relation) plus certain specific properties. In the case of initiation, the specific property is that the starting point of the initiating event equals the starting point of the initiated event. Culmination means that the subordinate event has been completed, whereas termination implies that the subordinate event has not been completed.

3. From Annotations to Formal Representations

We follow the "interpretation through translation" approach for interpreting ISO-TimeML annotations, and formulate a compositional translation from the XML representations of ISO-TimeML annotations into formulas of First-Order Logic. The translation is defined by a set of rules for translating ISO-TimeML subexpressions and a set of operations for combining these translations, ultimately leading to the construction of a formal representation of the annotated text.

We mentioned in the beginning of this paper that the proposed ISO-TimeML semantics in terms of Interval Temporal Logic (see Pratt-Hartman (2007) and the ISO (2007) document) is not fully compositional. In a nutshell, the problem of translating (XML-representations of) ISO-TimeML annotations into formulas of a logical language in a compositional way is the following.

Compositional translation means that every well-formed subexpression of the source language is translated into the target language independently of other subexpressions; these translations are subsequently combined in a way that is determined by the structure of the source expression as a whole, as encoded in the TLINK, ALINK and SLINK tags that link the various subexpressions. ISO-TimeML annotations contain two kinds of subexpressions: on the one hand the expressions corresponding to events and temporal objects (<EVENT .../EVENT> and <TIMEX3 .../TIMEX3> subexpressions) and on the other hand subexpressions that indicate temporal, aspectual, or subordinate relations (TLINK, ALINK, and SLINK expressions). The latter type of expressions contain attributes whose values are identifiers in the subexpressions denoting events or temporal objects, thereby 'linking' these subexpressions. Now when the various types of subexpressions are translated into logical formulas, this linking information is lost because the logical formulas do not have identifiers like the XML structures of the ISO-TimeML annotation. The following example illustrates the problem for the ITL-based semantics of ISO-TimeML provided in the ISO (2007) document.

(10) John
<EVENT eiid="ei" type="OCCURRENCE">
drove /EVENT>
to Boston on
<TIMEX3 tid="t1" >
Saturday TIMEX3>
<TLINK eventInstanceID="ei"
relatedToTime="t1"
relType="DURING">

The event tag is translated into $\exists\, I_{ei}$: $P_{ei}(I_{ei})$, which says that there is a temporal interval I_{ei} for which the predicate P_{ei} holds, i.e. for which it is true that John drove to Boston during that interval:

<EVENT eiid="ei" type="OCCURRENCE" >
drove /EVENT>
$\rightsquigarrow \exists\, I_{ei}$: $P_{ei}(I_{ei})$

The TLINK structure is subsequently translated in such a way that it takes this latter formula and conjoins it with a formula expressing that the interval I_{ei} is related to another interval I_{t1} (corresponding to Saturday) through the relation specified as the relType value in the TLINK expression:

<TLINK eventInstanceID="ei"
relatedToTime="t1" relType="DURING">
$\rightsquigarrow \exists\, I_{ei}$: $P_{ei}(I_{ei}) \wedge \exists\, I_{ei'}$: $\text{DURING}_r(I_{ei}, I_{ei'})$

Now note that this formula has not been constructed by indepently translating the TLINK structure into a formula which is combined with the formula that translates the event, but in fact the translation rule operating here says: *When translating a* TLINK *expression, find the* EVENT *expression that is identified by the value of the* eventInstanceID *attribute; take the translation of that structure, and build within the scope of the existential event quantifier of that formula a conjunction which adds the temporal relation encoded in the* TLINK *structure.*

Kiyong Lee (2007), in trying to provide an alternative semanticss for ISO-TimeML, struggled with the same problem, and adopted the solution that is described below. Katz' (2007) attempt to give a denotational semantic to ISO-TimeML also runs into scoping problems.

We present a solution to this problem and specify a fully compositional translation at the price of having to deal with more complex intermediate representational structures during the translation process. These intermediate representations are triples consisting of a FOL formula plus two components, that we call a *'combination index'* and a *'derivation index'*. The first of these is a list containing the ISO-TimeML identifiers of the subexpressions whose translations are to be combined with the present representation; the second is another list of ISO-TimeML identifiers, indicating the subexpressions whose translations have been used to construct the present representation. As such, they act as a kind of storage which allows to keep track of (a) which pieces of semantic information should be combined, according to the links in the ISO-Timeml/XML representations, and (b) which pieces have already been combined. With the help of these devices, we can make sure that those and only those translations of the ISO-TimeML subexpressions which are linked through TLINK, SLINK or ALINK structures are combined, and in a correct way.

3.1. Translating ISO-TimeML Subexpressions

Here we will deal with the translation of each type of ISO-TimeML tag. (We will not take into account the SIGNAL tag of ISO-TimeML, which has been left out of consideration in this paper, since all it does is assign an index to a signal word such that it can be referred to in other tags.)

3.1.1. The EVENT Tag

The translation of event tags is determined by their polarity. There are two translation rules, one for each polarity value. The notation $\exists e{\in}E$ is used here and throughout as a shorthand for $\exists e$: E(e).

<EVENT eiid="e" tense=T aspect=A
polarity="POS">
$\rightsquigarrow \lambda E\,.\,\lambda P\,.\,\exists e \in E : P(e) \wedge T'(e) \wedge A'(e)$

<EVENT eiid="e" tense=T aspect=A
polarity="NEG">
$\rightsquigarrow \lambda E\,.\,\lambda P\,.\,\neg\exists e \in E : P(e) \wedge T'(e) \wedge A'(e)$

The translations of the time and aspect values are given in Table 1. and Table 2, respectively.

tense value	Translation
tense="PAST"	λe . Before(ET(e),T_0)
tense="PRESENT"	λe . Inside(T_0,ET(e))
tense="FUTURE"	λe . Before(T_0,ET(e))

Table 1: Translation table for the EVENT tag attribute tense.

aspect value/ Translation
aspect="PROGRESSIVE" λe . Before(START(e),T_0) ∧ Before(T_0,END(e))
aspect="PERFECTIVE" λe . Before(END(e),RT)
aspect="PERFECTIVE_PROGRESSIVE" λe . Before(START(e),T_0) ∧ Before(T_0,END(e)) ∧ Before(END(e),RT)

Table 2: Translation table for the EVENT tag attribute aspect.

3.1.2. The TIMEX3 Tag

ISO-TimeML uses an adapted form of the TIDES 2002 standard (Ferro et al., 2002), called TIMEX3, for marking up descriptions of time points and intervals. In natural language, events are often temporally anchored to an underspecified moment or period. The temporal anchoring of events can be represented in such cases with the (polymorphic) Inside relation (where T2 stands for the underspecified moment or period):

<TIMEX3 tid="t2" type=TYPE value=VALUE
temporalFunction="TRUE" anchorTimeID="t1">
$\rightsquigarrow$ λP . λt_1 . ∃T_2 : Inside(t_1,T_2) ∧ P(T_2)

The translation of TIMEX3 tags with specified starting points and end points is quite straightforward:

<TIMEX3 tid="t1" type=TYPE value=VALUE
beginPoint="t2" end="t3">
$\rightsquigarrow$ λP . λt_2 . λt_3 . ∃T_1 : START(T_1) = t_2 ∧ END(T_1) = t_3 ∧ P(T_1)

3.1.3. The TLINK Tag

A TLINK tag, used to anchor an event in time, is structured in ISO-TimeML as follows:

(11) <TLINK eventInstanceID=e1 signalID=s1
related ToTime=t1 relType=R />

Here, the attribute relType has values corresponding to the use of temporal prepositions such as *at, before, in, during*; these values correspond to temporal relations in the underlying temporal ontology. The translation of such a TLINK tag has the following form:

λ e. λ t. R'(ET(e),t)

where R' is the translation of the *relType* value. Table 3 exemplifies the translation of these values. 'Before' is the polymorphic temporal ordering relation between instants and intervals.

In its other main use in ISO-TimeML, to represent a temporal relation between two events, a TLINK tag is translated as:

λ e_1. λ e_2. R'(e_1, e_2)

where e_1 and e_2 correspond to the two related events and R' translates the value of the relType attribute (which has values like *when, while, after*).

relType value	Translation
BEFORE	λx . λy . Before(x,y)
AFTER	λx . λy . Before(y,x)
AT	*lambda*x. λy. x=y
INCLUDES	λT . λe . Before(START(T),START(e)) ∧ ∧ Before(END(e),END(T))
IS_INCLUDED	λT . λe . Before(START(e),START(T)) ∧ ∧ Before(END(T),END(e))
DURING	λe_1 . λe_2 . Before(START(e_2),START(e_1)) ∧ ∧ Before(END(e_1),END(e_2))

Table 3: Translation table for some relType values of the TLINK tag.

3.1.4. The ALINK Tag

The different possible aspectual relations that can be marked up in an ALINK tag are encoded in the values of its relType attribute. Since an aspectual relation always seems to correspond to a thematic relation plus a temporal relation, we translate all ALINK tags to a formal representation of the form:

λe_1.λe_2. THEME(e_1,e_2) ∧ τ

where τ is the temporal component that depends on the value of the relType attribute. Table 4 specifies the translations of the various relType values.

relType value	Translation component
INITIATES	ET(e_1) = START(e_2)
TERMINTATES	ET(e_1) = END(e_2) ∧ ¬COMPLETED(e_2)
CULMINATES	ET(e_1) = END(e_2) ∧ COMPLETED(e_2)
CONTINUES	Before(START(e_2),ET(e_1)) ∧ ∧ Before(ET(e_1),END(e_2))

Table 4: Translation table for the ALINK tag.

4. Combining Translations

In order to compositionally translate an entire ISO-TimeML annotation into FOL, we need to combine the translations of its subexpressions. This poses a problem, as he following example shows.

(12) Igor arrived at 11 AM.
Igor
<EVENT eiid="e1" tense="PAST"
polarity="POS">
arrived </EVENT>
<SIGNAL sid="s1">
at </SIGNAL>

```
<TIMEX3 tid="t1" type="TIME"
value="T11:00">
11 AM </TIMEX3>
<TLINK eventInstanceID="e1"
signalID="s1" relatedToTime="t1"
reltype="BEFORE" />
```

The respective translations of the event tag and the TLINK tag are as follows (where z_c as before indicates the contextually relevant time zone for the clock time):

λP. $\exists e_1 \in$ARRIVE: $\exists t_1$:clockTime(t_1,z_c)=11:00 $\wedge$
$\wedge$ Before(ET(e_1),t_1) $\wedge$ P(e_1)
λe_1. λt_1. ET(e_1) = t_1

We would like to combine these representations, and in this case that's quite simple. However, the simplicity of the example is deceptive. When we consider a more complex example, such as *Amy was happy when Igor arrived before 11 AM*, then we get two translations of event tags and we must make sure that the translation of the TLINK tag is combined with that of the ARRIVE event, not with that of the REJOICE event. This is an instance of the problem of defining a compositional translation, pointed out above. Here, the problem is that the translations of the event- and TLINK tags have lost the linking information captured in the XML tags by the values of the `eventInstance` and `relatedToTime` attributes; the use of the same variables e_1 and t_1 in the translations of the tags only optically preserves the linking information; formally the names of these variables are insignificant.

We resolve this problem by keeping track of the linking information in the annotations and reformulating all translations as using intermediate representations in the form of triples

$$< ci, di, \varphi >$$

where *ci* (the 'combination index') contains XML identifiers such as the values of the `eventInstance` and `relatedToTime` attributes, for keeping track of the ISO-TimeML tags whose translations should be combined with the present representation, and where *di* (the 'derivation index') contains XML identifiers like the value of the `eiid` attribute in an event tag; this keeps track of which translations of ISO-TImeML subexpressions have already been used in the translation.

After translating the various tags in terms of such triples, the rest of the translation process consists of combining these triples, until a triple has been constructed whose combination index is empty and whose derivation index indicates that all the ISO-TimeML subexpressions have been linked together. For the combination of these triples we use a number of formal operations which are defined in the next subsection.

4.1. Combination operations

The operations that we use for combining the translations of ISO-TimeML subexpressions involve a few formula-manipulation operations defined in (Bunt, 2007). The most important one is a type of function application called *late unary application*, where a one-argument function is applied to an argument expression of the form $\lambda x_1,...,x_k$. E($x_1,...,x_k$). The definition of this operation, designated by '$\square$', is as follows:
F $\square$ $\lambda x_1,...,x_k$ λa . E = $\lambda x_1,...,x_k$ F(λa . E)

This operation and the others that we will describe below have to be extended to triples. In what follows, we will use the same symbols for the operations when applied to triples as when applied to formulas, except in the definitions where the subscript '3' is used to make clear that an operation is applied to triples. (We will use '$\cdot$' to indicate concatenation of lists, and' -' subtraction of lists.) For late unary application the triple-definition is:[2]

$<ci_1, li_1, \varphi_1> \square_3 <ci_2, li_2, \varphi_2> = <<ci_2\text{-}li_1>, <li_1 \cdot li_2 \cdot <ci_1>>, \lambda x_1,...,x_{k-1} . \varphi_1 (\lambda x_k \square \varphi_2)>$

Second, an operation called *lambda insertion-application* (designated by $\oplus$) is defined, which combines a lambda abstraction λa . F, where F is a function expression, with an expression of the form $\lambda x_1,...,x_k$. E_1 $\exists$z : E_2 into $\lambda x_1,...,x_k$. λa . E_1 $\exists$z : F(z) $\wedge$ E(z).

In terms of triples:[3]

$<ci_1, li_1, \varphi_1> \oplus_3 <<>, li_2, \varphi_2> = <ci_1\text{-}li_2, li_1 \cdot li_2 \cdot ci_1, \varphi_1 \oplus \varphi_2>$

A variant of this operation, designated by $\oplus$', swaps the order of its arguments in application, and is defined as follows, with its obvious extensions to triples:

$(\lambda x_1 . \lambda x_2 . F) \oplus' A = (\lambda x_2 . \lambda x_1 . F) \oplus A$

A third operation, called *cross-application* (designated by $\otimes$), merges two expressions of the form λv . $\exists$x : E_1(v,x) $\wedge$ E_2 and λw . $\exists$y : E_1(y,w) $\wedge$ E_3 into $\exists$x $\exists$y : E_1(y,x) $\wedge$ E_2 $\wedge$ E_3.
In terms of triples:
$<ci_1, li_1, \varphi_1> \otimes_3 <<>, k \cdot ci_1, \varphi_2> = <<>, k, \varphi_1 \otimes \varphi_2>$

Finally, an operation called *merge-application* (designated by $\odot$), is defined for any two representations E1 = $<ci_1, di_1, \alpha>$ and E2 = $<ci_2, di_2, \lambda z . \beta>$, where the set of first elements in the pairs constituting di_1 equals the set of identifiers in ci_1; β is not of the form λx..., and the length of the sequence of λ-abstractions in E2 equals the length of the list di_2. If α is a formula of the form γ Qz δ, where Q is a (generalized) quantifier, then the logical formula resulting from merge-application is γ Qz [λz . β](z) $\wedge$ δ.

In terms of triples: [4]

$<<>, li_1, \varphi_1> \odot_3 <ci_2, li_2, \varphi_2> = <ci_2\text{-}li_1, <li_1>, \varphi_1 \odot \varphi_2>$

These operations can be applied in any order to any triples that satisfy the properties required in the definitions of the operations, without any further constraints, thus ensuring the compositionality of the process. In the next subsection we will give some examples to illustrate the process.

4.2. Worked examples

(13) *Igor arrived at 11 AM.*

We considered the ISO-TimeML annotation of this example in the previous subsection (see (11)). We describe the translation step by step. The TIMEX3 tag and the TLINK tag:

[2] A condition on the applicability of the operation $\square_3$ is that the combination index $<ci_2$ of the second operand has the form $<ci_2' \cdot li_1$.

[3] A condition on the applicability of the operation $\oplus_3$ is that the combination index $<ci_1$ of the first operand has the form $<ci_{\_}' \cdot li_2$.

[4] See footnote 2.

T' = <<>, <t1>, λP. $\exists t_1$: clockTime(t_1,z_c)=11:00 $\wedge$ P(t_1)>
TL$_a$' = <<e1,t1>, <>, λa. λb. ET(a) = b>

Combination of the two translations using late unary application:

T' $\square$ TL$_a$' =
<<e1>, <t1>, λa. $\exists t_1$: clockTime(t_1,z_c)=11:00 $\wedge$ ET(a)=t_1>

Translation of the EVENT tag:

E' = <<>,<e1>,λQ. $\exists e_1 \in$ ARR: Before(ET(e_1),T_0) $\wedge$ Q(e_1)>

The EVENT translation is combined with that of the combination of the TIMEX3 tag and the TLINK tag using late unary application, which delivers the desired end result:

E' $\square$ (T' $\square$ TL$_a$') =
<<>, <t1,e1>, $\exists e_1 \in$ ARRIVE : $\exists t_1$: clockTime(t_1,z_c) = 11:00 $\wedge$ ET(e_1) = t_1 $\wedge$ Before(ET(e_1),T_0)>

Next we consider an example with two temporally ordered events:

(14) *Amy left before Igor arrived.*
Amy
<EVENT eiid="e1" tense="PAST"
polarity="POS">
left </EVENT>
<SIGNAL sid="s1">
before </SIGNAL>
Igor
<EVENT eiid="e2" tense="PAST"
polarity="POS">
arrived </EVENT>.
<TLINK eventInstanceID="e1"
signalID="s1"
relatedToEventInstance="e2"
reltype="BEFORE" />

The two EVENT tags:

E1' = <<>, <e1>, λQ. $\exists e_1 \in$ LEAVE: Before(ET(e_1),T_0) $\wedge$ Q(e_1)>
E2' = <<>, <e2>, λQ . $\exists e_2 \in$ ARRIVE: Before(ET(e_2),T_0) $\wedge$ Q(e_2)>

The TLINK tag:

TL$_e$' = <<e1,e2>, <>, λa . λb . Before(a,b)>

Combination of the translation of the second EVENT tag with that of the TLINK tag using late unary application:

E2' $\square$ TL$_e$' =
<<e1>, <e2>, λa . $\exists e_2 \in$ ARRIVE : Before(ET(e_2),T_0) $\wedge$ Before(a,e_2)>

Combination of the translation of the first EVENT tag (*Amy left*) with that of the second EVENT tag plus the TLINK tag (*before Igor arrived*) using late unary application, gives the desired end result:

E1' $\square$ (E2' $\square$ TL$_e$') =
<<>, <e1,e2>, $\exists e_1 \in$ LEAVE : Before(ET(e_1),T_0) $\wedge$ $\exists e_2 \in$ ARRIVE : Before(ET(e_2),T_0) $\wedge$ Before(e_1,e_2)>

We finally consider an example with three related events, two of which have an aspectual relation and two a temporal ordering relation.

(15) *Amy started to laugh when Igor arrived.*

Amy
<EVENT eiid="e1" tense="PAST"
polarity="POS">
started </EVENT>
to
<EVENT eiid="e2" tense="NONE"
vform="INFINITIVE" polarity="POS">
laugh </EVENT>
<SIGNAL sid="s1">
when </SIGNAL>
Igor
<EVENT eiid="e3" tense="PAST"
polarity="POS">
arrived </EVENT>.
<ALINK eventInstanceID="e1"
relatedToEventInstance="e2"
reltype="INITIATES" />
<TLINK eventInstanceID="e3"
signalID="s1" relatedToEventInstance="e1"
reltype="IDENTITY" />

The translation of *Amy started to laugh*:

E1' $\square$ (E2' $\square$ AL') =
<<>, <e1,e2>, $\exists e_1 \in$ START : Before(ET(e_1),T_0) $\wedge$ $\exists e_2 \in$ LAUGH : THEME(e_2,e_1) $\wedge$ ET(e_1) = START(e_2)>

The ARRIVE event tag:

E3' = <<>, <e3>, λQ . $\exists e_3 \in$ ARRIVE : Before(ET(e_3),T_0) $\wedge$ Q(e_3)>

The TLINK tag:

TL$_e$' = <<e1,e3>, <>, λa . λb . ET(a) = ET(b)>

Combination of the translation of the third EVENT tag with the that of the TLINK tag using late unary application:

E3' $\square$ TL$_e$' =
<<e1>, <e3>, λa . $\exists e_3 \in$ ARRIVE : Before(ET(e_3),T_0) $\wedge$ ET(a) = ET(e_3)>

Application of lambda-insertion applicaton with swapping of variables:

TL$_e$' $\oplus$' (E1' $\square$ (E2' $\square$ AL')) =
<<e3>, <e1,e2>, λb . $\exists e_1 \in$ START : Before(ET(e_1),T_0) $\wedge$ $\exists e_2 \in$ LAUGH : THEME(e_2,e_1) $\wedge$ ET(e_1) = START(e_2) $\wedge$ ET(e_1) = ET(b)>

Application of cross-application to this representation for *Amy started to laugh* and the translation of *when Igor arrived* gives the desired end result:

(E3' $\square$ TL$_e$') $\otimes$ (TL$_e$' $\oplus$' (E1' $\square$ (E2' $\square$ AL'))) =
<<>, <e1,e2,e3>, $\exists e_1 \in$ START : Before(ET(e_1),T_0) $\wedge$ $\exists e_2 \in$ LAUGH : THEME(e_2,e_1) $\wedge$ ET(e_1) = START(e_2) $\wedge$ $\exists e_3 \in$ ARRIVE: Before(ET(e_3),T_0) $\wedge$ ET(e_1) = ET(e_3)>

5. Discussion and Conclusions

The method described in this paper enables a larger part of ISO-TimeML to be formally interpreted than the ITL approach, including the interpretation of tense and aspect, the treatment of durations, and that of calendar years, clock times, and so on. A treatment of calendar years and the like in an ITL-based semantics would probably not be hard, adding predicates applicable to certain temporal intervals as we have have done here. It would be more difficult to extend would be difficult to extend the ITL-based semantics with the interpretation of tense and aspect, since tense interpretation for instance requires the representation of event times (as temporally related to speech times and reference times), which is a property of events and thus necessitates the availability of events as such. Even more difficult would be the addition of durations, since this requires new concepts (temporal units and amounts of time, defining equivalence classes of pairs of a temporal unit and a numerical value) to be added to the underlying ontology.

More important from a theoretical point of view, is that we have specified a fully compositional interpretation of ISO-TimeML. This has been achieved at the price of making use of more complex intermediate representations, but has, besides the obvious theoretical importance, the advantage of allowing a very flexible translation process, which consists of a number of operations that can be applied in any order.

The attempt to formally interpret ISO-TimeML annotations has also revealed interesting interferences with the annotation of other semantic information, such as semantic roles and quantification. As long as semantic annotation is restricted to temporal annotation only, it may be reasonable to annotate the relations between events for which ISO-TimeML uses SLINK structures in the temporal annotation language, but these relations are not really temporal in nature and would be better treated as semantic role relations which have certain temporal implications. Also, aspectual relations, as captured in ALINK tags, are by their very nature a combination of thematic and temporal relations. Temporal quantification does not have a fully satisfactory treatment in ISO-TimeML, and indeed this only seems possible by taking quantification into account more generally.

For ISO-TimeML interpretation only, it might be feasible to cast the formal semantics in terms of a description logic like OWL-DL; however this would restrict the extensibility of the approach. An important aspect of the ISO-TimeML semantics outlined in this paper is that it has a richer underlying ontology than Interval Temporal Logic, including events and nontemporal individuals, which makes it possible to extend the approach to the semantic annotation of other information related to events. This would notably include the roles that the participants in an event play ('semantic roles'), as well as other properties of such participants, such as referential relations among participants in different events, and aspects of quantification for dealing with cases where sets of participants are involved in sets of events. The possibilities in this direction are explored in (Bunt, 2007) and Bunt & Overbeeke (2008).

References

Allen, J. (1984). A General Model of Action and Time. *Artificial Intelligence*, 23-2.

Bach, E. (1981). On Time, Tense, and Aspect: An Essay in English Metaphysics. In Cole, P., editor, *Radical Pragmatics*. Academic Press, New York.

Bunt, H. (1985). *Mass Terms and Model-Theoretic Semantics*. Cambridge University Press.

Bunt, H. (2007). The Semantics of Semantic Annotation. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC21)*.

Bunt, H. and Overbeeke, C. (2008). Towards formal interpretation of semantic annotation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech. ELRA.

Bunt, H. and Romary, L. (2002). Towards Multimodal Content Representation. In Choi, K. S., editor, *Proceedings of LREC 2002, Workshop on International Standards of Terminology and Language Resources Management*, pages 54–60, Las Palmas, Spain. Paris: ELRA.

Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2002). *Instruction Manual for the Annotation of Temporal Expressions*. MITRE Washington C3 Center, McLean, Virginia.

Han, B. and Lavie, A. (2004). A Framework for Resolution of Time in Natural Language. *TALIP Special Issue on Spatial and Temporal Information Processing*, 3.

Hobbs, J. and Pan, F. (2004). An Ontology of Time for the Semantic Web. *TALIP Special Issue on Spatial and Temporal Information Processing*, 3-1:66–85.

Ide, N. and Romary, L. (2004). International Standard for a Linguistic Annotation Framework. *Natural Language Engineering*, 10:211–225.

ISO (2007). *Language Resource Management - Semantic Annotation Framework (SemAF) - Part 1: Time and Events*. Secretariat KATS. ISO Report ISO/TC37/SC4 N269 version 19 (ISO/WD 24617-1).

Katz, G. (2007). Towards a Denotatial Semantics for TimeML. In Schilder, F., Katz, G., and Pustejovsky, J., editors, *Annotation, Extraction, and Reasoning about Time and Events*. Springer, Dordrecht.

Lee, K. (2008). Against a Davidsonian Analysis of Copula Sentences. In Kadowaki, M. and Kawahara, S., editors, *NELS 33 Proceedings*.

Pratt-Hartmann, I. (2007). From TimeML to Interval Temporal Logic. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)*, pages 166–180, Tilburg, Netherlands.

Pustejovsky, J., Castano, J., Ingria, R., Gaizauskas, R., Katz, G., Saurí, R., and Setzer, A. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 337–353, Tilburg, Netherlands.

Pustejovsky, J., Knippen, R., Littman, J., and Saurí, R. (2007). Temporal and Event Information in Natural Language Text. In Bunt, H. and Muskens, R., editors, *Computing Meaning*, volume 3, pages 301–346. Springer.

Reichenbach, H. (1947). *Elements of Symbolic Logic*. Macmillan, New York.

Using Treebank, Dictionaries and GLARF to Improve NomBank Annotation

Adam Meyers

New York University
New York, NY
meyers@cs.nyu.edu

Abstract

In the field of corpus annotation, it is common to annotate text multiple times and then adjudicate the results. The resulting annotation is generally regarded as more consistent and more accurate than the results of a single pass. However, it is also very expensive to annotate in this way. Given text corpora that are annotated by many different research groups, another source of comparison is available: annotation of other linguistic information on the same corpora. By exploiting violations of expected relationships between the two annotation schemes, likely errors can be detected. This paper describes such an effort involving the NomBank annotation of noun arguments in the Wall Street Journal Corpus. These techniques made it possible to complete NomBank annotation efficiently and accurately.

1. Introduction

As with many annotation projects, NomBank took longer to finish than the creators initially expected. It eventually became necessary to find a way to complete the annotation in a way that minimized expenses, while maintaining high quality. In many projects involving the manual annotation of corpora with linguistic features, each text is annotated by two different annotators and the differences between their output are adjudicated. The resulting annotation is more consistent than singly annotated corpora and this increased consistency is usually assumed to indicate a corresponding improvement in accuracy. Due to practical constraints, this was not an option for NomBank.

Fortunately, the NomBank project was annotating a text corpus for which there was already previous annotation (in particular, Penn Treebank annotation). We established several expected relationships between the NomBank and the Penn Treebank annotation schemes. When any of these expected relationship did not hold, there were three possibilities: (1) there was an error in NomBank; (2) there was an error in the Penn Treebank; or (3) the expected relationship did not hold for this instance. Given these possibilities, annotation that violated an expected relationship was more likely to contain a NomBank error, than randomly selected annotation.

In addition, some parts of NomBank annotation had expected relationships with syntactic dictionaries, both ones created during the NomBank project (ADJADV, NOMLEX-PLUS) and existing ones (NOMLEX and COMLEX Syntax). By examining cases where these expected relationships were violated, we could predict likely NomBank (or dictionary) errors. As a result of these efforts, approximately 26% of NomBank manual annotation was predicted to contain likely errors and was examined and corrected by an expert annotator, a substantial savings in time and effort. Methods for evaluating the effectiveness of this effort are under consideration for future work.

NomBank annotators reviewed a total of 200,000 instances of nouns in the Penn Treebank corpus to produce 114,500 NomBank propositions. On average, they looked at about 20–25 noun instances per hour, working at a considerably slower pace than PropBank (Palmer et al., 2005)(less than one half the speed). This made double annotation of NomBank impractical. By comparing NomBank annotation to previous annotation, we were able to select approximately 30,000 propositions that were likely to contain errors and review those propositions in a focused way. This made for realistic and effective quality control.

We will now sketch an outline of the remainder of this paper. Section 2. provides an overview of NomBank annotation. Section 3. describes our approach to merging together various annotation schemes into a GLARF representation, which we use for our error detection system. Sections 4. through 7. describe the various constraints that we use to detect likely errors. Finally, Section 8. discusses ramifications and future research .

2. NomBank Annotation

NomBank.1.0 (Meyers et al., 2004a) provides a predicate argument structure representation of approximately 114,500 noun instances in the Wall Street Journal corpus. Like PropBank, this representation links particular word instances with words and phrases that are either arguments (ARG0, ARG1, ...) or belong to one of the classes of nonarguments (ARGMs) defined in the specifications. For each word, there is a dictionary entry (its frame file) which defines the set of possible arguments. The set of markable ARGMs are essentially those that have counterparts in verbal argument structure, e.g., temporal, locative, manner, etc.[1] In addition, we mark SUPPORT items, words that link arguments outside of the noun phrase to the nominal predicate. Some example sentences are provided below. The nominal predicate is underlined and the other parts of proposition are in bold. The labels following the arguments indicate the roles they play in the NomBank proposition. The set of support words in each of these examples forms a chain in that sentence connecting an argument outside the NP to the underlined predicate. For example the support chain, consisting of *gave* + *dozens* + *of*, links *John* to *kisses* – the chain should be viewed as filling a single SUPPORT slot in the NomBank proposition.

[1] See the NomBank manual, available from nlp.cs.nyu.edu/meyers/NomBank.html, for more information.

1. **Mary's**/*ARG0* **promise**/*ARG1-REF* **to John**/*ARG2*
2. **The Press's**/*ARG0* *criticism* **of the candidate**/*ARG1*
3. **John**/*ARG0* **gave**/*SUPPORT* **Mary**/*ARG2* **dozens of**/*SUPPORT* *kisses*
4. **They**/*ARG0* **accorded**/*SUPPORT* **minorities**/*ARG1* *an* **opportunity for**/*SUPPORT* *representation.*

Like PropBank, each word and phrase in NomBank is represented as a link to one or more nodes of Penn Treebank annotation (Marcus et al., 1994). This contrasts with most approaches to annotation such as: (a) inline annotation where the text is modified to include annotation features and (b) offset annotation which points to particular spans of text using another document (these text spans are usually referenced by byte offsets from the beginning of the target file). In this sense, NomBank is annotation of annotation, i.e., NomBank assigns features to units defined by pre-existing Penn Treebank annotation.

3. GLARFBANK

As part of the Unified Linguistic Annotation project (Pustejovsky et al., 2005), researchers at several United States universities are studying ways to merge together distinct annotation schemes. At New York University (NYU), we are taking an approach to merging that we call "aggressive" because we change incompatible aspects of the input annotation schemes so that they are compatible with each other, i.e., we change tokenization, phrase boundaries and text spans to maximize overlap between the input annotation schemes. In this respect, we are taking annotation created under different theoretical assumptions and converting them into a single-theory analysis. The output of the merging process is formalized as a Typed Feature Structure in the GLARF framework (Meyers et al., 2001a; Meyers et al., 2001b).[2]

The current GLARF'd version of the Wall Street Journal data annotated for NomBank includes the following annotation schemes: Penn Treebank, PropBank, NomBank, Penn Discourse Treebank (overt relations)(Miltsakaki et al., 2004) and BBN Named Entity tags. Future merged GLARFBANKs will also include Brandeis' TimeML (Pustejovsky et al., 2004) and University of Pittsburgh's Opinion annotation (Wilson and Wiebe, 2003). The WSJ GLARFBANK also includes various automatically generated features based on both heuristic rules and lexical lookup (COMLEX Syntax, NOMLEX, ADJADV, and others). GLARF rules correct parts of speech, mark focused constituents, fill gaps not covered by Treebank annotation, assign grammatical roles to constituents, add semantic features, etc.[3] A sample (simplified) GLARF representation is provided as Figure 1. It represents the merger of annotation for the sentence *Meanwhile, they made three bids*:[4]

```
(S
 (ADV (ADVP
       (HEAD (ADVX
              (HEAD (RB Meanwhile))
              (P-ARG1 (S (EC-TYPE PB)
                         (INDEX 0+0)))
              (P-ARG2 (S (EC-TYPE PB)
                         (INDEX 0)))))
       (INDEX 1)))
 (PUNCTUATION (, ,))
 (SBJ (NP (HEAD (PRP they)) (INDEX 2)))
 (PRD (VP
       (HEAD (VX
              (HEAD (VBN made))
              (P-ARG0 (NP (EC-TYPE PB)
                          (INDEX 2)))
              (P-ARG1 (NP (EC-TYPE PB)
                          (INDEX 4)))
              (P-ARGM-TMP (ADVP
                           (EC-TYPE PB)
                           (INDEX 1)))
              (INDEX 3)))
       (OBJ (NP (T-POS (CD three))
          (HEAD (NX
                 (HEAD (NNS bids))
                 (P-ARG0 (NP
                          (EC-TYPE PB)
                          (INDEX 2)))
                 (SUPPORT (VX
                           (EC-TYPE PB)
                           (INDEX 3)))))
          (INDEX 4)))))
 (PUNCTUATION (. .))
 (SENT-NUM 1)
 (INDEX 0))
```

Figure 1: GLARF for: *Meanwhile, they made three bids*

The GLARF representation[5] essentially adds structure to the Penn Treebank and if you delete this additional structure, the result would be the original Penn Treebank (with minor changes). We will highlight two of these elaborations

[2]Currently several applications are using GLARF'd data for Information Extraction including the systems described in (Zhao et al., 2004; Shinyama and Sekine, 2006) as well as NYU's recent Automatic Content Extraction (ACE) submissions. We have also begun a Machine Translation effort at NYU that uses Chinese, Japanese and English GLARF.

[3]We intend to make a GLARF representation of the ULA shared corpus available at nlp.cs.nyu.edu/wiki/corpuswg/ULA-OANC-1. Prior to the availability of hand annotation, automatically generated features are provided for PropBank, NomBank and the Penn Discourse Treebank. The author intends to make the Wall Street Journal GLARFBANK available either through the Linguistic Data Consortium, or by download should licensing restrictions on this corpus be relaxed.

[4]In the GLARF system the typed feature structure includes all the information in GLARF. A multi-level dependency representation is also available that is similar to the 2008 CONLL task representation (www.yr-bcn.es/conll2008/). In fact the latter is partially derived from the former.

[5]There are actually several different GLARF representations. The typed feature structure representation contains the most information and a dependency representation is the one that is most often used for Information Extraction and other applications.

here: (1) relational labels like HEAD, ADV, PRD, OBJ, that indicate relations between constituents, e.g., the constituent labeled SBJ is the subject of the sister constituent that is labeled PRD; and (2) Empty Categories that may or may not be part of the original Penn Treebank, e.g., the features prefixed with P- point to empty categories which bear PropBank, NomBank and PDTB relations with the HEAD constituent. These empty categories point to other GLARF constituents, e.g., the the NP *they* has an INDEX feature value of 2. The empty categories that are values of the P-ARG0 of *made* and *bids* both also have this index, representing that *they* is the PropBank ARG0 of *made* and the NomBank ARG0 of *bids*. The P-ARG2 of *Meanwhile* has a value of the entire sentence, which would appear to include itself. However, by convention, we assume that such arguments exclude what we call the *SELF-PHRASE*, the ancestor of the predicate (in this case *Meanwhile*) that is a child of the argument. This same rule is used for marking arguments of parenthetical predicates in PropBank and NomBank. Thus in the following two examples, the entire sentence can be marked as an argument of *claimed* and *request* because the self-phrases *Mary claimed* and *at John's request* can easily be accounted for: *Irving, Mary claimed, is ten feet tall*, *Mary, at John's request, made ridiculous claims about Irving.* The P-ARG1 of *Meanwhile* refers to the previous sentence (sentence 0, index 0).

Our system checks new NomBank data for its compatibility with other annotation frameworks, using the GLARF-BANK annotation as a way of incorporating the other annotation into a single representation. Following sections describe these compatibility tests and the subsequent adjudication.[6]

4. Structural Constraints on Internal Arguments of Nouns

We use the GLARF representation as a means to implement several types of constraints. First of all, by recognizing particular kinds of constituents, we can constrain how they appear in NomBank. Relative clauses typically are not markable in NomBank propositions. Thus, given a NomBank Proposition for a noun N, if one of the arguments (ARG0 ... ARG9) or ARGMs is a relative clause, this is flagged as a likely error, e.g., the *that* relative in *the banner that proclaims the renewal of socialism* was detected as a likely error and then removed during adjudication. It is easy to identify relative clause arguments because relative clauses are labeled as such in the GLARF'd version of the Penn Treebank. The GLARF-generating program uses a combination of the representation in the original Penn Treebank (the appearance of empty categories in that-clauses following nouns, the POS markings on *that*, etc.) and whether or not a *that*-clause is a possible complement for the head noun (using COMLEX Syntax) to determine if a structure is a relative clause (if a *that* phrase follows a noun that can't take *that* complements, the phrase is likely to be a relative clause).

NomBank annotators have the option of linking together constituents in the Penn Treebank to form a single NomBank argument. These combinations often correctly identify constituents not marked in the Penn Treebank, due to (for example) Penn Treebank's tendency to underspecify prenominal structure, e.g., in a phrase like *The ice cream man*, *ice* and *cream* would probably be left as separate constituents. However, it turns out that some constituent combinations are unlikely to be correct. For example, given D and N two adjacent prenominal modifiers of some head H, if D is a determiner or possessive and N is a noun or adjective, it is unlikely that D and N form a constituent. For example, one annotator marked *their financial* as a single constituent (an ARG1) of the predicate *viability* in the phrase *their financial viability*. In the corrected version, *their* is marked as an ARG3 and *financial* is marked as an ARG1. The reason for this error is clear. ARG3 and ARG1 are similar roles for nouns like *viability* which belong to the ATTRIBUTE class and the annotator opted to combine the two rather than mark them separately. The ARG1/ARG3 split in NomBank reflects that *viability* is an attribute of the *financialness* and *financial viability* is an attribute of *them*. In this case, the ARG3 is a secondary-theme a type of argument that has this interpretation (as per the NomBank manual). *Their financial viability* is a phrase that represents the degree or VALUE of the *viability* trait and therefore *viability* is marked as its own ARG2. This error detection routine occasionally identifies non-errors. For example, the GLARF generating program incorrectly marked the numeral 1 as a determiner in the sentence *CBS held the previous record for consecutive No. 1 victories*. The annotator had correctly marked *No. 1* as a single ARG1 – so this annotation was not changed during adjudication.

In a similar vein, annotations of discontinuous constituents are unlikely to be correct. Any series of constituents that form a NomBank argument are almost always consecutive. Nevertheless, NomBank annotators will occasionally mark discontinuous constituents, the most common reasons being: (1) one token is missed from a sequence, e.g., the comma was not included as part of the ARG1 *stock, bond and foreign exchange* in the initial marking of the phrase *its stock , bond and foreign exchange trading*; and (2) as in the determiner plus prenominal case above, the two arguments have similar relations to the head noun. For example, although one annotator marked a combination of *conversion* and *on the stock* as a single ARG1 of *rights* in the phrase *conversion rights on the stock*, the final version of NomBank makes *conversion* an ARG1 and *on the stock* an ARG3. The one consistent exception, discussed in Section 6., is where the entire sentence or NP is an argument of the noun (minus the self-phrase containing the nominal predicate). For example, in *Mr. Nadeau said discussions are under way with potential purchasers of each of the units*, the entire phrase minus *under way* is an ARG1. Apart from these carefully defined exceptions, there are also 10 cases involving the noun predicate *age*, where marking discontinuous constituents seemed unavoidable even though the examples did not fit into one of cases of external argu-

[6] Tests for compatibility between the structure of the GLARF-BANK and NomBank are mostly tests for compatibility between the Penn Treebank and NomBank. However, the GLARFBANK actually incorporates structures from other annotation. So the relation is not one to one.

ments of nouns, e.g., we marked *under 13* the ARG2 of *age* in the phrase *1,859 children under age 13*.

5. A Constraint on Empty Categories

Empty categories (Penn Treebank's way of representing gaps) are not typically noun arguments unless they are part of chains that link the empty category to a (pronounceable) word or phrase (the filler of the gap). Consider, for example, the NomBank annotation of *veto* in the following sentence: *Mr. Bush and some other aides are strongly drawn to the idea of trying out a line-item veto. Mr. Bush and some other aides* should be the ARG0 of veto as mediated by: (1) a number of empty categories in the Penn Treebank: the passive object of *drawn* and the subject of *trying*; and (2) the support verb *trying*. In the initial annotation, a NomBank annotator failed to make the final link from the passive object empty category to the lexical NP, but the error detection program predicted that this was a likely error. Exceptions do occur when an empty category represents an unfilled argument. For example, in the following definition of *stock-index arbitrage*, the ARG0 of *trades* should be the same as the empty subject of *executing*, which itself is unbound: *Stock-index arbitrage – Buying or selling baskets of stocks while at the same time executing offsetting trades in stock-index futures or options*.

The Penn Treebank resolves the referential properties of some, but not all empty categories. In the following example, a NomBank annotator needed to add the link between the possessive phrase *Illinois Supreme Court's* and the empty subject of *to institute*: *Illinois Supreme Court 's decision to institute the changes*. Here *institute* acts as a support verb linking its subject to the ARG0 position of the noun *changes*, i.e. the *Illinois Supreme Court* is assumed to be the AGENT of the changes. Therefore, it turns out that only some of the cases where empty categories are not bound in the Penn Treebank need to remain so and it turns out that unbound empty categories are unlikely to be correct as NomBank arguments – their presence signals a likely error.

6. Structural Constraints on External Arguments of Nouns

NomBank specifications place restrictions on the markability of a given potential argument A of a noun N that lies outside of the NP headed by N. It turns out that, for the most part, these restrictions were codable in terms of GLARF'd representations of the sentence and therefore could be automatically checked. Although there are some outliers that the automatic system did not handle correctly, the automatic detection system tended to overpredict errors, rather than underpredict. This made it possible to accurately identify many cases that we needed to review more carefully and it resulted in corrections of many NomBank propositions.

There are three environments in which External arguments can be licensed: (a) support; (b) predication; and (c) PP constructions containing the nominal predicate. Each of these configurations make specific requirements on how the NP-external arguments are linked to the nominal predicate. Furthermore, the absence of any of these configurations means that an NP-external argument is unlicensed and thus tagged as a likely error.

6.1. Constraints on Support Structures

A NomBank external argument A is a legal argument of a nominal predicate P, by virtue of support, if there exists a support chain S linking A to P. To be well-formed, a support chain must meet the following criteria[7]: (1) consist completely of lexical items (leaf nodes) in the Penn Treebank; (2) forms of *be*, auxiliaries, infinitival *to* and modals are skipped, i.e., for purposes of the support chain, we pretend that they do not exist and that the main verb, predicate adjective, or other predicative item is the main predicate of its clause[8]; (3) at least one item in the support chain must have as its part of speech: noun, adjective, verb or determiner[9]; (4) each link in the chain must be the head of the phrase containing it (after allowing for 2)[10]; (5) the first link in the chain must take A as its argument; (6) Each link N in the chain must take the phrase headed by link $N + 1$ as its argument; (7) the last link in the chain must take the phrase headed by P as its argument; and (8) the chain cannot cross any tensed clause phrasal boundaries. A schema of a support chain is provided as Figure 2. Some examples of legal support chains are provided as Figure 3.[11]

There are several ways which we use the constraints on support to verify the accuracy of NomBank annotation: (1) we verify that annotated support chains meet the criteria above; (2) we verify that there are external arguments that require support chains and propose the removal of annotated support chains that are extraneous; (3) we automatically generate a support chain and compare it to the one annotated. In each of these cases, we use the error detection procedures to identify potential errors. Should we determine that they are actual errors, we correct them.

Given a possible external argument A and a nominal predicate P, we assume that exactly one support chain is structurally possible. In simple cases, one can think of the typed feature structure as a labeled tree, although it is actually a rooted directed acyclic graph.[12] In most cases, to find the support chain, one first must identify the path derived by going up the tree from A to the common ancestor of A and P, and then down the tree to P. The support chain is the

[7]For simplicity, we ignore the complications caused by filler/gap constructions (passivization, WH, etc.) and coordination. Nevertheless, these phenomena are handled as well.

[8]This is roughly equivalent of a Verb Group Analysis, extended to cover copula constructions.

[9]The choice of noun, verb and adjective is more limited than the automatically implemented constraints currently allow. One could further limit support items to prepositions, transparent nouns (*a variety of problems*), determiners in partitive constructions (*all of the worst problems*), control predicates (*try, ability, able*), and lexically specific combinations of verbs and nouns (*take a walk, make a mistake*, etc.).

[10]For purposes of discussion, the main verb of a sentence is assumed to be the head.

[11]The final example includes *partner*, a CRISSCROSS noun which simultaneously is a support word for and an argument of *cooperation*.

[12]These graphs are like labeled trees, except they allow shared structure.

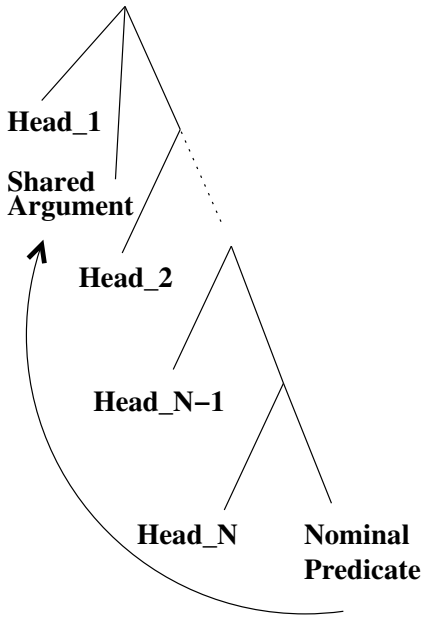

Figure 2: Schema for a Support Chain

1. **IBM**/*ARG0* **made**/*SUPPORT an agreement*
2. **This desk**/*ARG1* **has**/*SUPPORT a height* **of 25 inches**/*ARG2*
3. **their**/*ARG0* **responsibility for**/*SUPPORT* **hard**/*ARGM-MNR decisions.*
4. **The adjuster**/*ARG0* **does**/*SUPPORT a* **lot of**/*Support work* **by phone**/*ARGM-MNR*
5. **it**/*ARG1 is* **scheduled for**/*SUPPORT completion* **by Dec. 10**/*ARGM-TMP*
6. **I**/*ARG0* **take advantage of**/*SUPPORT this opportunity* **to make a plea to readers**/*ARG1*
7. **We**/*ARG0* **had lots of**/*SUPPORT* **internal**/*ARGM-MNR debate* **about this one**/*ARG1*
8. **Saab**/*ARG0 is* **looking for**/*SUPPORT a* **partner**/*ARG2+SUPPORT* **for**/*SUPPORT* **financial**/*ARGM-MNR cooperation*

Figure 3: Examples of Legal Support Chains

set of heads of all the phrases in this path. The complete algorithm for finding support chains must factor in filler gap constructions and coordination. Filler gap constructions complicate the simple algorithm because they are responsible for making the tree into a directed acyclic graph. The graph is derived by changing arcs that point to gaps so that they point to the fillers of those gaps instead. Nevertheless, in the entire Wall Street Journal corpus, we have not encountered a single instance in which multiple support chains linked a given A with a given P.[13] Special allowances are made so that conjoined predicates can both be part of the same support chain, e.g., in *Mary gave and received lots of kisses*, *gave* and *received* are assumed to be branches of the same support chain (*gave* + *received* + *lots* + *of*). It is as if the support chain splits in the middle and then merges together gain because, for the purpose of a support chain, coordinate structures are assumed to have multiple heads.

1. *The* **real**/*ARGM-ADV battle is* **over who will control the market**/*ARG2*
2. *This book is* **about his son**/*ARG1*
3. **Trying to time the economy**/*ARG1 is a mistake*
4. **They**/*ARG1 are* **some**/*ARG2 distance apart*

Figure 4: Linking External Arguments to Nouns Via Predication

6.2. Constraints on Predication

There are a number of instances in which predication licenses a connection between an argument and a noun predicate which we have determined are legitimate for marking NomBank arguments. We specifically avoid cases in which the argument can duplicate existing arguments, e.g., for argument nominalizations like *teacher*, we will always mark *teacher* as its own ARG0 and never NPs linked by predication, e.g., *Mary is John's teacher*.

We recognize the following markable instances of linking external argument to nouns via predication: (1) when the noun predicate is the subject of the sentence and one of its arguments follows a copula, e.g., Examples, 1 and 2 in Figure 4; and (2) when the noun predicate P follows the copula and its argument precedes the copula and P is either a nominalization of an adjective, an ATTRIBUTE noun (a NomBank class) or in a preposition plus noun construction that has an adjective-like distribution, e.g., 3–4, in Figure 4. A subset of the nouns in COMLEX Syntax that are marked with the feature (COUNTABLE :PVAL) combine with the preposition to form adjective-like constituents, e.g., the entry of *alert* is marked (COUNTABLE :PVAL ("on")). These entries can be used to identify instances of the aforementioned adjective-like PP construction.

Identifying these environments automatically is easy. One merely has to identify copulas, the subjects of those copulas (typically the NP or sentence immediately following the copula) and the underlying predicate (typically the phrase immediately following the predicate and often marked with the function tag -PRD). Other predicative environments, though rarer, are also easy to detect in the Penn Treebank: *small clauses* are *S* constituents consisting of an NP followed by another constituent marked with -PRD, *as* con-

[13] This is, at least in part, due to the constraint that a support chain cannot cross a tensed sentential node. This prevents, for example, support chains from including predicates on both sides of a relative clause boundary.

1. **Without**/*ARGM-NEG question,* **something intriguing is going on**/*ARG1* [PP Parenthetical]
2. **Some last-minute phone calls that Mr. Bush made**/*ARG1 (at the behest* **of some conservative U.S. senators**/*ARG0)* **to enlist backing for the U.S. position**/*ARG1* [PP Parenthetical]
3. **He**/*ARG1 was under consideration* **to succeed Joshua Lederberg**/*ARG2* [PP + Extraposition]
4. **ABC's baseball experience**/*ARG0 may be of interest* **to CBS Inc.**/*ARG1* [PP + Extraposition]
5. **they**/*ARG0 exercise for enjoyment* [Subject-Oriented PP]
6. **Garbage**/*ARG0 made its debut this fall with the promise* **to give consumers the straight scoop on the U.S. waste crisis**/*ARG1* [Subject-Oriented PP]
7. **Participants**/*ARG0 in the meeting* [Noun-Modifying PP]
8. *the* **bitterness**/*ARGM-MNR of the battle* [Noun-Modifying PP]
9. **That**/*ARG1 was in addition* **to $34,000 in direct campaign donations**/ARG2 [Discourse Connective]
10. **That $130 million gives us some flexibility**/*ARG1 in case* **Temple raises its bid**/*ARG2*. [Discourse Connective]
11. *In important particulars,* **the Soviets are different from the Chinese**/*ARG1* [Discourse Adverbial]
12. *In fact,* **they don't take it seriously at all**/*ARG1*

Figure 5: PP constructions that license External Arguments

stituents begin with the word *as*, etc.

6.3. PP constructions and External Arguments

When the NP headed by a predicate noun is the object of a preposition, the argument taking properties of that noun may change. This subsection describes a set of argument-taking environments in which such PPs license external arguments according to NomBank guidelines. These environments include: (1) The PP-parenthetical construction; (2) The PP + Extraposition construction; (3) Subject Oriented PPs; (4) Noun modifying PPs; and (5) Other Adverbial PPs including discourse connectives. Examples are provided in Figure 5. Although we can automatically detect most of these environments, we have not implemented ways of detecting all of them. Thus our automatic procedures still flag many of these as instances of unlicensed external arguments. As a result, many of the rarer PP constructions are always revisited during the error detection phase of annotation.

The PP-Parenthetical (Examples 1 and 2) and extraposed PP constructions (Examples 3 and 4) are both licensed by COMLEX Syntax dictionary entries and, in the former case, is limited to a short list of prepositions. The configurations are easily defined in terms of syntactic trees (or graphs). The PP-Parenthetical cases are licensed by nouns that take clausal complements and this lexical information is readily available from a combination of COMLEX Syntax and/or Nomlex (or Nomlex-Plus). These PP phrases (Examples 1 and 2 in Figure 5) are like their verbal counterparts (e.g., the *say* phrase in *Mary, John said, is an incredible botanist*) in that they can precede, follow or infix their sentential argument. In addition to the lexical subcategorization of the nominal predicate, another restriction is that only a narrow set of prepositions seem to license this construction: (*with, without, at, on, in* and possibly a few others). The PP is immediately dominated by the sentence that it takes as an argument (the PP is typically marked as a parenthetical in the Penn Treebank or offset by parentheses or commas). The Extraposition cases (Examples 3 and 4) are possible for a subset of nouns marked in COMLEX Syntax with the subcategorization features EXTRAP-P-NOUN-THAT-S. The COMLEX entry also specifies the preposition. For example, the COMLEX entry for *interest* includes the subcategorization feature (EXTRAP-P-NOUN-THAT-S :PVAL ("of")). In the Penn Treebank, the nominal predicate is the rightward argument of the copula and the subject of the copula is one argument of the noun. Using a combination of these lexical clues and configurational data, it is easy to see how correctly licensed instances of these constructions can be automatically identified.

Subject oriented adverbial PPs containing a NomBank predicate (Figure 5, Examples 5 and 6) can be identified by the following characteristics: (1) the subject of the sentence is an argument of the NomBank predicate (hence the name subject-oriented); (2) the PP is either a child of the sentential node or a child of the VP; and (3) the preposition belongs to a defined set which includes mainly temporal prepositions (*after*, *before*, *during*), instrumental prepositions (*with*, *without*, *through by*) and several others. These PPs are similar to other subject-oriented adverbs like *willingly*, *vengefully*, etc., which typically select for an animate subject.

The fourth case (Figure 5, Examples 7 and 8) involves a noun A that is modified by a PP containing a nominal predicate P, such that P takes A as an argument. This is an easy to recognize configuration and is limited to approximately the same set of prepositions as the others. We have yet to fully figure out the distribution of the nominal predicates that can occur in this configuration, although it does seem that adjective nominalizations and ATTRIBUTE nouns are the most common.

Finally, there are some NomBank frame entries that classify particular nouns as being either a discourse-connective (Examples 9 and 10) or discourse-adverbial (Examples 11 and 12). Similar entries are found in the NOMADV dictionary giving them one of the COMLEX Syntax classes applied to similar adverbs, i.e., the various sub-types of the META-ADV class (the connectives belong to the (META-ADV :CONJ T) class). The discourse adverbials can take entire sentences as arguments, whereas the discourse connectives link two arguments in a similar manner to the discourse connectives in the Penn Discourse Treebank (PDTB). Nom-

1. *After* **hours**/*SUPPORT+ARGM-TMP* **of**/*SUPPORT* *debate*, **the jury**/*ARG0 focuses on the facts*

2. **John**/*ARG1 is* **40**/*ARG2* **pounds**/*ARG2+SUPPORT* **in**/*ARG2* *weight*

Figure 6: Combining Support with Other Phenomena

Bank discourse connectives can link two sentences, two NPs or one NP and one sentence. This contrasts with PDTB connectives, which always link two sentences. The discourse adverbials, like the Parentheticals can preceded follow or be embedded in the sentence it modifies. NomBank discourse connectives have a similar configurational distribution as the PDTB connectives: the connective forms a constituent with one argument (e.g., *in case Temple raises its bid* in Example 10 and the other argument is either the rest of the superordinate phrase (the subject and the verb) or the subject of the sentence (e.g., *that* in Example 10). However, unlike PDTB, NomBank does not link predicates in one sentence with arguments outside that sentence, e.g., NomBank does not mark the sentence preceding an example like no. 12 as an argument of *fact*.

In summary, there are a number of configurations in which a PP containing a NomBank predicate (as the head of the prepositional object) that license external arguments of that noun. The configurations are easy to define and additional lexical restrictions makes it possible to identify the markable cases in NomBank. As of this writing, we recognize a subset of the admissible cases automatically. The remainder we must verify manually.

6.4. Combining Support with Other Constraints

We end this section with the examples in Figure 6, which combine support with some of the other external argument licensing environments. Both cases involve transparent noun constructions, which are viewed as a type of Support in NomBank. *After hours of debate* is treated as if *debate* is the main predicate of this subject-oriented PP construction (the subject of the sentence is an argument of *debate*). The support chain *hours + of* makes this treatment possible. In a similar way, the support chain *pounds + of* makes it possible for *weight* to be connected to the subject of the sentence by predication. The support chains serve to bring the nominal predicate into the position required to link them via these other types of constructions.

7. Lexical Constraints on NomBank

We will now describe one of the main dictionary-based constraints that we used to correct NomBank. At the same time, we used this constraint to correct the dictionary ADJADV (Meyers et al., 2004b), which we made along side of NomBank.

Although ARG1 ... ARG9 features were applied according to frames for particular words, the distribution of the ARGM features was left to the annotator's interpretation of the NomBank specifications. Nevertheless, to a large extent the ARGM features are also lexical in nature, but of a different sort. ARGMs tend to be the same for particular modifiers (the value of the ARGM itself), rather than

```
(ADJADV       :ORTH       "abject"
              :ADV        "abjectly"
              :FEATURES   ((MANNER-ADV)
                           (GRADABLE)))
(ADJADV       :ORTH       "actual"
              :ADV        "actually"
              :FEATURES   ((META-ADV
                            :VIEWPOINT T)))
(ADJADVLIKE   :ORTH       "big"
              :ADV        "immensely"
              :FEATURES   ((MANNER-ADV)
                           (DEGREE-ADV)))
```

Figure 7: Sample ADJADV Entries

nominal predicate. For example, the adjective *recent* is almost always marked ARGM-TMP due to lexical properties of *recent*, not lexical properties of the noun it modifies. Thus *recent* should be marked ARGM-TMP in *the recent destruction of the documents*, *their recent marriage* and *the recent knowledge*, regardless of what is in the frame entries of *destruction, marriage* and *knowledge*. We observed that the relevant information could not be found in the adjective entries of COMLEX Syntax, but could be found in related adverb entries. Specifically, *recently*, the adverb related to *recent* has the feature *TEMPORAL-ADV*. This motivated our construction of ADJADV. Some sample entries are given below in Figure 7. This dictionary was created in a semi-automatic way. For the most part, we simply found morphologically adjective adverb pairs and generated the entry based on the adverb. However, in some cases, e.g., *big*, we created an ADJADVLIKE entry based on a semantically related adverb.

Given the assumption that specific adjectives tended to be compatible with the same ARGM function tags, we could automatically detect likely errors by comparing the ARGMs assigned adjective premodifiers in NomBank against the ADJADV dictionary entries for those adjectives.[14] We assumed the table of compatibilities between function tags and COMLEX-SYNTAX features listed as Table 1. When an adjective was marked in a NomBank proposition in a way that was incompatible with the ADJADV entry, this would usually lead to either changing the NomBank annotation or changing the ADJADV lexical entry. In this way, we were able to simultaneously improve both NomBank and ADJADV.

8. Concluding Remarks

Above, we have outlined major ways in which we have improved NomBank by evaluating the compatibility of annotation with other resources. As a result of these and similar techniques, we have looked closely at over 30,000 of

[14] Some premodifiers were handled in other ways, e.g., prefixes were specially classified; numbers between 1000 and 2100 were assumed to be potential time modifiers, etc. Also, with respect to hyphenated items, we identified one hyphenated segment (typically the last segment) as the head and looked up the ADJADV entry for that segment. We omit a full description due to space limitations.

COMLEX Feature	ARGM
(META-ADV :CONJ T)	ARGM-DIS
other META-ADV	ARGM-ADV
MANNER-ADV	ARGM-MNR
DEGREE-ADV	ARGM-MNR
EVAL-ADV	ARGM-MNR
LOC&DIR-ADV	ARGM-LOC, ARGM-DIR
TEMPORAL-ADV	ARGM-TMP

Table 1: ADJADV/ARGM Compatibility

the 114,500 NomBank instances. We believe that these measures caused us to focus our efforts on the most likely causes of error, improving both the accuracy and efficiency of quality control. Had we annotated NomBank twice and then adjudicated instead of using this methodology, it would clearly have been a more expensive undertaking. Furthermore our attention would not have been as directed as it was using the error detection program.[15]

We have considered creating a degraded version of NomBank that consists of only pre-edited entries. We could then test to see if a automatic role labeling system (Jiang and Ng, 2006) trained on that version would not perform as accurately as a system trained on the final version. Better performance on the final system would confirm that we improved the system using our methods. However, this result would hardly be surprising because our technique does involve a selective second pass on the annotation by an expert annotator, methodology which is widely recognized to improve results. Clearer evaluation would require the annotation of additional data in a test setting in which duel annotation plus adjudication could be fairly compared with the method described here. This will be possible should we have the opportunity to annotate a substantial amount of additional NomBank data. However, given our limited resources, we are confident that we took the best possible approach.

This paper provides examples of how constraints on a new annotation scheme can be formulated in terms of previous annotation in order to provide quality control. Researchers who would like to take advantage of this methodology should consider annotating corpora that has already been annotated by other members of the annotation community

Acknowledgments

This research was supported by the National Science Foundation, award CNS-0551615, entitled Towards a Comprehensive Linguistic Annotation of Language.

9. References

Z. P. Jiang and H. T. Ng. 2006. Semantic role labeling of nombank: A maximum entropy approach. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australia.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

A. Meyers, R. Grishman, M. Kosaka, and S. Zhao. 2001a. Covering Treebanks with GLARF. In *ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*.

A. Meyers, M. Kosaka, S. Sekine, R. Grishman, and S. Zhao. 2001b. Parsing and GLARFing. In *Proceedings of RANLP-2001*, Tzigov Chark, Bulgaria.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004a. The NomBank Project: An Interim Report. In *NAACL/HLT 2004 Workshop Frontiers in Corpus Annotation*, Boston.

A. Meyers, R. Reeves, Catherine Macleod, Rachel Szekeley, Veronkia Zielinska, and Brian Young. 2004b. The Cross-Breeding of Dictionaries. In *Proceedings of LREC-2004*, Lisbon, Portugal.

E. Miltsakaki, A. Joshi, R. Prasad, and B. Webber. 2004. Annotating discourse connectives and their arguments. In A. Meyers, editor, *NAACL/HLT 2004 Workshop: Frontiers in Corpus Annotation*, pages 9–16, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani. 2004. The Specification Language TimeML. In I. Mani, J. Pustejovsky, and R. Gaizauskas, editors, *The Language of Time: A Reader*. Oxford University Press, Oxford.

J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. 2005. Merging PropBank, NomBank, TimeBank, Penn Discourse Treebank and Coreference. In *ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*.

Y. Shinyama and S. Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of NAACL/HLT*, New York, New York, USA. Association for Computational Linguistics.

Theresa Wilson and Janyce Wiebe. 2003. Annotating Opinions in the World Press. In *4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*.

S. Zhao, A. Meyers, and R. Grishman. 2004. Discriminative Slot Detection Using Kernel Methods. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, Geneva.

[15] Of course, we fixed errors that we found that were not detected by the program as well.

A Dictionary-based Model for Morpho-Syntactic Annotation

Cvetana Krstev[1], Svetla Koeva[2], Duško Vitas[3]
[1] University of Belgrade, Faculty of Philology, Studentski trg 3, RS - 11000 Belgrade,
[2]Bulgarian Academy of Sciences, 52 Shipchenski prohod, Bl. 17, BG - 1113 Sofia,
[3]University of Belgrade, Faculty of Mathematics, Studentski trg 16, RS - 11000 Belgrade
E-mail: cvetana@matf.bg.ac.yu, svetla@ ibl.bas.bg, vitas@matf.bg.ac.yu

Abstract

The main goal of this paper is to establish a proper and flexible method for morpho-syntactic annotation taking in consideration such language phenomena as multi-word units, complex word forms, regular and productive derivational processes, etc., which usually remain outside the scope of the morpho-syntactic annotation. We present the first results in the development of a multilingual resource that should enable the exploration of the possibility to apply various different lexical bases, such as inflectional dictionaries and multilingual lexical databases as Wordnet and Prolex that were developed during the last decade. This paper is limited to two Balkan languages, Serbian and Bulgarian.

1. Introduction

The paper outlines an approach for morpho-syntactic annotation and the first results in the creation and exploitation of an aligned and annotated corpus for the Bulgarian-Serbian pair. The main goal of this effort is to establish a proper and flexible method for morpho-syntactic annotation taking in consideration such language phenomena as multi-word units, complex word forms, regular and productive derivational processes, etc., which usually remain outside the scope of the morpho-syntactic annotation. Some of the existing morpho-syntactic annotation schemes consider only tokens thus neglecting the fact that a token is not always equal to a word form – namely, a word form can consist of several tokens (not necessarily contiguous) and several word forms can build a token. On the other hand, some of the proposed sets of morpho-syntactic attributes and their values are inconsistently composed, not taking into consideration the relative function of the chosen attributes or their relations with the higher language levels. In our approach we accept the assumption that the morpho-syntactic annotation has to be assigned to the word forms irrespective of their continuity and contiguity. Thus the term *word form* here means (following in general the MAF[1] prescriptions) contiguous or non-contiguous unit consisting of one or more tokens that refers to a single concept: single word, complex word forms (i.e. complex tenses, mood, aspect) and multi-word units. We also agree that the standardization in morpho-syntactic annotation has to cover both correspondences between different languages as well as language specific features (Ide et al., 2003). That is why two Balkan and South Slavic languages are taken in focus: Bulgarian and Serbian, for which similar language resources have been developed recently.

The particular research aims, stated in this paper, are as follows:

- Briefly to show some of the gaps in the existing annotation schemes;
- To offer some techniques for handling the morpho-syntactic annotation of word forms rather than tokens;
- To exploit parallel language resources.

These research aims are directed to the development of a complex method for morpho-syntactic annotation providing uniform and flexible way of treating word forms.

In the following section we present a short analysis of the related work. In the third section, we describe the different resources for Bulgarian and Serbian developed during the past years in the same or comparable format which are used in the course of the work. The forth and the fifth sections explain how we apply different techniques for morpho-syntactic annotation compatible to word forms corresponding to one ore more tokens. Finally, we discuss the presented study and propose future work to be done[2].

2. Previous research

One of the basic common resources for European languages during the last decade was developed within the Multext-East project (Erjavec, 2004)[3]. This resource consists of three main components for each of the languages included in the project, namely a proposed standard for morpho-syntactic description (further on, MSD), the text of the translation of Orwell's novel *1984* in the corresponding language, and the application of MSD on the annotation of the lemmatized version of the text of this novel. Besides these components, Multext-East encompasses the aligned versions of *1984* on the sentence level, by means of the Vanilla-aligner, for all languages included (and also added later) in the project. The results of this project on the level of the description of morpho-syntactic parameters were refined and enhanced several times, and the project results found a wide use in the research community.

At present, despite the success of this project, its

[1]ISO TC 37/ SC 4 N225

[2] The first version of this paper was presented before the very small audience at the Workshop '*A Common Natural Language Paradigm for Balkan Languages*' that was held in conjunction with RANLP Conference 2007.

[3] http://nl.ijs.si/ME/

shortcomings can be observed both regarding the content of the MSD, and in the way this description has been applied to specific languages, as presented, for instance, in (Przepiórkowski & Woliński, 2003). First of all, the principles taken into consideration when particular attributes and values are included in the Multext-East MSD are not always clear and consistent. Some of the attributes are properties of the lemma, some of them – properties of particular word forms only. The question is how the recommended attributes and values are chosen to be included in the MSD – are they those that determine the inflectional paradigm, or the agreement properties, or those relevant for the temporal and modal features, etc. If we consider the inflectional paradigms in Bulgarian and Serbian, we can give examples showing that other sets of categories than those defined in the Multext-East MSD determine these paradigms. The attribute Animateness with values human, animate and non-animate is not specified for Bulgarian but it determines the vocative and count slots in the noun paradigm. The word form *dvojica* 'two men' in Serbian is the nominative singular of the noun *dvojica* that behaves on the inflectional level as a noun of feminine gender in singular, while it actually represents the natural masculine gender in plural. This information has to be attached to the lemma and it determines the complex agreement conditions in Serbian which cannot be expressed within the Multext-East MSD.

Thus the criterion for the morpho-syntactic specifications of any language has not to be based on the set of the attributes shared with a group of other languages rather than on the set describing morpho-syntactic properties of a given language: a minimal set has to include those attributes and values that are relevant for the inflectional paradigms of the single word; more descriptive sets have to include attributes and values relevant for complex word forms as well as for the multi-word units, etc. The parallel processing of two or more languages has not to be limited to a predefined set of attributes and values that the languages share but to a flexible set that can be relevant for a particular NLP task.

Considering the application of MSD on text annotation of Orwell's novel, a unique method for obtaining annotated versions of the novel in different languages was not established, neither were methods for producing annotated text (automatically or manually) explicitly stated. This observation means that the information for resolving possible ambiguities in annotation is not explicitly represented and especially that the manner of disambiguation is not explained. As a consequence of this inconsistency the obtained annotated texts of Orwell's 1984 contain only the final result of the morphological and lexical analysis, where the mechanism of morphological analysis remains hidden, which means that the method of the assignment of the lemma and the MSD to the word form cannot be reproduced on a new text in the same manner. A possible application of a stochastic tagger trained on such a training set to a new text requires a thorough verification of the obtained results, which is in essence a more complicated task than the initial annotation of Orwell's text (because it has to be established whether the MSD attributed to a word in the text in such a manner is correct or false, instead of selecting the correct MSD among several possibilities).

To the great extend the ideas presented in this paper are synchronized with the proposal for an ambiguity handling through lattices based on a two level structuring for tokens and word forms involving the use of feature structures for morpho-syntactic content (Clément et al., 2005).

3. Parallel language resources

3.1. Parallel Bulgarian-Serbian corpus

The parallel corpus is compiled from the French text of Jules Verne's novel *Around the world in 80 days*, which has been aligned with its translations in a number of languages including English, Bulgarian and Serbian[4]. The alignment was accomplished using the Xalign system (Romary & Bonhomme, 2000)[5]. From the TEI-format obtained in this way, several versions of texts have been created in other formats such as TMX, Vanilla-like format and HTML (Appendix 1). The alignment was performed at the paragraph and segment levels in a manner that established a one-to-one correspondence between the original and the translation by means of additional manual segmentation, but preserving the segmentation of the original. This enabled the maintenance of the one-to-one correspondence between all the language pairs processed (Figure 1).

n506 : Сигурно добре знае, че в Индия, която е английска земя, няма да е в безопасност.	**n506 :** On dobro zna da necxe biti siguran u Indiji jer je to engleska zemlxa.
n507 : – Освен ако не е много опитен – отговори консулът.	**n507 :** -- Sem ako to nije vrlo odvazxan cyovek.

Figure 1: Aligned Bulgarian-Serbian parallel corpus

	Bulgarian	English	French	Serbian
# words	58 162	64 831	68 359	60 227
# sentences	4 435	4 435	4 435	4 435
# paragraphs	1 963	1960	1963	1963

Table 1: Statistical data for the parallel corpus

Although the aligned parallel corpus is relatively small at this stage (see figures in the Table 1) it is a part of the MaT[6] project, whose aims are directed to the compilation of a large multilingual parallel corpus of Balkan, South Slavic and bigger European languages that will be constituted of texts from different ranges (most of the existing parallel corpora of European languages as Acquis

[4] Besides these languages, the novel is fully or partially aligned in as many as nine other languages

[5] led.loria.fr/outils.php

[6] SEE-ERANET project *Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages*

Communautaire[7] consist of legislation documents only).

Starting from the text obtained in the above-mentioned manner we analyze different issues of morpho-syntactic annotation. All examples in this paper are taken from the Bulgarian and Serbian subparts of the parallel corpus.

3.2. Bulgarian and Serbian e-dictionaries

Various formalisms for the representation of linguistic knowledge are available, and at the first place, different types of morphological dictionaries and local grammars. The basic monolingual lexical resources for Bulgarian and Serbian considered in this paper are systems of morphological dictionaries in the so-called LADL-format (Courtois & Silberztein, 1990). This format is compatible with the draft of Lexical markup framework (LMF) standard[8], and an automatic conversion from this format into LMF is enabled (Krstev et al, 2006b). Automatic conversion from this format into Multext-East has also been successfully performed for Serbian (Krstev et al., 2004). The dictionaries for Bulgarian are described in (Koeva, 2004: Koeva, 2005) and some samples are presently available under the NooJ format[9], whereas the dictionaries for Serbian, developed under both the Unitex system[10] and NooJ, are outlined in (Vitas & Krstev, 2005)[11].

The common feature of these dictionaries is that they are developed within the same theoretical and methodological framework that enables a multi-level application of the results of the theory of finite state transducers to text processing. The basic form[12] of the entry in the morphological dictionary is described by the following pattern:

(*) ..... word form, lemma. K+SynSem:(mc;)*

where word form and lemma are simple words or continuous multi-word units, whereas K is a code that contains the information on the part-of-speech and inflective properties of the lemma, usually in the form of a corresponding finite transducer. SynSem is the sequence of syntactic and semantic attributes attached to the lemma, while mc represents the sequence that describes the relation between the word form and the lemma by means of specified values of grammatical categories. For instance, the following entries from the Serbian and Bulgarian dictionaries

lisca,lisac.N+Hum+Zool:ms2v;ms4v
лисица,лисица.N+F:s0

establish *lisca* 'fox' in Serbian as the genitive (2) or accusative (4) singular (s) form of masculine gender (m) noun (N) *lisac* that is marked as animate (v) and that can have the semantic feature Hum (for humans) (e.g. seg 1915: *Svoje brige poveri Fiksu, koji - prevejani* ***lisac*** *- pokuša...* '*He had confided his anxiety to Fix who--the sly* ***rascal****!--tried...*') and Zool (for animals) and respectively *лисица* 'fox' in Bulgarian as feminine (F) noun (N) whose form *лисици* is in singular (s) and indefinite (0) (e.g. seg. 1915: *Той бе доверил притесненията си на Фикс, а той – хитрата* ***лисица*** *– се опитваше ...*).

The WS4LR tool (Krstev et al, 2006b) can be used to enrich the SynSem field in the pattern (*) by transferring the information from semantic networks. For instance, the information on currencies was in this way transferred from Wordnet to e-dictionary:

gvineja,.N+Cur:fs1q:fp2q // guineas

Here, the marker *Cur* represents the names of currencies.

The attributes form the lexical database Prolex (Vitas et al., 2007) can be transferred into e-dictionaries by applying the same procedure:

Bombaj,.N+NProp+Top+Gr:ms1q:ms4q // Bombay
bombajskoj,bombajski.A+PosQ+NProp+Top+Gr:aefs3g

Here NProp represents a proper name, PosQ a relational adjective, Top a toponym, Gr a city.

4. Ambiguity

At least two types of PoS ambiguity can be distinguished in Bulgarian and Serbian.

Lexical ambiguity is observed when the ambiguous word forms pertain to different lemmas (usually with different POS) e. g. in Bulgarian the word *разходи* may either be the plural indefinite form of the masculine noun *разход* 'expense' - разходи,разход.N+M:p0,, or the third person singular present tense; second person singular aorist; third person singular aorist; and second person singular imperative of the verb *разходя* '*to take for a walk*' - разходи,разходя.V+F+T:P2s:R2s:R3s:I2s (i.e. seg 2345: *Това влиза в общите* ***разходи!*** *'This enters into ... general* ***expenses*****'** and seg 1761: *…госпожа Ауда, която бе проявила желание да се* ***разходи*** *'...Aouda, who betrayed a desire for a* ***walk****...'*).

Morphological ambiguity occurs when a given lemma has two or more identical distinct word forms, e.g. in Bulgarian inanimate masculine nouns such as *въпрос* 'question' whose singular definite short article (sh) and counted form (c) coincide: въпроса,въпрос.N+M:sh:c (seg. 1223: *Сър Франсис Кромарти му постави открито* ***въпроса***. '*Sir Francis frankly put the* ***question*** *to him*' vs. seg. 1896: *Зададе сто* ***въпроса*** *на капитана, офицерите, моряците Не* ***overwhelmed*** *the captain*). Assume the processing of the words in the following sentence (seg 56):

(1-sr) Pojavi se momak tridesetih godina i pozdravi.
(1-bg) Влезе един млад мъж на около тридесет години и поздрави.
(1-en) A young man of thirty advanced and bowed.

[7] http://langtech.jrc.it/JRC-Acquis.html
[8] http:// www.tc37sc4.org
[9] http://www.nooj4nlp.net/
[10] http://www-igm.univ-mlv.fr/~unitex/
[11] Dictionaries in this format exist for several other Balkan and South Slavic languages: Greek (Kyriacopoulou et al., 2002), Romanian (Dimitriu, 2005), Macedonian as well as for Albanian and Croatian in an initial stage.
[12] This is the format of the dictionary of inflected forms, the so called DELAF, derivable from a dictionary of non-inflected forms called DELAS

All its possible morphological interpretations will than be listed where among other things, we can see that (a) the form of the word *pojavi* can be interpreted as a form of the noun *pojava* 'appearance' or as a form of the verb *pojaviti (se)* 'to appear', (b) the form *pozdravi* as a form of the noun *pozdrav* 'greeting' or a form of the verb *pozdraviti (se)* 'to greet'. At the same time, both forms realize several different values of morphological categories. For example, if *pozdravi* is the form of the verb *pozdraviti*, then it can represent the third person of the present tense or the second person of imperative or aorist singular. Similar ambiguity is observed in Bulgarian: *появи* is either plural of the noun *поява* '*appearance*' or one of the four different forms of the verb *появя се* '*to appear*'; *поздрави* is either plural of the noun *поздрав* '*greeting*' or one of the four different forms of the verb *поздравя* '*to greet*'. These different interpretations are for Serbian represented by the graph in Figure 2.

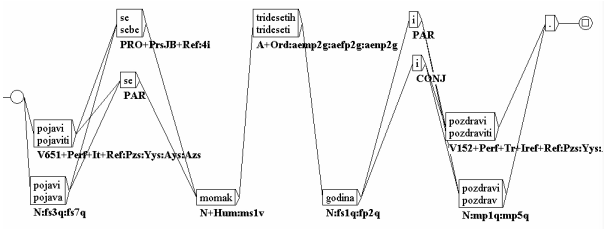

Figure 2: The sentence graph for the Serbian segment 53

This illustrates the problem of essential ambiguity in the interpretation of the incoming sentence. Let us compare the sentence (sr-1) with the result obtained by TnT (Brants, 2000), trained on the Serbian annotated text of *1984*:

Pojavi	Vm-p3s-an-n---e
se	Q
(.....)	
i	C-s
pozdravi	Ncmpn--n

Here MSD values are incorrectly established for the forms of *pojavi* (present instead of aorist) and *pozdravi* (noun instead of verb), and the problem of ambiguity of forms, characteristic for Slavic languages, remains completely hidden in the method used by TnT.

5. Annotation refinement

The parallel Bulgarian-Serbian corpus is annotated with the grammatical information available from Bulgarian and Serbian morphological dictionaries. After the annotation most of the unrecognized words are foreign proper names but there are also some words built by the regular derivation rules. Bulgarian and Serbian are languages with highly productive derivation concerning diminutives, relative adjectives, negative adjectives, adjective and adverb comparative forms (we are not going to discuss here whether comparison reflects in different lexemes or different word forms), verb aspect pairs, etc. Some of the words built by the regular derivation rules which are not included in the electronic dictionaries might be recognized by means of the respective morphological grammars. On the other hand, neither multi-word units (MWUs) nor complex word forms (both continuous and discontinuous) are recognized by the traditional electronic dictionaries. Continuous MWUs and complex word forms might be handled in an uniform way in morphological dictionaries together with the simple words, while discontinuous word forms might be processed by means of local grammars. Providing these techniques for a morpho-syntactic annotation of the word forms might bring new horizons in the POS tagging – to the best of our knowledge there are no POS taggers available that handle MWUs or complex word forms.

Thus the basic annotation assigned from morphological dictionaries can be refined in several ways. We shall indicate here only some of these techniques in order to show the directions towards a proper morpho-syntactic annotation.

5.1. Contiguous multi-word units

The issue of morpho-syntactic specifications of multi-word units (distributed among natural languages approximately equivalently and covering one fourth of the lexis according to the data represented in the European wordnets and one tenth of the words used in real texts according to the data coming from the Bulgarian sense tagged corpus) is very important. A multi-word unit can correspond to a single word in another language, for example: the multi-word unit *френско грозде* 'red currants' in Bulgarian corresponds to the single word *ribizlama* in Serbian or *groseilles* in French (seg 154: … *пълнен със стръкчета ревен и зелено* ***френско грозде****… kolača punjenih stabljikama ravente i zelenim* ***ribizlama****…un gâteau farci de tiges de rhubarbe et de* ***groseilles*** *vertes...* '*a rhubarb and* ***gooseberry*** *tart*'). Consequently multi-word units refer to a unique concept and have to be treated in a uniform way together with the single words. Attempts towards proper morpho-syntactic description of both single words and MWUs, were scarce so far. However, a description of the inflection of multi-word units based on dictionaries of simple words is given in (Vitas & Krstev, 2005), and further enhanced for some Slavic languages in (Koeva, 2004) and (Krstev et al., 2006).

An example of a MWU is presented by the expression *s vremena na vreme* in Serbian or *от време на време* in Bulgarian that represents an adverbial syntagma. On the level of simple word categories that sequence would be analyzed as Preposition Noun Preposition Noun, for instance by TnT:

s (Spsg) *vremena* (Ncnsg--n) *na* (Spsa) *vreme* (Ncnsa--n)

On the level of the dictionary of multi-word units such a sequence is described as adverbial syntagma, and the result of the annotation would add the following information:

s vremena na vreme.ADV+C
от време на време.ADV+C
(from time to time)

Here C indicates that a compound adverb is in question.

Another example of MWU is the sequence *Hong Kong* in the following sentence (seg 1963):

(2-sr) *Hong Kong je ostrvce koje je ... pripalo Engleskoj*
(2-bg) *Хонконг е островче под английско владение...*
(2-fr) *Hong-Kong n'est qu'un îlot (...) assura la possession à l'Angleterre*
(2-eng) *Hong Kong is an island which came into the possession of the English by the Treaty of Nankin*

In Serbian *Hong Kong* can be written in three different ways: *Hongkong* (as in Bulgarian), *Hong Kong* (as in English) or *Hong-Kong* (as in French). In the first case, it is a simple word (as a contingent sequence of alphabetic characters), whereas in the other two cases it is a multi-word unit or a compound word (composed of two simple words divided by a separator). As components of the MWU *Hong Kong* do not exist in the dictionary (neither *Hong*, nor *Kong*), the analysis on the level of simple words will mark this sequence as two unknown words. One solution of this problem would be the construction of a dictionary of MWUs with a structure analogous to the structure described by the pattern (*). A formalism is presented in (Savary, 2005) that enables the formalization of inflections of MWUs, analogous to the definition of the inflection of simple words.

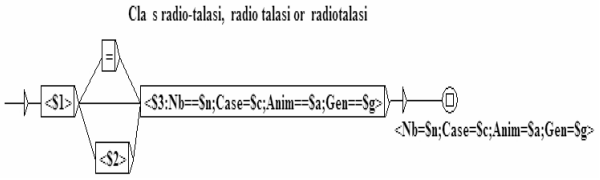

Figure 3: The inflectional graph for two-component compound for which the first component does not inflect, and a space between them can be either omitted or replaced by a hyphen.

The result of the application of this formalism is that, on the basis of the graph depicted in Figure 3, all forms of the inflectional paradigm will be generated for the three graphemic representations of the sequence Hong Kong. The analysis of the initial part of the sentence (sr-2) yields the graph given in Figure 4.

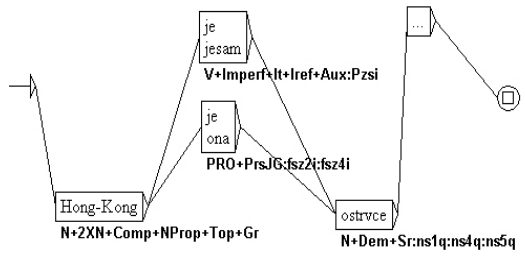

Figure 4: Sentence graph for the beginning of the sentence sr-2

In the interaction of the system of electronic dictionaries with the lexical database of proper names Prolex, the sequence Hong-Kong obtains also the attributes NProp (proper name), Top (toponym) and Gr (city). A comprehensive solution of the complex problem of numeral recognition is for Serbian presented in (Krstev & Vitas 2007).

5.2. Regular productive derivation

Another issue arises on the level of regular derivation. Namely, in the recognition of the results of the derivational processes the meaning of the derived form that is usually not described in the dictionary of the simple words is deduced from the meaning of the initial word (Vitas et al., 2007). In this way it is possible to associate the word forms that usually do not belong to the dictionary of simple words and thus remain in the category of unrecognized words after text analysis with the precise description (the level of precision is the same as that obtained by the word forms belonging to the dictionary). These processes are present both in Serbian and Bulgarian in deriving the diminutives, possessive adjectives, negative adjectives, verb aspect pairs, etc. among others.

This issue is illustrated in the example (sr-2) by the forms *ostrvce* in Serbian and (bg-2) by *островче* in Bulgarian, which are diminutive forms of the respective nouns *ostrvo* '*island*' in Serbian and *остров* in Bulgarian. The productivity of certain derivational processes such as the formation of diminutives, possessive and relational adjective, etc. are characteristic for Bulgarian and Serbian. From the angle of the completeness of electronic dictionaries, it is clear that all results of such derivational processes, which we will call regular derivation, cannot be described in the dictionary of simple words.

The forms generated by such processes can be described by a specific type of finite-state transducers, the so-called morphological grammars, which represent models of respective derivational processes. Such grammars are applied to words that remained unrecognized in the process of analysis, and enable the reduction of the unrecognized form to a lemma form missing from the dictionary. Thus, by applying the appropriate morphological grammar, for the word *ostrvce* in Serbian and *островче* Bulgarian the following sequence is generated on the output of the analyzer:

ostrvce,ostrvce.N+Dem+Sr:ns1q:ns4q:ns5q
островче,островче. N+NE+Dem:s0

where the attribute Dem, added by a morphological grammar, indicates that the word is a form of diminutive.

In the example of sentence (2) a problem in the multilingual context is also posed by the identification of proper names. Namely, in Serbian the toponym *Engleska* was used, whereas in (2-bg) the translation uses the adjective form *английско.* One solution that enables the linking of these two word forms in a multilingual context in a systematic way is analyzed in (Maurel et al, 2007).

5.3. Complex word forms and discontinuous MWUs

The third question concerns complex morphological categories which are usually excluded from the morpho-syntactic specifications. But a synthetic form in one language might correspond to an analytical one in another language, i.e. *ще чета* '*will read*' in Bulgarian corresponds to *će čitatiti* = *ćitaće* in Serbian, consequently they should also be treated in a uniform way. Most of the analytical forms are discontinuous – they allow other words – mainly clitics in Bulgarian to interrupt their parts.

Local grammars, as concepts defined in (Gross, 1993), enable the construction of finite transducers, which recognize and tag different structures in a text, on the basis of the content of the dictionary (and other local grammars). One example of local grammars for Bulgarian and Serbian are local grammars for the recognition of complex tenses (for Serbian see (Vitas & Krstev, 2003). These grammars enable not only the recognition of a compound tense in the sentence, but also the transformation of the sequence of words, or the transformation of the tense.

5.4. Named entities

Local grammars can be applied in other ways also. For example, let us observe the example of annotation of named entities on the aligned texts of Verne's novel in the sense of (Chinchor et al., 1999).

As a first example consider the regular expression of the following form:

(<A+NProp+Top>+<E>) <N+Cur>

Its meaning is: extract from a text any sequence of tokens that can be interpreted as a numeral, simple or compound, expressed by digits or words, that is followed by an optional adjective derived from a toponym that is followed by an obligatory noun that represents a currency. When this pattern is applied to the Verne's text the examples presented in Appendix 2 are obtained.

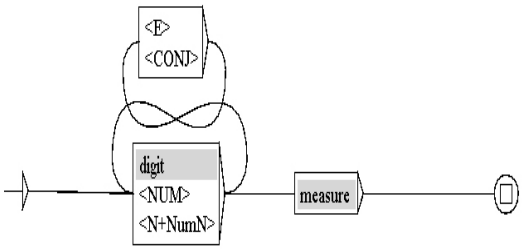

Figure 5: The graph *measure.grf* for the recognition of measure expressions

The annotation of named entities for some measures on the aligned texts of Verne's novel is more complex. The general expression for a measure is depicted by the graph *Measure.grf* in Figure 5 which describes it as a structure of a sequence of numbers written by words or digits followed by a measure indicator (kilometer, grade, mile, foot, etc).

Examples of sequences which correspond to this graph are *пет, шест или десет стъпки* in Bulgarian or in Serbian *hiljadu tri stotine osamdeset i dve milje* '*one thousand three hundred eighty two miles*'. The same graph refers to words that have the categories NUM (numbers) or N+NumN (number nouns) assigned in dictionaries of Bulgarian and Serbian. In the subgraph *digit* any sequence of digits is recognized. The difference between the Serbian and Bulgarian lexis of measures is described by the graph *measure* where the units of measure are named. Some examples of concordances extracted by the automaton in Figure 5 are given in Appendix 3. The graph produces the concordance lines that contain the number of segments where some entity appeared as well as the measure entity itself. Certain differences in recognition are a consequence of the phenomenon of regular derivation:
(seg 2256, seg 2280) bg: *двадесеттонен кораб* = sr. *brod od dvadeset tona = en. craft of twenty tons*
or inconsistency in the translation:
(seg 4397) bg. *трista и шестдесет градуса* = sr. *tri stotine šezdeset meridijana* = en. t*hree hundred and sixty degrees*.

6. Conclusion and further work

We have presented some techniques directed towards the establishment of a flexible and uniform method for morpho-syntactic annotation concerning not only single words but multi-word units, complex word forms and productive derivational rules. We have treated single words and continuous MWUs in a uniform way presenting them in a common inflexional dictionary format. We have applied morphological grammars for the morpho-syntactic annotation of unknown words that are derived by productive derivational rules, and local grammars for the recognition of the complex word forms and named entities.

- Further developments of the method include:
- Compilation of large and range balanced multilingual parallel corpus of Balkan and South Slavic languages;
- Development of large inflectional dictionaries including continuous multi-word units,
- Coverage of all productive and regular derivational rules by means of morphological grammars,
- Extensive coverage of complex word forms by means of local grammars,
- Analyzing the similar language phenomena in Balkan and South Slavic languages.

The further extension of the research is presupposed by the developing of equivalent language resources for other Balkan and South Slavic languages.

7. References

Brants, T. (2000) TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*, ANLP-2000, April 29 - May 3, 2000, Seattle, WA, pp. 224–231

Chinchor, N., Brown, E., Ferro, L. Robinson, P. (1999) *1999 Named Entity Recognition Task Definition* (version 1.4). Technical Report, SAIC, http://www.nist.gov/ speech/tests/ie-er/er_99/doc/ ne99_taskdef_v1_4.pdf

Clément Lionel and Éric de la Clergerie. (2005) MAF: a morphosyntactic annotation framework. In *Proc. of the Language and Technology Conference, Poznan, Poland*, pp 90--94.

Courtois, B., Silberztein M. (Eds.) (1990) *Dictionnaires électroniques du français*, Langue française 87, Paris: Larousse

Dumitriu, M. (2005) Grammaires de flexion du roumain en format DELA, Rapport interne 2005-02 de l'Institut Gaspard-Monge – CNRS

Gross, M. (1993) Local Grammars and Their Representation by Finite Automata. Data, Description, Discourse, *Papers on the English Language in honour of John McH Sinclair*, ed. by M Hoey. London: Harper-Collins. pp. 26-38

Erjavec, T. (2004) MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Fourth International Conference on Language Resources and Evaluation*, LREC'04

Ide, N., L. Romary, and E. Villemonte de la Clergerie, (2003). International standard for a linguistic annotation framework. In *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*. Edmonton. http://www.cs.vassar.edu/~ide/papers/ide-romary-clergerie.pdf

Koeva, S. (2004) Contemporary language technologies. In: *Laws of/for language*. Sofia, pp. 111- 135

Koeva, S. (2005) Inflection Morphology of Bulgarian Multiword Expressions. In: *Computer Applications in Slavic Studies – Proc.of Azbuki@net*, pp. 201-216, Sofia

Krstev, C., Vitas, D., Erjavec, T. (2004) Morpho-Syntactic Descriptions in MULTEXT-East - the Case of Serbian, *Informatica*, No. 28, The Slovene Society Informatika, Ljubljana, pp. 431-436

Krstev, C., Vitas, D., Savary A. (2006a) Prerequisites for a comprehensive Dictionary of Serbian Compounds. FinTAL, LNCS 4139, pp. 552-563

Krstev, C., Stanković, R., Vitas, D., Obradović, I. (2006b) WS4LR - a Workstation for Lexical Resources, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 1692-1697

Krstev, C. Vitas, D. (2007) Treatment of Numerals in Text Processing, in *Proceedings of 3nd Language & Technology Conference*, October 5-7, 2007, Poznań, Poland, ed. Zygmunt Vetulani, pp. 418-422

Kyriacopoulou, T., Mrabti S., Yannacopoulou, A. (2002) Le dictionnaire électronique des noms composés en grec moderne, *Lingvisticæ Investigationes* 25:1, Amsterdam/Philadelphia John Benjamins, pp. 7–28

Maurel, D., Krstev, C, Vitas, D., Koeva, S. (2007) Prolex: a lexical model for translation of proper names: Application to French, Serbian and Bulgarian. In *Slavic languages and French: formal approaches in contrastive studies, Bulag*, 32, Besancon, pp. 55-72

Przepiórkowski, A., Woliński, M. (2003) A Flexemic Tagset for Polish. In Erjavec, T., Vitas, D. (Eds.) *EACL workshop on Morphological Processing of Slavic Languages*, Budapest, pp. 33-40

Romary, L., Bonhomme P. (2000) Parallel Alignment of Structured Documents, In J. Véronis (Ed.) *Parallel text processing: Alignment and use of translation corpora*, Kluwer Academic Press, pp. 211-218

Savary, A. (2005) Towards a Formalism for the Computational Morphology of Multi-Word Units. In Vetulani (ed.) *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proceedings of the 2nd Language & Technology Conference.* Poznań, Poland, pp. 305-309

Vitas, D., Krstev, C. (2005) Regular derivation and synonymy in an e-dictionary of Serbian, *Archives of Control Sciences*, Volume 15(LI), No. 3, Polish Academy of Sciences, pp. 469-480

Vitas, D., Krstev C., Maurel, D. (2007) A note on the semantic and morphological properties of proper names in the Prolex project. In Sekine, Satoshi and Elisabete Ranchhod (Eds.): *Named Entities: Recognition, classification and use, Lingvisticæ Investigationes* 30 (1) pp. 115–133

Vitas, D., Krstev, C. (2003) Composite Tense Recognition and Tagging in Serbian. In Erjavec, T., Vitas, D. (Eds.) *EACL workshop on Morphological Processing of Slavic Languages*, Budapest, pp. 55 - 62

Vitas, D. (2004) Morphologie dérivationnelle et mots simples: Le cas du serbo-croate. *Syntax, Lexis & Lexicon-Grammar (Papers in honour of Maurice Gross)*, *Lingvisticæ Investigationes Supplementa* 24, Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 629-640

Appendix 1. Text fragment from Figure 1 in TMX format

<tu> <tuv xml:lang="BG" creationid="n506 " creationdate="20070801T123334Z">
<seg>Сигурно добре знае, че в Индия, която е английска земя, няма да е в безопасност. </seg>
</tuv>
<tuv xml:lang="SR" creationid="n506 " creationdate="20070801T123334Z">
<seg>On dobro zna da neće biti siguran u Indiji jer je to engleska zemlja. </seg>
</tuv>
<tuv xml:lang="FR" creationid="n506 " creationdate="20070801T123334Z">
<seg>Il doit bien savoir qu'il ne serait pas en sûreté dans l'Inde, qui est une terre anglaise. </seg>
</tuv>
<tuv xml:lang="EN" creationid="n506 " creationdate="20070801T123334Z">
<seg>He ought to know that he would not be safe an hour in India, which is English soil.</seg>
</tuv>

Appendix 2. Some of the entities extracted by the graph money.grf

n175 : Една пачка банкноти, възлизаща на огромната стойност от <*b_numex type="money"*>*петдесет и пет хиляди лири*<*e_numex*>, бе взета от масата на главния касиер на Банк ъф Ингланд
n175 : Svežanj novčanica u iznosu od <*b_numex type="money"*>*pedeset i pet hiljada livara*<*e_numex*> iščezao je sa stola glavnog blagajnika Engleske banke.
n175: Une liasse de bank-notes, formant l'énorme somme de <*b_numex type="money"*>*cinquante-cinq mille livres*<*e_numex*>, avait été prise sur la tablette du caissier principal de la Banque d'Angleterre.

n176 : ...в същия момент касиерът се е занимавал с вписването на приходи от <*b_numex type="money"*>три шилинга и шест пенса<e_numex> и че човек не може да държи всичко под око.
n176 : ... u tom trenutku blagajnik beležio primanje <*b_numex type="money"*>*tri šilinga i šest penija* <*e_numex*> i da se ne može na sve obratiti pažnja.
n176: ... à ce moment même, le caissier s'occupait d'enregistrer une recette de <*b_numex type="money"*>*trois shillings six pence*<*e_numex*>, et qu'on ne saurait avoir l'oeil à tout.

n2009 : ... меркантилна Англия продава годишно за <*b_numex type="money"*>*двеста и шестдесет милиона франка*<*e_numex*> от тази смъртоносна дрога, която се нарича опиум!
n2009 : ... trgovačka Engleska prodaje godišnje onu kobnu drogu nazvanu opijum za <*b_numex type="money"*>*dve stotine šezdeset hiljada franaka*<*e_numex*>!
n2009: ... la mercantile Angleterre vend annuellement pour <*b_numex type="money"*>*deux cent soixante millions de francs*<*e_numex*> de cette funeste drogue qui s'appelle l'opium!

n4342 : – Няма да си дам моята част от четири хиляди лири в облога – каза Андрю Стюарт, сядайки, – все пак ще получа <*b_numex type="money"*>*три хиляди деветстотин деветдесет и девет лири*<*e_numex*>.
n4342 : -- Ja svoj deo u opkladi ne bih dao pa da mi ko za njega daje <*b_numex type="money"*>*tri hiljade devet stotina i devedeset i devet livara*<*e_numex*> - reče Endrju Stjuart sedajući.
n4316: Je ne donnerais pas ma part de quatre mille livres dans le pari, dit Andrew Stuart en s'asseyant,-- quand même on m'en offrirait <*b_numex type="money"*>*trois mille neuf cent quatre-vingt-dix-neuf*<*e_numex*>!

Appendix 3. Some of the entities extracted by the graph measure.grf

<seg id="n50">... osamdeset četiri stepena Farenhajtovih ... осемдесет и четири градуса по Фаренхайт // 84°F
<seg id="n449">... dve hiljade osam sto tona ... две хиляди и осемстотин тона // 2 800 t
<seg id="n464">... sto šezdeset kilometara ... сто и шестдесетте километра // 160 km
<seg id="n493">... dve hiljade metara ... две хиляди метра // 2 000 m
<seg id="n839">... hiljadu do hiljadu sto milja ... хиляда до хиляда и сто мили // 1 000 - 1 100 knots
<seg id="n969">... sedamdeset i sedam stepeni ... седемдесет и седем градуса // 77 °
<seg id="n2689">... pet, šest, deset stopa ... пет, шест или десет стъпки // 5, 6, 10 feet
<seg id="n2961">... tri hiljade sedam stotina osamdeset šest milja ... три хиляди седемстотин осемдесет и шест мили // 3786 knots
<seg id="n3216">... sedam hiljada pet stotina dvadeset četiri engleske stope ... седем хиляди петстотин и осемдесет английски стъпки // 7524 feets
<seg id="n3664">... pola milje ... половин миля // 1/2 feet

Using inheritance and coreness sets to improve a verb lexicon harvested from FrameNet

Mark McConville and Myroslava O. Dzikovska

School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland
{Mark.McConville,M.Dzikovska}@ed.ac.uk

Abstract

We investigate two aspects of the annotation scheme underlying the FrameNet semantically annotated corpus — the inheritance relation on semantic types with its corresponding links between semantic roles of increasing granularity, and the specification of coreness sets of related semantic roles — against the background of our ongoing effort to harvest a lexicon of verb entries for deep parsing. We conclude that these aspects of the FrameNet annotation scheme *do* prove useful for reducing the complexity and ambiguity of verb entries, allowing for semantic roles of lower granularity for purposes of deep parsing, but need to be applied more systematically to make the lexicon usable in a practical parsing system.

1 Introduction

Semantically annotated corpora and wide-coverage semantic lexicons are an important resource for building NLP systems. They have been used to train shallow semantic parsers (Gildea and Jurafsky, 2002), provide paraphrases in question answering (Kaisser and Webber, 2007), and extend lexicons for deep parsing (Crabbé et al., 2006). All these applications use a 'frame-based' representation to express sentence semantics, where the semantic *type* corresponding to the meaning of a verb is related to its dependents by means of semantic *roles*. An essential task in building this representation is to make a connection between the surface form of the utterance and its semantics, usually by linking between syntactic and semantic structure.

Linking syntactic and semantic structure can be facilitated by a computational lexicon that describes possible mappings. McConville and Dzikovska (2007) report on an attempt to harvest a verb lexicon for deep linguistic processing from the FrameNet 1.3 semantically annotated corpus. We demonstrated that harvesting verb entries directly from annotations, as is done in the lexical entry files currently distributed with FrameNet, results in a number of subcategorisation frames which are unsuitable for inclusion in a computational lexicon used by a deep parser. We proposed a set of filtering rules to reduce the number of spurious subcategorisation frames generated by syntactic phenomena not directly captured in the FrameNet annotation.

In this paper we evaluate how this lexicon can be further improved by using two other aspects of the linguistic annotation underlying the corpus — the organisation of the semantic types (a.k.a. 'frames') and roles ('frame elements') into a hierarchy, and the specification of certain 'coreness sets' of related roles. The FrameNet ontology is very expressive and richly structured, with the aim of simplifying a number of reasoning tasks. However, we argue that FrameNet's level of role name granularity creates problems from the perspective of parsing, since it is traditionally assumed that verbs subcategorise for a relatively small number of arguments.

We first of all demonstrate that it is possible to use role inheritance to reduce the size of the role set (and hence the lexicon as a whole) without losing information, thus restricting the granularity of the semantic roles used in the output representation. We then describe an attempt to apply the coreness sets defined in the FrameNet ontology to eliminate ambiguity in lexical entries, making the FrameNet-based lexicon easier to use in a parsing system. We conclude that the FrameNet annotation scheme provides for useful mechanisms for reducing the complexity and ambiguity of verb entries, but needs to be applied more systematically to make the lexicon usable in a practical parsing system.

Section 2 provides some necessary background. **Section 3** discusses our investigations into the use of semantic role inheritance to reduce the vocabulary of roles invoked by arguments in verb entries. **Section 4** then turns to the topic of coreness sets in FrameNet, and the extent to which they can be used to eliminate redundancy in the harvested lexicon. Finally, **Section 5** discusses how our algorithms could be used in the future to benefit applications other than deep parsing.

2 Background

Regardless of the particular grammar formalism which they presuppose, lexicons used for parsing and semantic interpretation contain representations that map syntactic structure (a subcategorisation frame or a set of syntactic roles) to semantic structure (a predicate name and a set of arguments). For example, a lexical entry for the verb *move* would specify that: (a) the verb invokes a predicate which we might call 'motion'; (b) it subcategorises for a noun phrase subject which denotes the 'theme' (i.e. the object undergoing movement); and (c) it also subcategorises for a prepositional phrase complement headed by the preposition *to* which denotes the 'goal' (i.e. endpoint of the trajectory). This kind of information can be harvested automatically from semantically annotated corpora such as FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005) or OntoNotes (Hovy et al., 2006). The ultimate goal of our

project is to create a wide-coverage lexicon yielding representations that can be connected to the reasoning engine of a dialogue system. Thus, we chose FrameNet as our source for extracting lexical entries, since it includes an ontology which has already proved useful for information retrieval and question answering tasks (Surdeanu et al., 2003; Kaisser and Webber, 2007).

The FrameNet annotation scheme allows one to harvest a lexicon by reading the subcategorisation frames and their corresponding role assignments directly off the annotated sentences. The resulting lexicon contains 2,770 verb entries, each specifying a semantic type, an orthographic form, and a set of subcategorisation frames. Subcategorisation frames are sets of arguments, each of which specifies a syntactic role, syntactic category and semantic role.[1] Here is an example lexical entry for the verb *fry*, derived from an annotated sentence like *Matilde fried the catfish*:

$$\begin{bmatrix} \text{ORTH} & \langle \textit{fry} \rangle \\ \text{CAT} & \text{V} \\ \text{TYPE} & \texttt{Apply\_heat} \\ \text{ARGS} & \left\langle \begin{bmatrix} \text{ROLE} & \texttt{Ext} \\ \text{CAT} & \text{NP} \\ \text{ROLE} & \texttt{Cook} \end{bmatrix}, \begin{bmatrix} \text{ROLE} & \texttt{Obj} \\ \text{CAT} & \text{NP} \\ \text{ROLE} & \texttt{Food} \end{bmatrix} \right\rangle \end{bmatrix}$$

The subcategorisation frame lists two arguments, one for each annotated dependent in the sentence.

While collecting such entries is straightforward on the surface, not all of them would be usable with a deep parser. To begin with, all entries have to correspond to "canonical" syntactic subcategorization frames - i.e. to indicative mood, direct word order entries, and include only syntactic complements but not modifiers. Entries for other constructions, such as passives and clefts, are normally derived by syntactic transformations and are not included in the lexicon. We addressed these issues previously (McConville and Dzikovska, 2007; McConville and Dzikovska, 2008), developing methods to remove such spurious entries from the lexicon.

Secondly, we need to consider how well the syntax-semantics mappings harvested from the corpus fit with the representations traditionally used for parsing. We observed that the representations in the extracted entries manifest at least one significant difference in this respect. While there is no easily definable "canonical" representation for semantic roles, deep parsers, generally speaking, assume that the target semantic representation utilises a relatively small set of roles. There are several reasons for this. Firstly, restricting the vocabulary of semantic roles is convenient from a representational perspective — many existing lexicons are hierarchical (Copestake and Flickinger, 2000; McConville, 2006), and having a large number of distinct roles may make the lexicon less compact because it offers fewer opportunities for re-use through inheritance. Secondly, it has been proposed that the syntactic and semantic behaviour of verbs is correlated (Levin, 1993), and can be mediated through a small set of 'thematic roles', as for example encoded in the VerbNet lexicon (Kipper et al., 2000).

Finally, disambiguating between a large number of roles may require world knowledge and pragmatic information which is difficult to obtain and integrate in a domain-independent way. For example, the FrameNet semantic type `Closure` defines two distinct roles which can be denoted by the direct object of a transitive verb: `Container_portal` (e.g. *John closed **the tent flap***), and `Containing_object` (e.g. *Mary buttoned **her coat***). Human annotators are able to distinguish these roles based on common sense knowledge, and whilst it is true that such distinctions may be important for certain reasoning tasks, a deep parser would find this kind of ambiguity extremely difficult to resolve. Thus, a more compact roleset may be necessary to reduce the ambiguity in parsing and semantic interpretation.[2]

The importance of having a relatively small set of basic semantic roles has not been lost on the creators of FrameNet. Indeed, a lot of recent effort (between versions 1.1 and 1.3) has gone into organising the semantic types in the FrameNet ontology into an inheritance hierarchy and, in particular, into linking the fine-grained roles of child types with the more generic roles of their parent types. In addition, a number of 'coreness sets' of semantic roles have been specified, the idea being that only one member of a coreness set need be explicitly invoked in a well-formed, non-elliptical sentence, and hence that these roles are equivalent in some way. In the rest of this paper we describe how we used inheritance and coreness sets to eliminate redundancy in both the vocabulary of semantic roles and in the verb entries themselves.

As our general evaluation metric, we take the reduction in the number of individual roles and the reduction in the number of subcategorisation frames per verb entry in the lexicon. For comparison, we looked at two other lexicons: VerbNet, a lexicon of English verbs that aims to have a complete coverage of syntactic alternations for each verb covered, and the TRIPS lexicon (Allen et al., 2007) — a multi-domain lexicon used with a wide-coverage deep grammar. These lexicons were developed independently, but share the aim of explicitly representing the connections between syntax and semantics, with VerbNet focusing more on complete coverage, and TRIPS focusing on practical parsing applications that require syntactic and semantic disambiguation. Thus, while there is no way of determining the 'ideal' number of roles per se, comparison with these lexicons can give us some insight in the complexity or redundancy of the FrameNet-based lexicon compared to lexicons intended for parsing.[3]

The initial lexicon harvested from FrameNet (McConville

[1] We extracted this lexicon independently, but FrameNet contains an analogous set of lexical entries as part of the distribution, which we could have used as a starting point in the same way.

[2] Additional information can be brought in at a post-processing stage, linking the more generic semantic representation with more specific knowledge representation (Dzikovska et al., 2007).

[3] The various lexicons are not completely independent, in the sense that TRIPS contains an ontology of concepts inspired by an early version of FrameNet (Dzikovska et al., 2004), and it contains entries extracted from VerbNet (Crabbé et al., 2006). However, all entries were hand-edited to ensure that they conform to the independently developed lexicon design.

and Dzikovska, 2007) contains 9,180 subcategorization frames, invoking 362 distinct semantic types, and arguments invoking 441 distinct semantic role labels, an average of 1.2 semantic role labels per semantic type. In comparison with other deep verb lexicons, this ratio of roles to types is quite high. The TRIPS lexicon contains verb entries invoking 284 distinct semantic types and arguments invoking 48 distinct semantic roles, yielding a ratio of 0.17 roles per semantic type. Similarly, the VerbNet lexicon has 395 verb classes, with arguments instantiating just 33 distinct semantic/thematic roles, giving a ratio of 0.084 roles per verb class. In addition, the FrameNet-based lexicon contains 3.3 subcategorisation frames per verb entry, compared to 2.8 in VerbNet and 1.3 in TRIPS.[4]

3 Using inheritance to reduce the role set

We first consider how the inheritance relation encoded in the FrameNet ontology can be used to reduce the size of the vocabulary of semantic roles.

The FrameNet ontology of semantic types is organised into an inheritance hierarchy, where child types are connected to their parents by means of an `Inheritance` relation. For example, this relation partitions the `Motion` semantic type (encoding events involving a theme traversing a path) into a number of more specific subtypes such as `Self_motion` (the theme is a living being, acting under its own volition), `Fluidic_motion` (the theme is a fluid), etc. All the semantic roles associated with a parent type must be implemented by some role of each child type. For example, two of the roles associated with `Motion` are `Source` (start of the trajectory) and `Goal` (end of the trajectory). These roles are implemented directly by all child types of `Motion` using roles of the same name. On the other hand the `Theme` role associated with the `Motion` type is implemented by different roles in subtypes: in `Self_motion` it is implemented by `Self_mover`, in `Fluidic_motion` by `Fluid`, and so on. In addition, child types can introduce new roles which are *not* linked to roles of parent types.

The existence of this inheritance relation and its associated links between parent and child roles has important implications for the vocabulary of semantic roles in the lexicon we harvested from FrameNet. For example, the transitive verb *dismiss* invokes the FrameNet semantic type `Firing`, and its subject and object instantiate the associated semantic roles `Employer` and `Employee` respectively, hence the following subcategorisation frame:

(1) `Sbj:Employer Obj:Employee`

However, the semantic type `Firing` is subsumed by the parent type `Intentionally_affect` in the FrameNet ontology, with the `Employer` role linked to the superrole `Agent` and the `Employee` role linked to the `Patient` superrole. Thus, an alternative way of representing the transitive subcategorisation frame for *dismiss*, using the information contained in the inheritance hierarchy, is:

(2) `Sbj:Agent Obj:Patient`

Note that the semantic roles specified in this lexicon are much more generic, and are similar to the kinds of role names used in the VerbNet and TRIPS lexicons.

The aim of the first part of our project was to investigate the extent to which we can use information about supertypes and 'superroles' in the FrameNet 1.3 ontology to decrease the number of distinct semantic roles invoked by arguments in the harvested lexicon, thus creating a less redundant verb lexicon for deep parsing.

3.1 Methodology

We went through each argument of each subcategorisation frame of each verb entry in the harvested lexicon and, where the entry's semantic type was linked to some parent type in the FrameNet ontology and the argument's semantic role was linked to some role of the parent type, we replaced the original role with the superrole. We repeated this until we reached the root type in the ontology, which in this case involved five cycles (i.e. the maximum depth of the relevant part of the inheritance hierarchy is 5). In the cases where a role is linked to two or more distinct superroles (because of multiple inheritance in the FrameNet ontology), we included all of them.

3.2 Results

The results are presented in Table 1 in the 'full lexicon' column. Each row represents a level of recursion, i.e. '0' means that no supertypes are taken into account, '1' means that we move one level up the hierarchy etc. The first column represents the number of distinct semantic role labels across the entire lexicon at each cycle, and the second column represents represents the number of distinct types of subcategorisation frame in the lexicon (where a subcategorisation frame is abstracted to a set of semantic roles).

Thus, taking the lexicon we harvested from FrameNet as a whole, we can reduce the number of distinct semantic role labels by 21%, from 441 to 347. The five most common roles which are the beneficiaries of this process are presented in Table 2.

Note that the number of distinct role labels, 347, still appears to be very high in comparison with the selection found in other deep verb lexicons like TRIPS and VerbNet. In addition, Table 2 demonstrates that, although the three most popular roles to be introduced are the generic roles `Theme`, `Patient` and `Agent`, familiar from both the VerbNet and TRIPS lexicons and from mainstream theories of thematic roles, there are still some overly specific roles in evidence, for example `Communicator` and `Sought_entity`,

We hypothesised that the very small reduction in the number of semantic roles is a function of the incomplete nature of the inheritance relation in the FrameNet ontology. Recall that the FrameNet 1.3 ontology contains 362 verbal types. However, a large proportion of these, 145, are 'orphan types', in the (strong) sense that they are not linked to any other type in the ontology, neither as child nor as

[4] Note that the TRIPS figure is significantly lower in part because the TRIPS lexicon has been built based on the subcategorisation frames attested in spoken dialogue corpora, so it does not contain many frames that are included in VerbNet but only rarely appear in speech and dialogue.

	full lexicon		restricted lexicon	
cycle	roles	frames	roles	frames
0	441	1256	289	807
1	364	1129	196	653
2	348	1083	177	596
3	347	1083	176	596
4	347	1083	176	596
5	347	1083		

Table 1: Results of the inheritance experiments

full lexicon		restricted lexicon	
frequency	role	frequency	role
1254	`Theme`	1843	`Agent`
1150	`Patient`	1486	`Theme`
777	`Agent`	1189	`Patient`
709	`Communicator`	827	`Communicator`
225	`Sought_entity`	591	`Goal`

Table 2: Most common role labels in the resulting lexicon

parent. In order to determine whether the disappointingly small reduction in distinct semantic roles as we climb the hierarchy is a result of the existence of these orphan types, we eliminated all verb entries from the harvested lexicon which invoke one of the 145 orphan types, and repeated the process.

Our restricted lexicon now contains 1,729 verb entries invoking 217 distinct semantic types. There are 6,253 subcategorisation frames distributed across these entries. The results of substituting more general roles for more specific ones, according to the inheritance relation underpinning the FrameNet 1.3 ontology, are presented in the 'restricted lexicon' half of Table 1.

The five most common roles which are now the beneficiaries of this process are presented on the right hand side of Table 2.

Thus, assuming the subset of the FrameNet-harvested lexicon which only includes types which are incorporated into the inheritance relation underpinning the FrameNet 1.3 ontology, we can reduce the number of distinct semantic role labels by 39%, from 289 to 176. This is significantly higher than the 21% reduction we managed using the full lexicon, thus supporting our hypothesis that the more 'connected' the FrameNet inheritance relation is, the more useful it will be in allowing us to harvest a deep verb lexicon with a manageable set of semantic roles. The fact that only 975 of the 2,770 verb entries in the harvested lexicon have semantic types which are rooted in either the `State` or `Event` supertypes shows that the FrameNet ontology still has a way to go in this respect.

4 Using coreness sets to filter subcategorisation frames

As discussed in the introduction, after filtering out modifiers and frames derived from non-canonical usages of target verbs, the lexicon we harvested from FrameNet contained 9,180 subcategorisation frames, distributed among 2,770 verb entries.

One interesting feature of the FrameNet ontology which we have not considered until now involves the specification of certain kinds of dependency between the semantic roles associated with a particular semantic type. For example, in certain semantic types, a particular subset of the semantic roles may be grouped together in a 'coreness' set, only one of which need be expressed in order to produce a complete, non-elliptical sentence. The most prevalent example of this involves the following semantic roles within the `Motion` semantic type and its subtypes:

- `Source` (e.g. *from Cairo*)
- `Goal` (*to Khartoum*)
- `Path` (*down the Nile*)
- `Area` (*around the country*)
- `Direction` (*towards Alexandria*)

The fact that these five roles are grouped together into a coreness set, captures the fact that they are in some sense equivalent, or that they instantiate the same underlying role, that of "trajectory".

The existence of coreness sets has implications for lexical concision. For example, the harvested lexicon contains 115 entries invoking the `Self_motion` semantic type, and these entries involve *eleven* distinct types of subcategorisation frame (ignoring syntactic categories) with the `Self_mover` role as subject and these 'trajectory' roles as oblique dependents, for example:

(3) `Sbj:Mover Dep:Source`
`Sbj:Mover Dep:Goal`
`Sbj:Mover Dep:Source Dep:Goal`
...

However, if we assume that the trajectory roles are actually just alternative realisations of the same underlying semantic role, then we can condense all these frames into just the one, where the Kleene star denotes an unbounded number of instances of the specified argument type:

(4) `Sbj:Theme Dep:Trajectory*`

The FrameNet 1.3 ontology specifies 210 coreness sets for 174 verbal semantic types. Each coreness set brings together an average of 2.5 semantic roles. The aim of the second part of our project was thus to investigate to what extent we can use the coreness sets defined in the ontology to consolidate the harvested lexicon, in terms of reducing the number of subcategorisation frames that need to be specified.

4.1 Methodology

We proceeded in two stages. First of all, we went through every argument of every subcategorisation frame of every verb entry and, where the argument's semantic role was part of some relevant coreness set, we replaced the semantic role name with the coreness set name. Then we went through every verb entry and eliminated duplicate frames, assuming that two frames are identical if and only if they have the same arguments, and that two arguments are identical just in case they have the same syntactic role, syntactic category and semantic role/coreness set.

4.2 Results

The first stage of the procedure, where we replaced semantic role labels with relevant coreness sets, affected 1,542 of the 2,770 verb entries in the lexicon, and 5,954 of the subcategorisation frames found in these entries. After eliminating duplicate subcategorisation frames, we were left with 7,804 frames across the lexicon as a whole (down from 9,180).

Of the 7,804 subcategorisation frames left in the lexicon, 1,253 have potentially duplicate arguments, i.e. where two or more arguments have semantic roles from the same coreness set. Thus, we next eliminated all duplicate arguments from individual subcategorisation frames, resulting in a decrease in the total number of arguments across all extant subcategorisation frames, from 16,795 to 16,406. Finally, after again eliminating duplicate subcategorisation frames from within each verb entry, the lexicon contained 7,672 frames across the 2,770 verb entries. This constitutes an average of 2.8 subcategorisation frames per entry and a reduction of 16% on the original number of 9,180.

4.3 Evaluation

We wanted to evaluate whether the use of coreness sets to consolidate pairs of subcategorisation frames corresponds with linguistic intuitions about which subcategorisations frames in a verb entry are really 'equivalent' and hence 'collapsible'. To this end, we selected 100 random cases where our procedure had used coreness sets to make a judgment that two distinct subcategorisation frames were essentially the same. We ensured that our sample contained only one instance from each semantic type, so as to counteract the bias in the FrameNet corpus whereby certain types include more verbs than others and certain verbs have been more fully annotated. Where necessary, we referred to the equivalent verb entries in VerbNet and the TRIPS lexicon.

Of the 100 entries chosen, 17 involved variations of the 'trajectory' coreness set discussed above, associated with an assortment of motion, orientation and spatial extension predicates. It is important to note, first of all, that this coreness set is independently motivated, for example in the ontology of paths outlined in Jackendoff (1983), where source, goal, and other unbounded path expressions are treated as equivalent in the sense that they are alternative realisations of one and the same thematic role in conceptual structure. We verified that in all 17 cases, the coreness set *did* in fact correlate with this linguistic intuition, and hence that combining the two subcategorisation frames was valid. Take for example, the following subcategorisation frames of the verb *buzz* from the `Motion_noise` semantic type:

(5) `Sbj:NP:Theme Dep:PP:Goal`
`Sbj:NP:Theme Dep:PP:Path`

The first of these includes a `Goal` argument (e.g. *buzz into the room*) and the second a `Path` (e.g. *buzz across the room*). Since the FrameNet ontology lists these in a coreness set for `Motion_noise`, the two subcategorisation frames are combined into the following unified representation:

(6) `Sbj:NP:Theme Dep:PP:Goal/Path`

This decision corresponds with our linguistic intuitions about the argument structure of the verb *buzz*, which subcategorises for an unbounded number of trajectory expressions (e.g. *The fly buzzed from the doorway across the room to the window*). We used similar reasoning with the other 16 instances involving the 'trajectory' coreness set in our sample.

Of the remaining cases in our sample, a substantial number (around 40) involve what can loosely be termed 'part-whole' alternations in the relevant argument. For example, the verb *claw* from the `Manipulation` type subcategorises for subjects with two distinct semantic roles, `Agent` and `Bodypart_of_agent`, related through a coreness set. These two usages are exemplified in the following two sentences:

(7) Jane clawed at his back
Fingers clawed at his back

Other examples are somewhat more abstract. For example, the verb *eclipse* from the `Surpassing` type subcategorises for two kinds of subject in the FrameNet lexicon, `Profiled_item` and `Profiled_attribute`, again related through a coreness set, and where the latter can be approximated as a 'part' (or possibly 'feature') of the former:

(8) John eclipsed Mary
John's talent eclipsed Mary's

Again, the consolidation of these arguments was judged to be linguistically valid, in part because VerbNet treats them as encoding the same thematic role (i.e. `Theme1`).

Other coreness sets which occurred repeatedly throughout our sample involved agent-cause alternations (e.g. *John/The blackout disabled the alarm system*) and speaker-medium alternation (e.g. *The critics/survey labelled her a has-been*). Again the intuitiveness of these coreness sets is supported by VerbNet thematic roles.

However, there were at least ten cases where the coreness sets lead to an invalid consolidation of arguments, in general caused by the fact that FrameNet syntactic information, and hence our lexical entry extraction procedure, does not distinguish between preposition phrases headed by different prepositions. For example, consider the following two example sentences involving the verb *jab* from the `Cause_impact` type:

(9) Mary jabbed John with a bayonet
Mary jabbed a bayonet at John

In both these sentences, *John* would be annotated as an `Impactee` and *a bayonet* as an `Impactor`. Since these two roles are part of the same coreness set, the subcategorisation frames underlying both sentences are consolidated into the following unified representation:

(10) `Sbj:NP:Agent`
`Obj:NP:Impactor/Impactee`
`Dep:PP:Impactor/Impactor`

This is clearly undesirable, since it leads to an unnecessary level of ambiguity for a parser, a conclusion reinforced by the fact that VerbNet treats the impactee and impactor arguments with distinct thematic roles (i.e. `Destination` and `Instrument` respectively).

It is worth dwelling a little on the possible reasons for FrameNet annotators formulating such an obviously unintuitive coreness set. In previous work (McConville and Dzikovska, 2007), we have noted the tendency to incorporate *all* uses of a particular verb into the same frame, even when syntax disagrees. For example, take the two uses of the verb *rip* in the following sentences:

(11) John ripped his trousers below the knee
John ripped the top off his packet of cigarettes

In both sentences, annotators have judged that the target verb *rip* evokes the `Damaging` frame, which has two important 'core' roles — `Agent` (i.e. the 'ripper') and `Patient` (the object that gets ripped). In this respect, annotation of the first sentence is simple — *John* is the `Agent`, *his trousers* is the `Patient`, and the prepositional phrase *below the knee* is assigned to a 'non-core', locative role called `Subregion`.

Assuming that the use of the target verb *rip* in the second sentence also involves the `Damaging` frame causes problems however — *the top*, is assigned to the non-core `Subregion` role, despite being realised as a (syntactically obligatory) direct object. Thus, in this case the syntactic generalisation that subjects and onjects realise core roles is overuled in favour of keeping all uses of the target verb within the same frame. A more appropriate analysis would have been to assign the use of the target verb in the second sentence to the `Removing` frame.

Considering again the examples involving the target verb *jab* in (9), we see that similar forces are at work. The hypothesised reason for grouping roles into coreness sets is where a number of distinct roles are realised by the same syntactic role — in this case, the direct object can realise either the `Impactee` (i.e. *John*) or the `Impactor` (i.e. *a bayonet*), so the formulation of a coreness set `Impactor/Impactee` makes sense. Note however that this is purely an artifact of the decision to treat both uses of the target verb *jab* as evoking the same frame. If the second sentence were treated as involving the `Cause_motion` frame, the undesirable coreness set would not have been formulated.

Therefore, we can conclude that, although the FrameNet coreness sets correspond in the vast majority of cases with valid underlying thematic roles, there are a number of problematic cases, at least some of which involve target verbs being assigned to suboptimal frames by annotators.

Note that information about the particular kind of preposition which can head a given PP argument is often considered to be a part of a subcategorisation frame, especially for deep parsers (c.f. the commonly used PFORM feature). If such information were available in FrameNet annotation, this would have the side effect of avoiding some of the problems caused by this kind of unintuitive coreness set, since the argument structures derived from the two sentences in (9) would not be identical — the first would have a `PPwith` dependent, whereas the second would have a `PPat`. However, it would also make it more difficult to merge arguments from some of the intuitive coreness sets such as that involving trajectory arguments, since these can be introduced by a large variety of prepositions.

In the future, we are planning to improve our lexicon extraction algorithm so that prepositions are taken into account in extracting and differentiating subcategorisation frames. This would require a more detailed investigation which arguments can be merged despite using different prepositions, and in which cases they should be kept separate. One possible solution is suggested by the approach taken in the VerbNet. The arguments in the VerbNet subcategorisation frame can either be associated with a single preposition (such as *with*), or with a class of prepositions (such as *P:loc* corresponding to a set of locative prepositions). This encodes the intuition that in some cases the preposition is fixed by the verb, and therefore 'meaningless', while in other cases the preposition is 'meaningful' in that it corresponds to a specific predicate (*on*, *in*, *under*) and can be drawn from a large set of possibilities. We therefore are considering using the FrameNet corpus data to see if a preposition associated with a given role appears to be fixed, or can be drawn from a larger set, and using this as a basis for making the distinction between meaningless and meaningful prepositions associated with coreness sets.

5 Discussion

In this paper, we argued that for purposes of parsing and semantic interpretation, a less specific set of semantic roles would ease lexicon construction and disambiguation. Consider an analogy with word sense distinctions: Palmer et al. (2004) argue that different levels of granularity are

needed for different applications. For example, information retrieval may require coarser distinctions, at the level of PropBank sense groupings, while machine translation may require much more fine-grained distinctions, such as those found in WordNet (Miller, 1995). Similar reasoning can be applied to semantic roles: coarser distinctions, such as the argument labelling assumed in PropBank (i.e. ARG0, ARG1, etc.), may be the easiest to disambiguate and annotate; thematic roles as used in VerbNet (i.e. AGENT, THEME, etc.) may provide an appropriate level of generalisation when linking syntactic and semantic structure; and the fine distinctions encoded in FrameNet (i.e. COOK, FOOD, etc.) may be useful for reasoning. Ideally, these different levels could be mapped to each other, similarly to the way WordNet senses are linked to VerbNet and PropBank entries. Our study is a first step in evaluating to what extent the different levels of generalisation could be linked in FrameNet through the use of features defined in its ontology, and in attempting to automatically derive a set of semantic roles and lexical entries at lower granularity.

While our research is primarily centered on the needs of a deep parser and lexicon, the algorithms we developed could also contribute to ongoing research on linking various lexical resources and annotated corpora, for both manual and automatic linking approaches. In case of manual linking, the SemLink project[5] aims to develop correspondences between the semantic types and roles underlying PropBank, VerbNet and FrameNet. In the future, we plan to compare results of our automatic procedure with the correspondences made by human coders. Assuming that there is sufficient agreement, this automatic approach could be adapted in the future to reduce the need for manual linking. For automatic linking, Kwon and Hovy (2006) propose an automatic algorithm for aligning role names between semantic lexicons, which achieves around 78% accuracy in aligning FrameNet and PropBank roles based on corpus evidence. It may be interesting to consider whether using either inheritance or coreness set information could improve the accuracy of the alignment algorithm.

Finally, statistical parsers and semantic role labellers (Gildea and Jurafsky, 2002) could benefit from having a smaller set of semantic roles, because this would reduce the data sparsity problem. Using the hierarchy to reduce the role set could be useful under the circumstances, without loss of data. It is admittedly less clear how the coreness set information could be used, but this too may be worth exploring if it could be utilised as a way of backing off to more general role names in a statistical model.

6 Conclusion

The aim of the project reported in this paper was to take a verb lexicon harvested fairly directly from the FrameNet semantically annotated corpus, and to apply some of the mechanisms within the FrameNet ontology to make this lexicon more effective for use with a deep parser. We argued that the lexicon would be improved with a more concise and generic role set, because it will simplify making links between syntax and semantics in the lexical entries. We examined: (a) the inheritance relation on semantic roles, and the corresponding links between semantic roles of increasing granularity, as a means of reducing the size of the vocabulary of roles across the lexicon as a whole; and (b) the coreness sets of related semantic roles specified within the FrameNet ontology, with the aim of consolidating subcategorisation frames within individual verb entries. In both cases, we concluded that the annotation scheme provide useful, though not perfect, mechanisms for our purposes. This is in part due to the fact that the relevant aspects of the scheme are not always applied in systematic manner across the FrameNet ontology. Making this part of the FrameNet annotation more consistent could benefit not only our application, but also applications that support linking between different resources, and potentially semantic role labelling applications.

[5] http://verbs.colorado.edu/semlink

Acknowledgements

The work reported here was supported by grants N000140510043 and N000140510048 from the Office of Naval Research.

7 References

James Allen, Myroslava Dzikovska, Mehdi Manshadi, and Mary Swift. 2007. Deep linguistic processing for spoken dialogue systems. In *Proceedings of the ACL'07 Workshop on Deep Linguistic processing*, pages 49–56.

C. F. Baker, C. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL'98, Montreal*, pages 86–90.

Ann Copestake and Daniel Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of LREC'00, Athens, Greece*, pages 591–600.

Benoit Crabbé, Myroslava O. Dzikovska, William de Beaumont, and Mary D. Swift. 2006. Increasing coverage of a domain independent dialogue lexicon with VerbNet. In *Proceedings of the Third International Workshop on Scalable Natural Language Understanding (ScaNaLU 2006)*.

Myroslava O. Dzikovska, Mary D. Swift, and James F. Allen. 2004. Building a computational lexicon and ontology with FrameNet. In *Proceedings of the LREC'04 Workshop on Building Lexical Resources from Semantically Annotated Corpora*.

Myroslava O. Dzikovska, Mary D. Swift, and James F. Allen. 2007. Linking semantic and knowledge representations in a multi-domain dialogue system. *Journal of Logic and Computation, Special Issue on Natural Language and Knowledge Representation*.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press.

Michael Kaisser and Bonnie Webber. 2007. Question answering based on semantic roles. In *Proceedings of the ACL'07 Workshop on Deep Linguistic processing*.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of AAAI'00*.

Namhee Kwon and Eduard Hovy. 2006. Integrating semantic frames from multiple sources. In *Proceedings of CICLing'06*.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Mark McConville and Myroslava O. Dzikovska. 2007. Extracting a verb lexicon for deep parsing from FrameNet. In *Proceedings of the ACL'07 Workshop on Deep Linguistic processing*, pages 112–119.

Mark McConville and Myroslava O. Dzikovska. 2008. Evaluating complement-modifier distinctions in a semantically annotated corpus. In *Proceedings of LREC'08*.

Mark McConville. 2006. Inheritance and the CCG lexicon. In *Proceedings of EACL'06*, pages 1–8.

G. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(5).

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, pages 49–56, Boston, Massachusetts, USA, May 2 - May 7.

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

Mihai Surdeanu, Sanda M. Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL-03*, pages 8–15.

An Entailment-based Approach to Semantic Role Annotation

Voula Gotsoulia
Department of Language and Linguistics
University of Essex
Colchester, United Kingdom
pghots@essex.ac.uk

Abstract

In this paper, we consider an entailment-based view of the notion of semantic role and propose an annotation scheme for associating arguments with fine-grained, prototypical properties entailed by the semantics of predicators. We empirically investigate the potential of incorporating an entailment-based layer to semantic role corpus annotation for acquisition and formalization of linguistic knowledge at a general syntax-semantics interface.

1. Introduction

Large-scale lexical semantic resources that provide relational information about lexical items are at the heart of current research in natural language processing (NLP). In particular corpora with predicate-argument structure annotation constitute the basis for development of semantic parsing algorithms that automatically identify the semantic roles conveyed by sentential constituents [8] furnishing a shallow semantic level of text interpretation. Yet, on a parallel track, corpora with semantic role annotation are essential data for acquisition of linguistic knowledge at a principled syntax-semantics interface. Formalization of corpus-induced linking information at a suitable level of generality or abstraction can be useful in a variety of ways. Besides providing insight into lexical semantic phenomena, generalizations over specific mappings of argument structure to syntactic form can be incorporated into alternative systems e.g. applying meta-learning strategies such as active learning for semi-automatic acquisition of extended sets of annotated data [6]. Extraction of linking regularities across various predicate senses and constructions can thus be used as a remedy for the severe problem of sparse data in lexical semantic corpus annotation (i.e. the insufficient coverage of specific senses and constructions within sensible sizes of manually annotated data).

In this context, we consider the implications of different theoretical approaches determining essential design aspects of semantic role annotation. Relying on the insights of Dowty's [3] theory of proto-roles we propose an annotation scheme that associates arguments with prototypical properties entailed by the semantics of predicates. We discuss the potential of implementing an entailment-based approach for extraction of general information about possible syntax-semantics mappings.

2. Corpora and Semantic Roles

Corpora with semantic role annotation available for English represent distinct approaches to the notion of semantic role.

The Proposition Bank (PropBank) [12] is a one million word corpus in which predicate-argument relations are annotated for every occurrence of every verb in the Wall Street Journal part of the Penn Treebank [15]. Different verb senses are distinguished mostly on syntactic grounds. For each sense, arguments are numbered sequentially. Although the same argument labels are used for all verbs (ARG0, ARG1, …, ARG5), these labels are defined on a per-verb basis, i.e. they have a verb-specific meaning and are only consistent across syntactic alternations of a single verb sense. Example PropBank annotations:

(1) [ARG0 Blue-chip consumer stocks] [*rel* provided] [ARG1 a lift] to [ARG2-TO the industrial average].

(2) In addition, [ARG0 the bank] has an option to [*rel* buy] [ARG1 a 30% stake in BIP] from [ARG2-FROM Societe Generale] [ARGM-TMP[1] after Jan.1, 1990] at [ARG3-AT 1,015 francs a share].

PropBank makes no attempt to formalize the semantics of the role labels it employs. This is particularly clear with higher-numbered labels: ARG2, for instance, indicates *benefactive* with the verb *provide* (1), while with the verb *buy* (2) it indicates *source*. Lower-numbered labels denote various roles as well, although they are less arbitrary across verbs: ARG0 generally corresponds to traditional agents, experiencers, certain types of theme, etc. that surface as subjects of transitive verbs and a class of intransitives called unergatives; ARG1 is assigned to objects of transitive verbs and subjects of unaccusatives and is the equivalent of patient, theme, etc. Nonetheless, there are still inconsistencies even for ARG0 and ARG1. In effect, since no consistent mapping is ensured between a label and a semantic role, PropBank labels do not lend themselves to any formalization of linguistic knowledge. Currently, an attempt[2] is made to map argument labels to semantically more coherent roles by ensuring their consistency within verb classes defined by VerbNet[3].

[1] ARGM indicates adjuncts. It is generally followed by one of a set of functional tags denoting the role of the element in question, e.g. ARGM-LOC for locatives, ARGM-TMP for temporals, etc.

[2] http://verbs.colorado.edu/semlink/

[3] VerbNet [13] is an implementation of Levin's [14] verb classes defined on the basis of the ability of verbs to participate or not in pairs of syntactic frames representing alternations in the expression of their arguments (so-called *diathesis alternations*).

FrameNet[4], on the other hand, is creating an online semantic lexicon based on Fillmore's [4] theory of frame semantics. It describes word meaning in terms of underlying conceptual structures encoded in the form of *frames*, i.e. schematic representations of stereotyped situations capturing real-world knowledge. Each frame is associated with a set of lexical items that *evoke* it and a set of roles (*frame elements*) corresponding to the participants in the designated prototypical situation. A distinction is made between *core* and *non-core* (marginal) roles.

FrameNet includes manually annotated example sentences from the British National Corpus providing additional layers of phrase structure and grammatical function annotation [5]. It also includes two small corpora of full-text annotation intended to facilitate statistical analysis of frame-semantic structures. Currently, it contains more than 625 frames covering more than 8,900 lexical units. The following sentence exemplifies the SUPPLY frame in which "a SUPPLIER gives a THEME to a RECIPIENT to fulfill a need or purpose (PURPOSE_OF_RECIPIENT) of the RECIPIENT".

(3) [SUPPLIER Russia] will *provide* [RECIPIENT Syria] [THEME with equipment and high technology] [PURPOSE_OF_RECIPIENT for this peaceful purpose].

FrameNet avoids the difficulties of attempting to pin down a small set of general roles. Instead, frame elements are defined locally, i.e. in terms of frames that are situated in semantic space by means of directed (asymmetric) relations. Each frame-to-frame relation generally associates a less dependent or more general frame (Super_frame) with a more dependent or less general one (Sub_frame)[5]. The formulation of generalizations about possible mappings of frame elements to grammatical functions (i.e. linking generalizations) essentially relies on the establishment of a frame hierarchy and a theory of frame element identities or analogs across frames.

3. An Entailment-based Approach to Semantic Role Annotation

3.1 Background

A substantially different approach to semantic roles is put forth by Dowty [3]. Refraining from the idea of semantic roles as discrete categories[6] Dowty describes argument selection (i.e. the question of what principles determine which argument of an *n*-place relation denoted by a predicate is expressed by which grammatical relation) in terms of fine-grained, prototypical properties entailed by the semantics of predicates. He gives the following lists of entailments classified in two cluster concepts that he calls Proto-Agent and Proto-Patient:

(4) Contributing properties for the Agent Proto-Role:
- a. volitional involvement in the event or state
- b. sentience (and/or perception)
- c. causing an event or change of state in another participant
- d. movement (relative to the position of another participant)
- e. (exists independently of the event named by the verb)

(5) Contributing properties for the Patient Proto-Role:
- a. undergoes change of state
- b. incremental theme
- c. causally affected by another participant
- d. stationary relative to movement of another participant
- e. (does not exist independently of the event, or not at all)

In Dowty's model, Proto-Agent and Proto-Patient are conceptualized as suitable abstractions that define a semantic continuum mapping directly onto syntax on the basis of a numerical comparison linking principle[7]. That is, no unifying semantics is implied for either of the lists in (4)-(5). Semantic properties are associated with grammatical categories in a many-to-one fashion and arguments are allowed to have different degrees of membership to both proto-roles. It thus misinterprets Dowty to speak of a particular argument of a predicate as *the* Proto-Agent or *the* Proto-Patient.

In a related vein, Wechsler [19, 20] analyzes argument structure in terms of universal semantic primitives, i.e. concepts independently required by the semantics of natural language. One such primitive is Notion, which reconstructs Dowty's entailment of sentience (or perception) by assuming an asymmetric relation of notion. Linking for predicates that entail notion is constrained in accordance with a consistent pattern suggesting that the individual denoted by the subject of a transitive verb is entailed to have a notion of the individual denoted by the object, while the reverse entailment does not necessarily hold. While Dowty's proto-role analysis is restricted to the domain of monotransitive verbs, Wechsler considers prepositional complements as well. He argues in favour of the view that many prepositions heading complement PPs are semantically contentful (rather than syntactically tagging a complement of the verb) and that their semantics must unify with the semantics of the predicate[8].

[4] http://framenet.icsi.berkeley.edu/

[5] For a detailed description of these relations see the FrameNet Book [17] pp. 104-111.

[6] Dowty [2] has argued that role types like agent, patient, theme, etc. are ill-founded inasmuch as it is difficult to establish sets of properties that pick out unified (undecomposable) notions.

[7] Argument selection principle: In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of Proto-Agent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of Proto-Patient entailments will be lexicalized as the direct object.

[8] This view of prepositional complements has originally been developed by Gawron [7], who points out that prepositions that occur felicitously with a particular verb are not random. According to Gawron, a necessary (though not sufficient) condition for a preposition to be selected for a given complement of a verb is that the prepositional semantics be compatible (or a component of) the semantics of the verb. Certain degree of arbitrariness

A more or less similar view of argument structure is espoused by Davis [1]. He reifies proto-roles as attributes into lexical semantic structures that capture linguistically relevant aspects of a verb's meaning in the formalism of Head-driven Phrase Structure Grammar. Each attribute is associated with one or more of a specified set of entailments holding of the denoted participant[9]. Davis' model builds upon Dowty's and Wechsler's sets of entailments. By imposing some internal structure to lexical semantic representations it provides an account for the causal relationships between events (and their participants) and the combined effect of entailments with respect to linking[10].

3.2 Annotating the Entailments

So far, proto-role properties have been used in semantic role annotation to characterize the underlying semantics of the roles used for the markup. PropBank, for example, defines the semantic content of ARG0 and ARG1 in terms of Dowty's Proto-Agent and Proto-Patient entailments, respectively; ARG0 and ARG1 are, in effect, clusters of various types of participants defined syntactically as well as semantically, similar to Davis' proto-role attributes (yet, crucially less coherent).

FrameNet, on the other hand, defines specific frames and frame elements in terms of fine-grained lexical entailments shared by individual lexical units. Proto-role entailments such as notion, causation, volition, etc. form the basis for the definition of abstract frames from which more specific ones inherit within the frame hierarchy (e.g. AWARENESS, INTENTIONALLY_ACT, etc.).

In the rest of this section, we propose to mark arguments of predicates with proto-role properties *explicitly*, i.e. in an unmediated way. We describe a tentative set of entailments (conceptualized within abstract lexical semantic relations) intended to cover a broad range of verbs with various syntactic patterns beyond transitivity and include some annotations that exemplify our scheme. Prepositional complements filling necessary slots of predicate semantics are marked with the corresponding entailments, to which they might be contributing additional information. Prepositional semantics is thus represented in terms of the common basis it shares with predicate semantics. The implications of such an approach are discussed in the next section.

i. A **notion** relation similar to that proposed by Wechsler classifies arguments (more accurately, the participants denoted by them) into *conceivers* (entailed to necessarily have some notion or perception of others) and *conceived* ones.

(6) [CONCEIVER I] noticed [CONCEIVED their appearance] and also noticed [CONCEIVED that, left alone, they disappeared too].

(7) It appears that [CONCEIVER Schwarzenegger] will renege [CONCEIVED on an agreement he made with teachers].

In accordance with Dowty's theoretical assumptions, semantic properties are generally meant to be associated with arguments in a many-to-one fashion. Thus participants that share the same entailment are distinguished in terms of any additional entailments they might have. For instance, verbs entailing that a conceiver has a notion of more than one entity might involve a complex notion relation in which conceived arguments are related to each other by means of an internal *predicative* relation (within the conceiver's mental or perceptional structures). These arguments are to be marked with corresponding additional entailments such as *entity* and *predicate*.

(8) [CONCEIVER I] soon considered [CONCEIVED, ENTITY him] [CONCEIVED, PREDICATE part of my family].

(9) [CONCEIVER The police] suspect [CONCEIVED, ENTITY Noah Rogers] [CONCEIVED, PREDICATE of aiding the robbery last night].

(10) [CONCEIVER Some] labeled [CONCEIVED, ENTITY him] [CONCEIVED, PREDICATE a womanizer].

In event types, on the other hand, in which no internal relation is entailed to hold of participants of whom a conceiver has a notion, we distinguish arguments that share this entailment in terms of their semantic salience: arguments that are secondary or less salient in terms of the essential lexical semantic relation denoted by a predicate are associated with a more specific property termed *conceived_background_state_of_affairs* (*conceived_bsoa*)[11]. For example, the primary focus in the semantics of a verb like *scour* (11) is intuitively on the scoured entity, not the sought entity (contrary to verbs like *search* or *look for*).

(11) [CONCEIVER Scientists] scoured [CONCEIVED the ice samples] [CONCEIVED_BSOA for signs of life].

(12) [CONCEIVER The coach] could distinguish [CONCEIVED the twins] [CONCEIVED_BSOA by their hair].

ii. An abstract relation of **causation** is entailed by the semantics of predicates that involve affected participants. In the denoted events, a *causer* is usually entailed to affect a *causee* in a physical or mental manner. Causally affected participants may additionally have more specific proper-

is imposed by individual lexical stipulations.

[9] Proto-role attributes are used as an appropriate level of representation for stating a small number of linking principles.

[10] Davis models the fact that the causal structure of the semantics of a predicate takes precedence over all other entailments for purposes of linking. Note that the lexical semantic representations he proposes implicitly rely on previous work by Gruber [9], Jackendoff [10, 11], Talmy [18] and Pinker [16]. All of these works advocate a relational view of thematic roles defining them in terms of sets of positions in semantic structures, themselves based upon sets of grammatically relevant elements of lexical semantics, i.e. recurring meaning components that have some effect on a verb's grammatical behavior. A fairly similar intuition underlies the semantic predicates (e.g. *motion*, *contact*, or *cause*) associated with the semantics of verb classes in VerbNet.

[11] Such arguments are often syntactically optional. Nevertheless, syntactic properties do not independently yield sufficient diagnostics for identifying the semantic status of arguments.

ties such as *change-of-state* or *incremental themehood*[12] referring to readily observable changes in their (physical or mental) states; furthermore, predicates that entail a *change-of-state* may also lexicalize a *source* and/or *end* state of the affected entity.

(13) [CAUSEE His home and car] had been attacked in the past.
(14) [CAUSER My dad] changed [CAUSEE, CH_OF_STATE his hair color] [SOURCE_STATE from red] [END_STATE to blue] today.
(15) [CAUSER Samantha] terrorized [CAUSEE, CH_OF_STATE the children] [END_STATE into screaming].
(16) [CAUSEE This service] will diminish [CAUSEE, CH_OF_STATE in quality].
(17) [CAUSER He] coated [CAUSEE, INCR_THEME the wall] [CAUSEE with paint].

A causal event may alternatively involve a participant that either brings about or affects a *resulting event or state*. Preserving much of what is valuable in Davis' model, we keep track of the internal status of entailments within the caused or affected event by representing them in square brackets. Roughly, a *conceiver* differs from a *[conceiver]* in that the latter is an entity that is *caused to conceive*. For instance, a communication verb like *report* is represented in accordance with the entailment that a speaker necessarily conceives of an addressee and certain message and causes that the addressee conceives of that message too. Similarly, verbs such as *aid* and *constrain* involve causally affected states of affairs described in terms of their internal entailments (cf. verbs of caused motion and possession discussed below). The roles associated with *aid*, for instance, are represented by means of the underlying entailment that some entity (*discussion of the film*) affects the internal relation between an intentional conceiver (*the teacher*) and a conceived participant (the *evaluation of the effectiveness of the film*).

(18) [CAUSER, CONCEIVER He] reported [[CONCEIVED], CONCEIVED the matter] [[CONCEIVER], CONCEIVED to the security].
(19) [CAUSER The painting] inspired [CAUSEE, [CONCEIVER, INTENTIONAL] me] [[CONCEIVED] to take the risk and use the intense green for the sky].
(20) [CAUSER Discussion of the film] can aid [[CONCEIVER, INTENTIONAL] the teacher] [[CONCEIVED] in evaluating the effectiveness of the film].
(21) [CAUSER The presence of exams] seems to constrain [CAUSEE, [CONCEIVER, INTENTIONAL] them] [[CONCEIVED] in their approach to classroom teaching].

iii. Dowty's entailment of volitional involvement in an event or state is replaced with the more clearcut property of **intentionality**. *Intentional* participants are characterized by conscious choice, decision or control over the course of action denoted by the verb; specific *intentions* might also be lexicalized.

(22) [INTENTIONAL, CAUSER The company] manufactured [CAUSEE, CH_OF_STATE T-shirts].
(23) [INTENTIONAL, CONCEIVER He] used [CONCEIVED his influence] [CONCEIVED_BSOA, INTENTION to favour a contemporary of Keeton's].
(24) [INTENTIONAL Science] aims [INTENTION at theories with a large informative content].

iv. A relation of **motion** generally involves a *moving entity* and a stationary reference frame (*path*) within which various points (start, end, or intermediate) may be further specified. Verbs of caused motion, in particular, involve participants that are both moving and are causally affected (i.e. set to motion).

(25) [MOVING A woman in uniform] entered [PATH the room].
(26) [MOVING The uranium particles] radiate [PATH_SOURCE from the nuclear plant].
(27) [MOVING We] approached [PATH_GOAL the house].
(28) [CAUSER You] can use it to shoot [CAUSEE, [MOVING] heavy balls of metal] [[PATH_SOURCE] from large guns].
(29) [CAUSER John] ran [CAUSEE, [MOVING] the car] [[PATH_GOAL] into the field].
(30) [MOVING The rock] hit [PATH_GOAL the sand] with a thump.
(31) [MOVING, CONCEIVER The squirrel] chased [CONCEIVED the nut] [PATH across the road].
(32) [CONCEIVER, INTENTIONAL, MOVING Several Indian peasant leaders] fled [CONCEIVED, PATH_SOURCE the country] in the early hours of the coup.

v. **Inclusion** relies on one of Wechsler's primitive relations and captures the entailment of some entity necessarily being a constituent *part* or member of a *whole* (i.e. a physical, social, or mental entirety).

(33) [WHOLE The box] holds [PART three hundred pictures].
(34) [WHOLE The collaboration] incorporates [PART movement, dance, music and vocal techniques] to explore Chekhov's text.
(35) [PART Several of the countries] were unable to participate [WHOLE in the market].
(36) [CAUSER He] has merged [CAUSEE, [PART] the two companies] [[WHOLE] into a single organization].

vi. Finally, a relation of **possession** accounts for the semantics of transitive verbs like *have*, *own*, *acquire*, *inherit*, *lack* and ditransitives like *give*. The latter are represented as meaning *cause-to-possess*, i.e. in terms of causation and possession. In addition to a *possessor* and a *possessed* entity, a *source* of possession (or the initial possessor)

[12] In the terminology of Dowty [3], *incremental themes* identify roles for which a change of state in the participants filling them reflects the temporal structure (i.e. the progression) of the denoted event in question.

might also be lexicalized[13].

(37) [POSSESSOR Iran] had acquired [POSSESSED four nuclear weapons] [SOURCE from former Soviet Moslem republics].

(38) [CAUSER Hunting] provides [[POSSESSOR] the men] [[POSSESSED] with a public stage] for the stylized display of virility.

(39) [CAUSER They] submitted [[POSSESSED] their evidence] [[POSSESSOR] to the committee].

Possession is also entailed by predicates of commercial transaction that include two transfer events (i.e. the transfer of goods and the transfer of money), either of which might be highlighted as the main event.

(40) [SOURCE, CAUSER Ben] sold [[POSSESSED] the car] [[POSSESSOR] to Lisa].

(41) [SOURCE, CAUSER, CONCEIVER Lisa] paid [[POSSESSOR] Ben] [[POSSESSED] 15,000 dollars] [CONCEIVED_BSOA for the car].

The list of proto-role properties described above is by no means complete. Additional abstract lexical semantic relations and corresponding entailments might be necessary to represent the semantics of classes of predicates that have not been discussed here[14].

Yet, certain predicates raise interesting questions for an entailment-based view of semantic roles. For instance, so-called symmetric verbs have arguments that are indistinguishable in terms of entailments displaying significant variability of syntactic patterns (42)-(43). Such arguments are traditionally described by roles (e.g. figure, ground) that Dowty refers to as perspective-dependent (contrary to event-dependent roles). The semantic properties of participants that seem to vary across different perspectives of viewing an event are most probably outside the scope of a proto-role theoretical approach that inherently involves asymmetric relations between entities.

(42) [FIGURE The house] is near [GROUND the sea].

(43) [FIGURE The sea] is near [GROUND the house].

4. Evaluation and Future Directions

As a preliminary evaluation of our approach we addressed a case study comparing it with one of the existing annotation approaches. We concentrated on a portion of English lexical items (verbs for the moment) ensuring a linguistically representative dataset for each of them. We used the FrameNet full-text corpora as source of our data. Additionally, for each verb found in the corpora and the semantically related ones belonging to the corresponding, invoked frames we extracted collections of example annotated sentences from the FrameNet lexicon.

A new annotation layer was added to this data mapping proto-role properties to FrameNet roles (frame elements) identified in the sentences in accordance with the previously described scheme. We automatically produced annotations by mapping frame elements to entailments at a frame level and then manually checked these annotations for consistency in terms of the semantics of individual verbs. In general, members of the same frame share a minimum of common properties. Yet, they might differ in specific aspects of their meaning. For instance, *obtain* differs from a passive sense of the verb *acquire* in that it entails action on the part of the eventual possessor.

(44) "[POSSESSOR It] needs to *acquire* [POSSESSED some teeth] [SOURCE from somewhere]'', he said.

(45) In some cases, [CAUSER, [POSSESSOR] the BGS libraries] *obtained* [[POSSESSED] copies of these] [[SOURCE] from the authors].

We considered a total of 241 frames. Each of the core frame elements within a given frame was mapped to a unique entailment or a combination of entailments[15]. Non-core frame elements (such as Time, Place, Purpose, Reason, Manner, etc.) whose semantics are independent of individual frames and predicates were dismissed from consideration. Specific frame elements posited interesting issues for refinement of the semantic content of entailments:

(46) The policy will be implemented [INTENTIONAL by a new computer system][16].

A first issue is to do with the coverage of the set of properties we assumed. We identified 19 frames for which none of the entailments of the previous section seemed to hold. The majority of these includes stative verbs such as *exist*, *happen*, *occur*, *remain*, *continue*, *lie*, *stand*, *depend*, *rank*, *resemble*, *match*, etc., some of which involve perspective-dependent roles. The rest of problematic cases include verbs like *can*, *should*, *must*, *follow*, *precede*, *rival*, *equal*, *respond*, *demand*, *deserve*, etc. Further analysis of these is necessary to shed light to the scope of an entailment-based approach. In general, predicates whose linking patterns depend on pragmatic or contextual information are expected to require a different treatment.

From the new annotated dataset we additionally extracted mappings of entailments to syntactic categories, a portion of which is summarized in Table 1 along with the

[13] The entailments *source*, *path_source* and *source_state* can be replaced by a more general property termed *source* entailed for participants that undergo some kind of change (i.e. of possession, location, or state). Jackendoff [10, 11] building on Gruber [9] has put forth a relevant analysis that treats linguistic uniformities across various semantic fields in terms of extensions of motional and locational (conceptual) structures (an analysis known as the *thematic relations hypothesis*).

[14] Davis advocates a relation of *surpassing* for the semantics of verbs such as *exceed*, *dwarf*, *outscore*, *outplay*, etc. that entail that a superior entity outranks an inferior one. On the other hand, he describes verbs like *hit*, *strike*, *poke*, *tap*, *press*, etc. in terms of an entailment of impingement or forceful contact. We refrain from adopting the latter as an independently motivated entailment and represent verbs of forceful contact in terms of motion.

[15] An exception involves metonymically related frame elements, i.e. closely related, mutually exclusive roles distinguished solely by ontological criteria. Such frame elements are mapped to the same entailments.

[16] In accordance with the definition of intentionality, a computer might be intentional in a sense that it controls the course of an inherently intentional action (i.e. its start, intermission, or end points).

FrameNet frames from which it was acquired[17]. As can be seen in this table, the distribution of realizations of entailments (or combinations of entailments) in the dataset readily abstracts over a wide range of semantic and syntactic combinatorial properties of individual verbs belonging to semantically related or even unrelated frames. Contrary to fine-grained lexical semantic distinctions that underlie the FrameNet frames, an asset of proto-role entailments is that they identify grammatically pervasive semantic elements suitable for defining an abstract, principled syntax-semantics interface. In a frame-wise approach, generalizations about mappings to syntactic form should emerge by analyzing the distribution of role assignments for each frame separately, abstracting over specific mappings of the corresponding lexical units and attempting to unify abstractions across the frame hierarchy, which is not a trivial task. Proto-role properties practically decouple linking information from aspects of lexical semantics that have no effect on it. They thus capture systematic linking patterns (including entailment-preposition correspondences) that can be formalized into classes of non-lexicalized frames.

In sum, proto-role entailments are semantic notions firmly grounded in linguistic intuition that have a wide coverage over lexical semantic relations that humans express *systematically*. Contrary to catch-all labels, they pin down predicate-argument structure relations in a general, yet coherent way. The list of entailments in 3.2 is presented as a first attack to the methodological issues related to annotation of such properties. Further analysis should refine and extend the current set to cover an even wider range of predicate types.

In a subsequent phase of evaluating our proposal, we intend to employ an entailment-based annotation layer for semantic parsing. Using the above annotated data we plan to implement a semantic analyzer that identifies the entailments instead of the frame elements associated with arguments. Since frame elements are uniquely mapped to entailments within each frame, we can eventually evaluate the entailment-based system comparing its performance to standard frame-semantic models.

5. Acknowledgements

Much of this work has been supported by the Greek State Scholarships Foundation. I am indebted to Angelos Nikolaou for his valuable support in programming. I am also grateful to Doug Arnold and Massimo Poesio for their insightful comments on parts of this work.

6. References

[1] Davis, Anthony. 2001. Linking by types in the hierarchical lexicon. CSLI Publications.

[2] Dowty, David. 1986. On the semantic content of the notion thematic role. In Gennaro Chierchia, Barbara Partee and Ray Turner, eds. Property theory, Type theory, and Natural language semantics. Dortrecht: Reidel.

[3] Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. Language 67.3. 547-619.

[4] Fillmore, Charles J. 1985. Frames and the semantics of understanding. Quaderni di Semantica 6.2. 222-254.

[5] Fillmore, Charles J., Christopher R. Johnson, Miriam R. Petruck. 2003. Background to FrameNet. International journal of lexicography, 16. 235-250.

[6] Frank, Annette. 2004. Generalizations over corpus-induced frame assignment rules. In Charles Fillmore, Manfred Pinkal, Collin Baker and Katrin Erk (eds.): Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora. Lisbon, Portugal. 31-38.

[7] Gawron, Jean Mark. 1986. Situations and prepositions. Linguistics and philosophy 9. 327-382.

[8] Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational Linguistics 28 (3) 245-288.

[9] Gruber, Jefferey. 1965. Studies in Lexical Relations. Ph.D. dissertation, MIT (reprinted in Lexical Structures in Syntax and Semantics. Amsterdam, North-Holland, 1976).

[10] Jackendoff, Ray. 1983. Semantics and Cognition. Cambridge, MA, MIT Press.

[11] Jackendoff, Ray. 1990. Semantic Structures. Cambridge, MA, MIT Press.

[12] Kingsbury, Paul and Martha Palmer. 2002. From Treebank to PropBank. In Proceedings of the LREC, Las Palmas, Canary Islands, Spain.

[13] Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000), Austin,TX, July-August.

[14] Levin, Beth. 1993. English verb classes and alternations: A preliminary investigation. University of Chicago Press.

[15] Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In Proceedings ARPAHLT Workshop.

[16] Pinker, Steven. 1989. Learnability and Cognition. Cambridge, MA, MIT Press.

[17] Ruppenhofer, Josef, Michael Ellsworth, Miriam R. Petruck, Christopher R. Johnson, Jan Scheffczyk. FrameNet II: Extended theory and practice: http://www.icsi.berkeley.edu/framenet/book/book.html

[18] Talmy, Leonard. 1985. Lexicalization patterns: semantic structure in lexical form. In Language Typology and Syntactic Description, vol.3, ed. Timothy Shopen. Cambridge, UK, Cambridge University Press.

[19] Wechsler, Stephen. 1991. Argument Structure and Linking. Ph.D. Dissertation, Stanford University.

[20] Wechsler, Stephen. 1995. The semantic basis of argument structure. Stanford, CA. CSLI Publications.

[17] Phrase structure and grammatical function tags are the ones used by FrameNet; the two annotation layers are represented as a compound tag with a dot separating them. The tag Ext is used for external arguments which include subjects of finite verbs, Obj refers to objects, while Dep is assigned to dependents of the governing verb. Parentheses denote optional properties.

Lexical semantic relation	**Linking generalizations**	**FrameNet frames**
Notion	CONCEIVER: NP.Ext CONCEIVED: NP.Obj PP[on].Dep PP[upon].Dep PP[of].Dep PP[over].Dep PP[about].Dep PP[for].Dep PP[after].Dep PP[in].Dep	Locating, Desiring, Becoming_aware, Activity_ongoing, Coming_to_believe, Certainty, Experiencer_subj, Feeling, Awareness, Expectation, Perception_experience, Waiting, Opinion, Be_in_agreement_on_assessment, Trust, Cogitation, Emotion_active, Religious_belief, etc.
	CONCEIVER: NP.Ext CONCEIVED: NP.Obj CONCEIVED_BSOA: PP[by].Dep PP[for].Dep	Categorization, Differentiation, Judgment, etc.
	CONCEIVER: NP.Ext CONCEIVED, ENTITY: NP.Obj CONCEIVED, PREDICATE: PP[of].Dep PP[as].Dep NP.Dep AJP.Dep	Categorization, Judgment, Judgment_communication, Suspicion, Labeling, Referring_by_name, etc.
Notion, Intentionality	CONCEIVER, INTENTIONAL: NP.Ext CONCEIVED: NP.Obj PP[for].Dep PP[on].Dep PP[upon].Dep	Taking_sides, Place_weight_on, Using, Ratification, Sign_agreement, Execute_plan, Deciding, Going_back_on_commitment, Leadership, Intentionally_act, Attempt, Employing, Operating_a_system, Accomplishment, Change_of_leadership, Piracy, Examination, Seeking, Claim_ownership, Activity_prepare, Collaboration, Discussion, Make_agreement_on_action, Hiring, Avoiding, etc.
	CONCEIVER, INTENTIONAL: NP.Ext CONCEIVED: NP.Obj CONCEIVED_BSOA: PP[for].Dep PP[on].Dep PP[over].Dep	Scrutiny, Arrest, Assessing, Inspecting, Revenge, Operational_testing, Scouring, Hostile_encounter, Justifying, etc.
	CONCEIVER, INTENTIONAL:NP.Ext CONCEIVED: NP.Obj CONCEIVED_BSOA, INTENTION: PP[for].Dep VPto.Dep	Using, Practice, Employing, Hiring, Needing, etc.
Causation	CAUSER: NP.Ext CAUSEE: NP.Obj (CAUSEE, CH_OF_STATE): PP[in].Dep	Objective_influence, Causation, Attack, Experiencer_obj, Change_event_time, Hindering, Preventing, Thwarting, Cause_to_continue, Cause_harm, Cause_to_experience, Eclipse, Cause_change_of_position_on_a_scale, etc.
	CAUSER: NP.Ext CAUSEE, CH_OF_STATE: NP.Obj (SOURCE_STATE): PP[from].Dep (END_STATE): PP[to].Dep PP[into].Dep	Creating, Cause_to_end, Cause_expansion, Render_nonfunctional, Destroying, Cause_change_of_position_on_a_scale, Resolve_problem, Cause_to_resume, Killing, Cause_to_start, Damaging, Grinding, Cause_to_fragment, Cause_change, Reshaping, etc.
Causation, Notion, (Intentionality)	CAUSER, CONCEIVER, (INTENTIONAL): NP.Ext CAUSEE, (CH_OF_STATE), CONCEIVED: NP.Obj	Cause_to_make_progress, Intercepting, Processing_materials, Activity_resume, Committing_crime, Activity_stop, Activity_pause, Activity_finish, etc.
	CAUSER, (INTENTIONAL): NP.Ext (CAUSEE), [CONCEIVER, (INTENTIONAL)]: NP.Obj [CONCEIVED]: PP[in].Dep PP[into].Dep	Assistance, Subjective_influence, Manipulate_into_doing, Suasion, Hindering, etc.

Table 1: Entailment-based linking generalizations and corresponding FrameNet frames

A French Corpus Annotated for Multiword Expressions with Adverbial Function

Eric Laporte, Takuya Nakamura, Stavroula Voyatzi
Université Paris-Est
Institut Gaspard-Monge - LabInfo
5, Boulevard Descartes, Champs-sur-Marne
77454 Marne-la-Vallée Cedex 2 (France)
E-mail: eric.laporte@univ-paris-est.fr, nakamura@univ-mlv.fr, voyatzi@univ-mlv.fr

Abstract

This paper presents a French corpus annotated for multiword expressions (MWEs) with adverbial function. This corpus is designed for investigation on information retrieval and extraction, as well as on deep and shallow syntactic parsing. We delimit which kind of MWEs we annotated, we describe the resources and methods we used for the annotation, and we briefly comment the results. The annotated corpus is available at *http://infolingu.univ-mlv.fr/* under the LGPLLR license.

1. Introduction

Recognising multiword adverbs such as *à long terme* 'in the long run' in texts is likely to be useful for information retrieval and extraction because of the information that such adverbials can convey. In addition, it is likely to help resolving prepositional attachment during shallow or deep parsing: most multiword adverbs have the superficial syntax of prepositional phrases; in many cases, recognising them rules out analyses where they are arguments or noun modifiers.

The quality of the recognition of multiword adverbs depends on algorithms, but also on resources. We created a corpus of French texts annotated with multiword adverbs. In this article, we survey related work, we define the target of our annotation effort, we describe the method we have implemented and we analyse the corpus obtained. This corpus will be made freely available on the web under the LGPLLR license when this article is published.

2. Related work

Corpora annotated with multiword adverbs are rare and small[1]. In the Grace corpus (Rajman *et al*., 1997), most multiword units are ignored. In the French Treebank (Abeillé *et al*., 2003), prepositional phrases and adverbs are annotated with a binary feature ('compound') which indicates whether they are multiword units; the distinction between whether prepositional phrases are verb modifiers, noun modifiers or objects appears only in the function-annotated part of the Treebank (350 000 words). We are not aware of other available French corpora annotated with multiword adverbs. In other languages, including English, corpora annotated with multiword units are rare and small as well.

3. Target of annotation

The target of our annotation effort is defined by the intersection of two criteria: (i) multiword expressions and (ii) adverbial function. In this section, we define both criteria in more detail, we define the features that we included in the annotations, and we describe the corpus.

3.1 Multiword expression criterion

For this work, we considered a phrase composed of several words to be a multiword expression if some or all of their elements are frozen together in the sense of (Gross, 1986), that is, if their combination does not obey productive rules of syntactic and semantic compositionality. In the following example, *de nos jours* ('nowadays', lit. 'of our days') is a multiword adverb:

(1) *Il est facile* ***de nos jours*** *de s'informer*
'It is easy to get informed **nowadays**'

This criterion ensures a complementarity between lexicon and grammar. In other words, it tends to ensure[2] that any combination of linguistic elements which is licit in the language, but is not represented in syntactic-semantic grammars, will be stored in lexicons.

Syntactic-semantic compositionality is usually defined as follows (Freckleton, 1985; Machonis, 1985; Silberztein, 1993; Lamiroy, 2003): a combination of linguistic elements is compositional if and only if its meaning can be computed from its elements. This is also our conception. However, in this definition, we consider that the possibility of computing the meaning of phrases from their elements is of any interest only if it is a better solution than storing the same phrases in lexicons, i.e. if

[1] Several reasons explain this lack of interest. Firstly, adverbials are usually felt as less useful than nouns for information retrieval and extraction. Secondly, many multiword adverbs are difficult to distinguish from prepositional phrases assuming other syntactic functions, such as arguments or noun modifiers: the distinction is hardly correlated to any material markers in texts and lies in complex linguistic notions (Villavicencio, 2002; Merlo, 2003). The task is therefore felt as too difficult by most researchers in language processing, whose main background is in information technology. However, the distinction in question is essential to identifying the semantic core of a sentence, and the availability of a larger corpus of annotated text is likely to shed light on the problems posed by this task.

[2] That can be empirically checked only after a lexicon and a grammar for the same language are complete and compatible.

they rely on grammatical rules with sufficient generality. In other words, we consider a combination of linguistic elements to be compositional if and only if its meaning can be computed from its elements **by a grammar**. In example (1) above, the lack of compositionality is apparent from distributional restrictions[3] such as:

* *Il est facile de nos **semaines** de s'informer*
*'It is easy to get informed nowa**weeks**'

Multiword expressions include many different subtypes, varying from entirely fixed expressions to syntactically more flexible expressions (Sag *et al*., 2002). We annotated expressions undergoing variations[4]. In (2), the possessive adjective agrees obligatorily in person and number with the subject of the sentence:

(2) ***De (ses + *mes) propres mains**, il a construit une maison*
'**With (his + *my) own hands**, he built a house'

3.2 Adverbial function

We annotated only expressions with adverbial function, or circumstantial complements, i.e. complements which are not objects of the predicate of the clause in which they appear. We recognised them through criteria (Gross 1986, 1990a, 1990b) involving the fact that they are optional, they combine freely with a wide variety of predicates and some of them pronominalize with specific forms. Phrases with adverbial function are often called 'circumstantial complements', 'adverbials', 'adjuncts', or 'generalised adverbs'. They assume several morphosyntactic forms: underived (*demain* 'tomorrow') or derived adverbs (*prochainement* 'soon'), prepositional phrases (*à la dernière minute* 'at the last minute') or circumstantial clauses (*jusqu'à ce que mort s'ensuive* 'until death comes'), and special structures in the case of named entities of time (*lundi 20* 'on Monday 20'). We annotated NEs only when they have an adverbial function, as in: *Jean arrive lundi 20* 'John arrives on Monday 20'. NEs of other categories, such as places, persons, events, etc., are usually not adverbials.

3.3 Features

Two types of features were included in the annotations.
(i) Each occurrence of a multiword adverb was assigned one internal morphosyntactic structure or semantic type among 19. The definition of the morphosyntactic structures is based on the number, category and position of the frozen and free components of the adverbial. They are described as a sequence of parts of speech and syntactic categories. For example, *à la nuit tombante* 'at nightfall' is assigned a structure identified by the mnemonic acronym *PCA*, and defined as *Prép Dét C (MPA) Adj*, where *C* stands for a noun frozen with the rest of the adverbial, *Adj* for a post-posed noun modifier (e.g. an adjectival phrase or a relative clause), and *MPA* for a pre-adjectival modifier, empty in this lexical item. For named entities, this feature encodes the semantic type: date, duration, time or frequency, in conformity with the typology of the Infom@gic project (Martineau *et al*., 2007). The 19 structures and semantic types are listed in Table 1. In this table, *N* stands for a free noun phrase, and *W* for a variable ranging over verb complements. Other symbols are easy to interpret: *Prép*, *Dét*, *Adj*, *V*, *Conj*...

Identifiers	Structures	Examples
PC	Prép C	*en bref*
PDETC	Prép Dét C	*de nos jours*
PAC	Prép Adj C	*à la dernière minute*
PCA	Prép C Adj	*à la nuit tombante*
PCDC	Prép C de C	*dans la limite du possible*
PCPC	Prép C Prép C	*des pieds à la tête*
PCONJ	Prép C Conj C	*en tout et pour tout*
PCDN	Prép C de N	*au moyen de N*
PCPN	Prép C Prép N	*par rapport à N*
PV	Prép V W	*à dire vrai*
PF	P (frozen clause)	*jusqu'à ce que mort s'ensuive*
PECO	(Adj) comme C	*comme ses pieds*
PVCO	(V) comme C	*comme un cheveu sur la soupe*
PPCO	(V) comme Prép C	*comme dans du beurre*
PJC	Conj C	*mais enfin et surtout*
DATE	Named Entities	*le 22 mai 2008*
DURATION	Named Entities	*pendant vingt-quatre heures*
TIME	Named Entities	*à huit heures du soir*
FREQUENCE	Named Entities	*deux fois par jour*

Table 1: Morphosyntactic structures and semantic types of MWEs with adverbial function

(ii) The second feature is binary and encodes whether the adverbial assumes a conjunctive function in discourse, i.e. it connects the clause in which the adverbial occurs with the previous clause, as *en dernier lieu* 'finally'. The positive value is indicated by identifier 'Conj' in attribute 'fs'. Example: *<ADV fs='PAC Conj'>*.

3.4 The corpus

The corpus we annotated includes: (a) the complete minutes of the sessions of the French National Assembly on October 3-4, 2006, transcribed into written style from oral French (hereafter *AS*)[5] and (b) Jules Verne's novel *Le Tour du monde en quatre-vingts jours*, 1873 (hereafter *JV*). Errors (e.g. *mis enoeuvre* for *mis en oeuvre* 'implemented') have not been corrected. Statistics on the corpus are displayed in Table 2.

[3] The point is that this blocking of distributional variation (and other syntactic constraints) cannot be predicted on the basis of general grammar rules and independently needed lexical entries. Therefore, the acceptable combinations are meaning units and have to be included in lexicons as multiword lexical items.

[4] We annotated phrases which comprise a frozen part and a free part, e.g. *au moyen de ce bouton* 'with the aid of this switch', in which *au moyen de* 'with the aid of' is frozen, and *ce bouton* 'this switch' is a distributionally free noun phrase embedded in the global phrase. In such cases, we delimited the embedded free part with tags (cf. section 4.2). Finally, we annotated named entities (NEs) of date and duration. The status of named entities with respect to compositionality is not fully consensual: however, we complied with the usual view that, since they follow quite specific grammatical rules, they should be considered as multiword expressions.

[5] http://www.assemblee-nationale.fr/12/documents/index-rapports.asp.

	size (Kb)	sentences	tokens	types
corpus *AS*	824	5 146	98 969	18 028
corpus *JV*	1 231	3 648	69 877	19 828
total	2 055	8 794	168 846	37 856

Table 2: Size of the corpus

4. Methodology

In order to annotate the corpus, we tagged the occurrences of the expressions described in a syntactic-semantic lexicon of adverbials, as Abeillé *et al*. (2003), Baptista (2003) for Portuguese, and Català & Baptista (2007) for Spanish; we tagged NEs of date, duration, time, and frequency through a set of local grammars, as Friburger & Maurel (2004); then, we revised the annotation manually.

4.1 The lexicon

We used the same syntactic-semantic lexicon (Gross, 1990a) as Abeillé *et al*. (2003), so that the two corpora can be used jointly for further research. This lexicon has 6 800 entries. It is freely available[6] for research and business under the LGPLLR license. It was constructed on the basis of conventional dictionaries, grammars, corpora and introspection, within the Lexicon-Grammar methodology (Gross, 1986; 1994). It takes the form of a set of Lexicon-Grammar tables such that of Table 3, which displays a sample of the lexical items with the *PCA* morphosyntactic structure.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	N0 =: Nhum	N0 =: N-hum	Nég obl			Prép	Dét	C	Modif pré-adj	Adj	Prép Dét C	Prép Dét MPA Adj C	Prép MPA Adj C	Conjonction
170	+	-	-	<E>	agir	dans	les	délais	les plus	brefs	-	-	+	-
171	+	-	-	<E>	agir	dans	les	délais	les plus	courts	-	-	+	-
172	+	-	-	<E>	agir	dans	les	délais	les	meilleurs	-	-	+	-
173	+	-	-	<E>	rire	<E>	toutes	dents	<E>	dehors	-	-	-	-
174	-	+	-	:se	produire	à	cette	époque-	<E>	ci	+	-	-	-
175	-	+	-	:se	produire	à	cette	époque-	<E>	là	+	-	-	+

Table 3: Sample of the table of entries with the *PCA* morphosyntactic structure

In this table, each row describes a lexical item, and each column corresponds:
- either to one of the elements in the morphosyntactic structure of the items (columns with identifiers 'Prép', 'Dét', 'C', 'Modif pré-adj' and 'Adj');
- or to a syntactic-semantic feature (columns with binary values), for example the conjunctive function of the adverbial in discourse (column with identifier 'Conjonction'), or the constraint that the adverbial obligatorily occurs in a negative clause (column with identifier 'Nég obl');
- or to illustrative information provided as an aid for the human reader to find examples of sentences containing the adverbial (e.g. columns D and E giving an example of a verb compatible with the adverb).

There are 15 such tables, one for each of the morphosyntactic structures. The features provided by the lexicon were used to annotate the occurrences.

4.2 Tagging

We tagged the corpus with the Unitex system (Paumier, 2006). Many multiword adverbs are entirely fixed expressions, but others present variations, such as grammatical agreement (cf. example (2), section 3.1), permutations and omissions. Due to these variations, we tagged them with finite-state transducers (FST): the input part of these transducers recognises the expressions and their variants, and the output part inserts the tags. Like Català & Baptista (2007), we used lexicalised transducers, i.e. one for each lexical item, and we generated them with the technique of parameterised graphs (Roche, 1999) modified by Silberztein (1999).

Multiword adverbs with a free prepositional phrase modifier (morphosyntactic structures *PCDN* and *PCPN*) were annotated semi-automatically as follows ('*N*' if the free complement is occupied by a noun phrase, '*S*' if it is occupied by a clause):

(i) *<ADV fs='PCDN'>**compte tenu de** <NP>**vos ambitions**</NP></ADV>*
'taking into account your ambitions'

(ii) *<ADV fs='PCDN'>**compte tenu de** <S>**ce que tout va bien**</S></ADV>*
'taking into account that everything is OK'

Named entities with temporal value (cf. section 3.2) were automatically tagged by using FST methods similar to those applied for multiword adverbs.

4.4 Manual revision

The annotation was manually reviewed by three experts. This validation followed guidelines, which are available along with the corpus. It involved two operations.

(i) The sequences tagged with the aid of the lexicon and Unitex were checked in order to detect cases when the recognised sequence is in fact a part of a larger MWE. For instance, when *de force* 'forcibly' occurred within the compound noun *ligne de force* 'thrust', the tags around *de force* were deleted.

When the embedded free part of a multiword adverb is a coordination, we tagged it manually:

<ADV fs='PCDN'>en termes de <NP>santé</NP>
et d'<NP>éducation</NP></ADV>
'in terms of health and education'

(ii) The text was integrally reviewed in search for multiword adverbs absent from the lexicon, and thus undetected by Unitex, e.g. *de plus* 'moreover' or *pour le moins* 'at least'.

This required for the annotators to identify the syntactic structure of each sentence in the corpus. We had meetings during the annotation process in order to make it consistent.

[6] http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Lexiques-Grammaires/View.html.

5. Results

This corpus is annotated with 4 247 occurrences of MWEs with adverbial function. They represent about 6 % of the overall of simple word occurrences occurring in the whole corpus. Table 4, below, shows the number of occurrences of annotated MWEs. The lines of the table correspond to the morphosyntactic structures and semantic types.

identifiers	*JV* corpus	% *JV* cover	*AS* corpus	% *AS* cover
PC	338	1.38	420	1.28
PDETC	257	1.16	165	0.64
PAC	77	0.35	127	0.51
PCA	55	0.30	53	0.22
PCDC	38	0.17	36	0.12
PCPC	37	0.15	59	0.20
PCONJ	13	0.07	21	0.08
PCDN	248	1.00	834	2.52
PCPN	103	0.41	107	0.32
PV	53	0.21	54	0.17
PF	11	0.04	23	0.07
PECO	1	0.00	1	0.00
PVCO	8	0.04	3	0.00
PPCO	2	0.00	1	0.00
PJC	2	0.00	3	0.00
DATE	258	1.00	383	1.04
DURATION	120	0.49	111	0.31
TIME	128	0.50	29	0.06
FREQUENCE	31	0.11	37	0.10
Total	1 780	6.28	2 467	6.64

Table 4: Annotated occurrences of MWEs with adverbial function in the corpus

6. Conclusion

This paper described the design of a French corpus annotated for MWEs with adverbial function. Various types of features are included in the annotations: the morphosyntactic structure, special functions in discourse (e.g. the conjunctive function) and the semantic types of named entities of time. This annotated corpus can be used jointly with the French Treebank (Abeillé *et al*., 2003) for research on information retrieval and extraction, automatic lexical acquisition, as well as on deep and shallow syntactic parsing.

7. Acknowledgment

This task has been partially financed by CNRS and by the Cap Digital business cluster. We thank Anne Abeillé for making the French Treebank available to us.

8. References

Abeillé, A., Clément, L., and Toussenel F. (2003). Building a Treebank for French. In A. Abeillé (Ed.), *Building and Using Parsed Corpora*, *Text*, *Speech and Language Technology*, 20, Kluwer, Dordrecht, pp. 165--187.

Baptista, J. (2003). Some Families of Compound Temporal Adverbs in Portuguese. In *Proceedings of the Workshop on Finite-State Methods for Natural Language Processing*, *EACL 2003*, Budapest, Hungary, pp. 97--104.

Català, D., Baptista, J. (2007). Spanish Adverbial Frozen Expressions. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, *ACL 2007*, Prague, Czech Republic, pp. 33--40.

Freckleton, P. (1985). Sentence idioms in English, *Working Papers in Linguistics*, University of Melbourne, pp. 153--168 & appendix (196 p.).

Friburger, N., Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1), pp. 93--104.

Gross, M. (1986). Lexicon-Grammar. The representation of compound words. In *Proceedings of the Eleventh International Conference on Computational Linguistics*, Bonn, West Germany, pp. 1--6.

Gross, M. (1990a). *Grammaire transformationnelle du français: 3. Syntaxe de l'adverbe*. Paris, ASSTRIL.

Gross, M. (1990b). La caractérisation des adverbes dans un lexique-grammaire. *Langue Française*, 86, pp. 90--102.

Gross, M. (1994). Constructing Lexicon-Grammars. In Atkins & Zampoli (Eds.), *Computational Approaches to the Lexicon*, Oxford University Press, pp. 213--263.

Lamiroy, B. (2003). Les notions linguistiques de figement et de contrainte, *Lingvisticae Investigationes*, 26:1, Amsterdam/Philadelphia: John Benjamins, pp. 1--14.

Machonis, P. (1985). Transformations of verb phrase idioms: passivization, particle movement, dative shift. *American Speech*, 60:4, pp. 291--308.

Martineau, C., Tolone, E., Voyatzi, S. (2007). Les Entités Nommées : usage et degrés de précision et de désambiguïsation. In *Proceedings of the Twenty Sixth International Conference on Lexis and Grammar*, Bonifacio, Corse du Sud, pp. 105--112.

Merlo, P. (2003). Generalised PP-attachment Disambiguation using Corpus-based Linguistic Diagnostics. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 251--258.

Paumier, S. (2006). *Unitex Manual*. Université Paris-Est. http://igm.univ-mlv.fr/~unitex/manuel.html.

Rajman, M., Lecomte, J., Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Rapport GRACE GTR-3--2.1.

Roche, E. (1999). Finite-state transducers: parsing free and frozen sentences. In Kornai (Ed.), *Extended finite-state models of language,* Cambridge University Press, pp. 108--120.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbuk (Ed.), *Computational Linguistics and Intelligent Text Processing: Proceedings of the Third International Conference CICLing 2002*, Springer-Verlag, Heidelberg/Berlin, pp. 1--15.

Silberztein, M. D. (1993) Les groupes nominaux productifs et les noms composés lexicalisés. *Lingvisticae Investigationes*, 17:2, Amsterdam/ Philadelphia, John Benjamins, pp.405--426.

Silberztein, M. (1999). *Manuel d'utilisation d'Intex version 4.12*.

Villavicencio, A. (2002). Learning to distinguish PP arguments from adjuncts. In *Proceedings of the Sixth Conference on Natural Language Learning*, Taipei, Taiwan, pp. 84--90.

On Construction of Polish Spoken Dialogs Corpus

Agnieszka Mykowiecka[*†], Krzysztof Marasek[†], Małgorzata Marciniak[*], Joanna Rabiega-Wiśniewska[*], Ryszard Gubrynowicz[†]

[*]Institute of Computer Science, Polish Academy of Sciences
J. K. Ordona 21, 01-237 Warsaw, Poland
[†]Polish Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland

Abstract

The paper concerns construction of the Polish spontaneous spoken dialogs corpus built within the LUNA project. It elaborates on the process of collecting conversations, their transcription and annotation at morpho-syntactic and concept levels. Corpus annotation is performed using a mixture of manual and automated techniques.

1. Introduction

In this paper we describe the process of construction and annotation of Polish spoken dialogs corpus. Collecting corpora of spontaneous speech in French, Italian and Polish is one of the goals of LUNA (spoken **L**anguage **UN**derstanding in multilingu**A**l communication systems) 6th Framework 33549 project. The general assumptions of this task are described in (Raymond et al., 2007).

The Polish corpus is maintained by two partners: the Institute of Computer Science Polish Academy of Sciences and the Polish-Japanese Institute of Information Technology. The task of building the corpus is twofold. First, a sub-corpus of 500 human-human conversation has been collected and annotated at obligatory levels (agreed by the project partners). The annotation scheme takes advantage of the previous works experience (Mengel et al., 2000; Cattoni et al., 2001). Now, a sub-corpus of 500 human-machine dialogs is being collected, see (Koržinek et al., 2008), and will be annotated with methods elaborated for the first sub-corpus. The chosen domain of conversations is public transport in Warsaw. A brief description of the corpus recordings of the first sub-corpus is presented in section 2.

In section 3. we present rules of dialogs transcription that are common for all three languages. Some of them were invented in order to preserve phenomena occurring in Polish dialogs. An account of morpho-syntactic annotation is given in section 4. The level of concept annotation is presented in section 5.

2. The Polish corpus collection

The corpus of human-human dialogs contains spontaneous dialogs recorded at Warsaw City Transportation Information Center in spring 2007. We have selected about 500 dialogs from around 12 thousands collected calls. The call center receives about 250 calls every day, but not all of them are relevant to our project's scope, and a part of them is of very low signal quality. An average conversation lasts 2 minutes. Before further processing, chosen calls were classified according to the main dialog topic. There are five classes of dialogs:

- STOPS – A caller asks about:
  - a stop nearest to the point in the city,
  - the name of the stop to get on or to get off,
  - a transportation mean (a bus or a tram) that stops at a given stop,
  - a stop appropriate for transfer between communication means.
- FARES_REDUCTION – A caller asks about rules concerning fares reduction in Warsaw.
- WHEN – A caller asks about timetables and travel duration.
- HOW_TO_GET – A caller asks how to get to a given place in the city or about details of transportation means' routes.
- DOES_IT_GO_TO – A caller asks if any transportation mean goes to a particular place in the city.

Naturally, many conversations refer to more than one described topic class. The classification was done in view of the dominating subject of the dialog. Still, the whole dialog is transcribed and annotated.

Topic class	Nb of dialogs	Nb of turns
DOES_IT_GO_TO	91	2667
HOW_TO_GET	140	4694
WHEN	99	2512
STOPS	51	1383
FARES_REDUCTION	83	1868
All dialogs	464	13124

Table 1: The distribution of the dialogs' topics

3. Data transcription

After the dialogs were chosen, an annotator converted them into texts using Transcriber (Barras et al., 1998). Every conversation was divided into turns referred to *a caller* and

an operator, respectively. The transcription output is an XML file which includes the dialog text and some metadata referring to articulation distortions, speaker and non-speaker noises, and time-stamps of the beginning and the end of each turn. General rules of transcription were agreed by all the project partners (Rodriguez et al., 2007) and they are presented in the context of Polish data in (Mykowiecka et al., 2007). To cover some phenomena significant for Polish dialogs, a few additional rules were defined. The most important Polish additions are [1]:

- It was agreed to transcribe spellings (and spelled acronyms) with capital letters tagged with a symbol *pron=SPELLED* as in *[pron=SPELLED-] PKO [-pron=SPELLED]*. However, it is typical of Polish to syllabify words, especially proper names. Therefore the symbol *pron=SYL* was introduced, e.g. *Bank [pron=SYL-] Narodowy [-pron=SYL] (National Bank)*.
- Acronyms pronounced as words are written in capitals, e.g. *PEKAES*. Some of acronyms in Polish undergo inflection, e.g. *ZUS*, *ZUS-u*, *ZUS-em*. In these cases, an inflection suffix is added to the basis in small letters, e.g. *ZUSu*, *ZUSem*, etc.
- Foreign words or acronyms are transcribed in their original orthographic form and tagged with a symbol *lang=* and the label of the language, e.g. *[lang=English-] Blue City [-lang=English]*. In case they are inflected by a Polish speaker, an inflection suffix appears directly after the closing tag *lang=*, e. g. *Plac [lang=English-] Wilson [-lang=English]a (Wilson's Square)*.
- A tag *lex=FIL* represents pause fillers, hesitations and articulatory noises as breath, laugh, cough, etc. In order to capture significant non-verbal answers as confirmation, which could be helpful at dialog acts annotation, it was decided to distinguish here a subtype marked with a tag *lex=FIL+*.

An example of the transcribed utterance is presented in Fig. 1[2].

user: [lex=FIL] chciałam się dowiedzieć jak długo jedzie autobus [silence] linii sto pięćdziesiąt siedem z ulicy Grójeckiej przy Bitwy Warszawskiej na Plac [lang=English-]Wilson[-lang=English]a

operator: [lex=FIL] do Placu [lang=English-]Wilson[-lang=English]a jedzie od dwudziestu sześciu do trzydziestu dwóch minut

Figure 1: Example of the text transcription

[1] All examples come from the dialog corpus

[2] Translation: **U:I** wanted to know how long it takes the bus 157 to go from Grójecka Str at Warsaw Battle Str to Wilson's Square; **O:** To Wilson's Square it goes 26-32 minutes

4. Morphosyntactic annotation of dialogs

After transcription, the set of dialogs was annotated morphologically with POS tags and inflectional characteristics. As the project concerns three different languages, the partners have adopted the recommendations of EAGLES (Leech and Wilson, 1996) for the morphosyntactic annotation and have defined for each language a core set of tags consistent with international standards.

There are several inflectional analyzers available for Polish (Hajnicz and Kupść, 2001) from which we have chosen AMOR (Rabiega-Wiśniewska and Rudolf, 2003). It was easy to extend it with the domain vocabulary and proper names; and to adapt it to the project annotation guidelines. The most important changes made to the analyzer's lexicon are described below:

- In the dialogs there are a lot of proper names, however, they split into a few POS classes. Originally, the AMOR contained only nominal proper names, now there are also proper adjectives, e.g. *Afrykańska (African)*, *Centralna (Central)*, proper prepositions, e.g. *Przy (At)*, and proper numerals, e.g. *Siedem (Seven)*. At present, the set of proper names in the corpus consists of 6500 words that belong to 820 lemmas.
- Sometimes a caller is not sure what the name of the street (a building etc.) is or how to pronounce its name. Names that were not recognized by an operator or heavily distorted are transcribed according to their real pronunciation. At the morphological level such words get a POS tag 'PropName' but no additional characteristics. The aim is to be able to represent every, correct and mistaken, proper name in the corpus. Compare the examples below:

  (1) <w id='37' word='Bliżna' lemma='-' POS='PropName' morph='-' />

  (2) <w id='58' word='Wólkę' lemma='Wólka' POS='Np' morph='acc.sg.fem' />

- The spoken language is rich in colloquial expressions (*se* instead of *sobie*) and ungrammatical (*poszłem* instead of *poszedłem*) word forms. Those appearing regularly and frequently in collected texts, were added to the lexicon.

The automatic morphological analysis gives approximately three different interpretations per word. As there is no Polish tagger which accounts for proper names and which was tested on speech data, disambiguation of morphological tags is done manually. However, it is planned to train a tagger on a sample of the annotated corpus in the next stage of the project. The morphological analysis results in obtaining for every word a set of tags: *id*, *word*, *lemma*, *pos* and *morph*. They are stored in XML files in a format presented in Fig. 2[3].

[3] Translation: I wanted to ask (about) bus 143 from the direction of Ursynów

chciałam zapytać autobus sto czterdzieści trzy z kierunku Ursynowa
<words>
<w id='10' word='chciałam' lemma='chcieć' POS='VV' morph='1.sg.fem.past.ind.imperf' />
<w id='11' word='zapytać' lemma='zapytać' POS='VV' morph='inf.perf' />
<w id='12' word='autobus' lemma='autobus' POS='Nc' morph='nom.sg.m3' />
<w id='13' word='sto' lemma='sto' POS='NUM' morph='nom.nm1' />
<w id='14' word='czterdzieści' lemma='czterdzieści' POS='NUM' morph='nom.nm1' />
<w id='15' word='trzy' lemma='trzy' POS='NUM' morph='nom.nm1' />
<w id='16' word='z' lemma='z' POS='PreP' morph='-' />
<w id='17' word='kierunku' lemma='kierunek' POS='Nc' morph='gen.sg.m3' />
<w id='18' word='Ursynowa' lemma='Ursynów' POS='Np' morph='gen.sg.m3' />
</words>

Figure 2: Example of the morphological annotation

Morphologically annotated texts of dialogs are next segmented into elementary syntactic chunks. The aim of syntactic description is to group the words into basic nominal phrases and verbal groups. As there exists no chunker suitable for the analysis of Polish spoken texts, a program used in the project was designed especially for the purpose. In order to find phrases within an utterance of one speaker, information about turns is used. The parser uses also some domain knowledge, which helps for example to recognize transportation line numbers. The following phrase: *autobusy pięćset dwanaście sto siedemdziesiąt cztery 'buses five hundred twelve one hundred seventy four'* can be theoretically divided in many ways, but we know that all buses in Warsaw have three-digit numbers so we can divide the phrase properly into two numbers: *pięćset dwanaście 'five hundred twelve'* and *sto siedemdziesiąt cztery 'one hundred seventy four'*. The chunker also recognizes compound verbal phrases, *będzie jechać 'will go'*, and nominal phrases (without prepositional modifiers), *następny przystanek autobusowy 'the next bus stop'*. For these phrases it indicates the main word i.e., the word semantically most significant. In the previous examples it is *jechać 'go'* and *przystanek 'stop'* respectively. In the case of a nominal phrase it coincides with the head of the phrase. The syntactic segmentation of previously morphologically annotated example is shown in Fig. 3.

5. Semantic annotation of dialogs

Semantic annotation of the dialogs consists in assigning attributes and their values to phrases. The principles of the annotation in our project are similar to the attribute annotation in MEDIA corpus (Hardy et al., 2003). The set of attributes was defined specially for the project and contains general transportation system features, some details on Warsaw public transport and some concepts related to

chciałam zapytać autobus sto czterdzieści trzy z kierunku Ursynowa
<chunks>
<chunk id='8' span='word_10' cat='VP' main='word_10' />
<chunk id='9' span='word_11' cat='VP_INF' />
<chunk id='10' span='word_12' cat='NP' main='word_12' />
<chunk id='11' span='word_13..word_15' cat='NUM' />
<chunk id='12' span='word_16' cat='PP' />
<chunk id='13' span='word_17' cat='NP' main='word_17' />
<chunk id='14' span='word_18' cat='PN' />
</chunks>

Figure 3: Example of the syntactic annotation

different type of questions occurring in the recorded conversations. The domain model specification started with defining a general ontology of public transport in OWL (`http://www.w3.org/TR/owl-ref/`). On its basis, a set of attributes representing concepts was defined. It contains simple notions:

- bus, tram, metro lines, routes, their ends and stops;
- places in the city: districts, streets, squares, parks, important buildings;
- fare reduction's concepts;
- basic time points specifications.

And more complex ideas like:

- trips' beginnings and endings,
- trips' durations and other time specifications,
- questions concerning different attribute values.

Within the chosen domain there are a lot of proper names (there are over 4 thousands names of streets, buildings, city districts, etc). Their recognition is not easy, see (Mykowiecka et al., 2008b):

- A lot of names are inflected, e.g. street names: *Francuska*, *Francuskiej*, *Francuską* etc. *(French)*, building names: *Teatr Dramatyczny*, *Teatru Dramatycznego* etc. *(Dramatic Theater)*
- For many names there is more than one variant, e.g. names of persons in the street names are often omitted: *Krasińskiego* instead of *Zygmunta Krasińskiego*.
- Complex names are simplified: *Bitwy Warszawskiej 1920 r. (Warsaw Battle in 1920)* to *Bitwy*, *Plac Powstańców Warszawy (Warsaw Uprising Square)* to *Plac Powstańców*.
- Proper names are sometimes ambiguous as bus stops have frequently the same names as streets, squares or buildings names where they are situated.

In Polish, nouns, adjectives and numerals undergo inflection, so the recognized proper names had to be lemmatized. As the final quality of the annotation was the primary target, we introduced proper names elements into our inflectional dictionary which enabled us to obtain lemmas for all

names elements. We also prepared a lexicon which relates sequences of basic forms of name elements to the basic forms of the entire names (Mykowiecka et al., 2008b).
The next planned step is to annotate collected dialogs with concept names. To realize this task we use rule-based Information Extraction approach. Therefore, the annotation is done automatically on the basis of manually created rules that define patterns for recognizing attributes and their values. At the moment there are 950 rules that recognize 134 attributes. Most attributes can have only a few possible values but there are a few attributes that can have many values like: destination, bus number, time description. An example of the output is shown in Fig. 4.

chciałam zapytać autobus sto czterdzieści trzy z kierunku Ursynowa
<concept id="4" span="word_10" attribute="Action" value="Request" />
<concept id="5" span="word_12..word_15" attribute="BUS" value="sto czterdzieści trzy" / >
<concept id="6" span="word_16..word_18" attribute="SOURCE_DIR_TD" value="Ursynów" / >

Figure 4: Example of the concept annotation

The first evaluation of the set of rules was done on the 26 dialogs and showed the overall concept error rate at the level of 20.5% (Mykowiecka et al., 2008a).

6. Summary

In the paper we described the collection process and the annotation practice underlying the first multi-level annotated spontaneous speech corpus of Polish. The corpus is being developed as a part of the LUNA project.
The procedures adopted within the project combine manual and automatic approach. The automatically obtained morphological annotations are disambiguated manually and randomly verified. The annotation on the syntactic level is automatic but only very basic chunks are built, so the number of introduced errors is rather low. Automatic semantic annotation is much more difficult, but as the size of the corpus is not too big, this annotation level will be checked manually.
In the next step we are going to annotate the collected data with predicates' roles, coreferences and dialog acts.

7. References

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

R. Cattoni, M. Danieli, A. Panizza, V. Sandrini, and C. Soria. 2001. Building a corpus of annotated dialogues: the ADAM experience. In *Proceedings of the Corpus Linguistics 2001 conference*, pages 109–119, Lancaster, UK.

E. Hajnicz and A. Kupść. 2001. *Przegląd analizatorów morfologicznych dla języka polskiego*. Wydawnictwo IPI PAN, Warszawa.

H. Hardy, K. Baker, H. Bonneau-Maynard, S. Rosset, L. Devillers, and T. Strzalkowski. 2003. Semantic and dialogic annotation for automated multilingual customer service. In *Proceedings of Eurospeech-2003*, pages 201–204, Geneva, Switzerland.

D. Koržinek, Ł. Brocki, R. Gubrynowicz, and K. Marasek. 2008. Wizard of Oz Experiment for a Telephony-Based City Transport Dialog System. In *Proceedings of the IIS 2008 Workshop on Spoken Language Understanding and Dialogue Systems*, Zakopane, Poland. Springer Verlag. To appear.

G. Leech and A. Wilson. 1996. Eagles. Recommendations for the morphosyntactic annotation of corpora, EAG-TCWG-MAC/R. Technical report, ILC-CNR, Pisa.

A. Mengel, L. Dybkjaer, J.M. Garrido, U. Heid, M. Klein, V. Pirrelli, M. Poesio, S. Quazza, A. Schiffrin, and C. Soria. 2000. *MATE Dialogue Annotation Guidelines*. MATE Deliverable 2.1.

A. Mykowiecka, K. Marasek, M. Marciniak, R. Gubrynowicz, and J. Rabiega-Wiśniewska. 2007. Annotation of Polish spoken dialogs in LUNA project. In *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of 3rd Language & Technology Conference. October 5-7, 2007, Poznań, Poland*.

A. Mykowiecka, M. Marciniak, and K. Głowińska. 2008a. Automatic Semantic Annotation of Polish Dialog Corpus. In progress".

A. Mykowiecka, M. Marciniak, and J. Rabiega-Wiśniewska. 2008b. Proper Names in Polish Dialogs. In *Proceedings of the IIS 2008 Workshop on Spoken Language Understanding and Dialogue Systems*, Zakopane, Poland. Springer Verlag. To appear.

J. Rabiega-Wiśniewska and M. Rudolf. 2003. Towards a Bi-Modular Automatic Analyzer of Large Polish Corpora. In R. Kosta, J. Błaszczak, J. Frasek, L. Geist, and M. Żygis, editors, *Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages – FDSL IV, held at Potsdam University, November 28-30th, 2001*, pages 363–372.

Ch. Raymond, G. Riccardi, K. J. Rodriguez, and J. Wisniewska. 2007. The LUNA Corpus: an Annotation Scheme for a Multi-domain Multi-lingual Dialogue Corpus. In R. Artstein and L. Vieu, editors, *Decalog 2007: Proceedings of the 11th Workshop on Semantics and Pragmatics of Dialogue, Trento, Italy, 30 May – 1 June 2007*, pages 185–186, Trento, Italy.

K. J. Rodriguez, S. Dipper, M. Götze, M. Poesio, G. Riccardi, C. Raymond, and J. Rabiega-Wiśniewska. 2007. Standoff coordination for multi-tool annotation in a dialogue corpus. In *Proceedings of the Linguistic Annotation Workshop*, pages 148–155, Prague, Czech Republic. Association for Computational Linguistics.

A RESTful interface to Annotations on the Web

Steve Cassidy

Centre for Language Technology
Department of Computing
Macquarie University
Sydney
steve.cassidy@mq.edu.au

Abstract

Annotation data is stored and manipulated in various formats and there have been a number of efforts to build generalised models of annotation to support sharing of data between tools. This work has shown that it is possible to store annotations from many different tools in a single canonical format and allow transformation into other formats as needed. However, moving data between formats is often a matter of importing or exporting from one tool to another. This paper describes a web-based interface to annotation data that makes use of an abstract model of annotation in its internal store but is able to deliver a variety of annotation formats to clients over the web.

1. Introduction

There has been considerable work in recent years on building generalised models of annotation and defining interchange file formats such that data can be moved between tools. This work offers the hope that annotation data can be released from the project or discipline specific dungeons it is often locked in due only to the difficulty in understanding data from foreign tools. However, while data sits in files on a researcher's disk it remains hard to discover it and get access, let alone collaborate on the development of a corpus.

A second problem is that annotations, even in well known and widely distributed corpora, can't be cited in the same way that we might cite a result in a research paper. Exceptions to this are cases where the authors of a corpus have taken care to define reference codes for segments of the corpus (e.g. the line numbers of the Brown corpus).

We propose that both of these problems can be addressed by defining a well structured interface to corpora and annotations over the web. Such an interface would have the advantage of defining a public URI for every corpus and annotation within the corpus that could be cited in a research paper. It could also allow widespread access to data from remote locations to facilitate collaboration and sharing of annotations. Using the infrastructure of the web allows technologies such as caching and access control to be layered on top of the basic interface.

This paper describes the core of a web based interface to corpora. At present this interface only supports reading of annotations from a central annotation store. However, the design has been built with a view to enabling read/write access to data over the web.

2. Background

A number of proposals have been made in recent years for generalised data models for Linguistic annotation. These models provide an abstract representation of annotation data that subsumes practices in the majority of research areas where language data is annotated or marked up in some way. While there are some differences in the proposals they are largely compatible with each other; this is perhaps not surprising since they are designed to support transformation to and from a similar set of end-user formats.

Two examples whose design is particularly focussed on interchange of annotations between formats are Annotation Graphs (Bird and Liberman, 2001) and the Linguistic Annotation Framework (Ide and Romary, 2007). Both are structured as directed graph structures with annotations as nodes in the graph; annotations are distinct objects carrying arbitrary feature structures (attribute-value pairs) and may be related to each other by many kinds of relations. Both formats make use of so called *stand-off* markup where the annotations are stored separately to the primary data itself. Locations in the primary data are indicated by pointers; for audio and video data these are time values, for textual data they can be character offsets or XPointer references.

The use of annotations that point into primary data instead of being embedded in it was motivated in part by the need to be able to represent overlapping hierarchies. Since XML, a common format used for annotation, can only directly represent a single hierarchy, a solution that separated the different hierarchies into different XML files was used. A side effect of this change is that annotations can be managed separately to the primary data, paving the way for an annotation architecture that uses an abstract interface rather than an application specific file format.

The work described here develops on this idea of an abstract interface to an annotation store as an alternative to reading and writing annotation files. Instead of thinking of annotations as elements in files and corpora being collections of these files we abstract these ideas to make all of these things *resources* within an annotation store. Internally in our system, we store annotations as assertions in an RDF triple store and provide an abstract interface for creation, deletion and query of annotation data. The proposal in this paper though, does not make any assumptions about the kind of store that is used; only that it supports the idea of annotations as separate entities. This is true of the Annotation Graph system for example and will generally be true of any tool that displays and manipulates annotation data.

This work has been implemented in a development system that is being used as part of a larger project to support collaborative annotation on language resources. A demonstration of the service may be available at the URI `http://dada1.ics.mq.edu.au/` depending on the

current status of the software.

This paper first highlights the capabilities of the HTTP transport layer, then develops the design of an interface to annotation data over HTTP and finally describes some extensions to this interface that we are currently exploring.

2.1. HTTP and the Web

The Hypertext Transfer Protocol (HTTP) is the base protocol of the World Wide Web and defines the conversation that takes place between a web server and a client such as a web browser. The original web was conceived as a read/write medium and the design of HTTP reflects this in the provision of actions for creating, updating and deleting resources as well as retrieving them. Until recently, the two-way nature of HTTP was not widely exploited but the development of web services following the REST (Representational State Transfer) architecture (Fielding, 2000) has highlighted the power of the original design.

The REST view of the web is as a means to provide access to *resources* that are identified by unique addresses (the Uniform Resource Identifier or *URI*). Resources are accessed through a constrained set of operations for transferring state information between client and server; be it a GET request to retrieve the current state of a resource or a POST request to update it. State information can range from the content of an HTML web page to the contents of a shopping cart or a value in a data store. It is also common to differentiate the internal form of the resource from the surface form that is transferred over the network. Hence, the current temperature on a web accessible device could be transferred as a simple text file, an XML document or an HTML web page. The form of the response is determined by the request that is sent from client to server.

The most common request in HTTP is *GET* which retrieves the current state of a resource. A *POST* request is often used to submit form data to a web service but in general is intended to submit data to a resource and can be interpreted as creating a subsidiary resource (e.g.. a file within a folder) or updating an existing resource. Less commonly used are *PUT* and *DELETE* which create new resources and delete them; since these generally imply creating and deleting files on a server they are not generally implemented for security reasons. HTTP supports a few other kinds of request and there are a number of extensions to the protocol to support additional applications (for example WebDAV to support remote file stores).

While HTTP is an inherently open protocol, it is able to support secure and authenticated access to resources. Encrypted connections using the Secure Sockets Layer (SSL) mean that traffic over the network cannot be intercepted. Authentication can be layered on top of the basic HTTP protocol using cookies - additional headers exchanged with every transaction. In combination, these can provide secure access to resources mediated via appropriate authentication and authorisation controls. This is an important feature for working with language resources which often need to be protected from general access; some work relating to this will be outlined later in the paper.

3. Annotations on the Web

3.1. What gets a URI?

The first question in designing an interface to annotations over the web is that of designing the *URI space* – the logical structure of URIs used to retrieve and modify annotations. Closely tied to this is the question of what should have a URI of its own. Our proposal is for a three-level abstraction of resources from the annotation store: *corpora*, *annotation sets* and *annotations*. We also include an explicit representation of an annotation end point (start or end time or pointer to a document location) called an *anchor*.

Each of these kind of resource is identified by a unique URI. This is both a canonical name for the resource and a means of accessing a description of it over the HTTP interface.

Corpora represent collections of documents whose annotations are stored on the server. A corpus might be a traditional curated collection such as the TIMIT or BNC corpora, or an ad-hoc collection by a single researcher. A corpus has a URI of the form `http://example.org/corpora/NAME` where *NAME* is a symbolic name for the corpus[1] A collection of corpora housed on a given server will also have a URI (`http://example.org/corpora/` here) that could be used to discover what data is available on this server.

Annotation Sets are containers for the annotations on a single document or media file. It is common to have this level of abstraction when using a tool such as ELAN (Wittenburg et al., 2006) or Transcriber (Barras et al., 1998) that stores all annotations on a media file in a single XML file. Annotation sets might correspond to more than one of these XML files in the case when multiple kinds of annotation are stored in different files. An annotation set is always part of a corpus and has the corpus URI as a prefix of its URI which is of the form `http://example.org/corpora/NAME/ASID`; ASID here is a unique identifier for the annotation set.

Annotations are the individual annotations that make up an annotation set. A single annotation might store the part of speech of a word or a phonetic label for a segment of a speech signal. The URI of an annotation has an annotation set URI as a prefix: `http://example.org/corpora/NAME/ASID/ANNID` where ANNID is an annotation identifier.

Anchors are the endpoints of annotations and are represented as explicit resources to allow them to be shared between annotations. For example, one anchor may be the end point of one annotation and the start point of a second. Anchors appear in some form in many annotation formats including Annotation Graphs (Bird and Liberman, 2001) and ELAN (Wittenburg et al., 2006) which calls them *time slots*. Since anchors are also contained within annotation sets, they also have a URI that has an annotation set URI as a prefix: `http://example.org/corpora/NAME/ASID/ANCHID` where ANCHID is an anchor identifier.

[1] In these examples we use a common prefix of `http://example.org/corpora/` in all URIs, this is arbitrary and will depend on the server used to store the corpora.

Each of these kind of resources can be described by a feature structure (in the TEI or ISO 24610-1 sense (M. Laurent Romary and TC 37/SC 4/WG 2, 2006)) containing information about the resource. This structure supports attaching feature sets to any level of detail from the corpus to the annotation itself. Feature values can include relations between resources; these are easily expressed since each resource has a unique URI that can appear as the value of a feature. The vocabulary used in defining features is of course important; we note that the Linguistic Annotation Framework (Ide and Romary, 2007) is directly addressing this need in setting up standards for a Data Category Registry that would allow mapping of feature names between resources.

3.2. Responses to URIs

Having said that these resources have unique URIs that can be published and accessed to allow sharing of annotations, we still need to define what exactly will be returned if someone enters one of these URIs into a web browser.

By default, the response to a request for a URI from the annotation server will be an HTML representation of the resource being referenced. This means that someone can access one of these published URIs in a web browser and see a human readable representation of the corpus, annotation set or annotation. The actual representation that is returned is the concern of the implementer of the server and need not be uniquely defined; for example, a server that holds annotations of video data might be able to serve a representation of an annotation set as a page with the video embedded alongside a browseable version of the annotations similar to that developed by the EOPAS project (Thieberger and Schroeter, 2006) for ethnographic data.

Our current implementation includes links to all of the subordinate resources in the HTML representation. So, the page generated for a corpus links to all of the annotation sets in the corpus while the annotation set links to all of the annotations. The page for an annotation includes all of the properties associated with the annotation and links to any other associated annotations (e.g.. parents, dependancies, etc.).

3.2.1. Content Negotiation

A little used option in HTTP is the ability to have the web browser request certain types of content when requesting a resource. For example, I can ask for `http://example.org/data` while saying that I will accept plain text or PDF. The server can then respond with whichever of these it is able to produce. This process is called *content negotiation* and is not widespread partially because of the lack of support for it in all browsers.

The web service described here makes use of content negotiation to serve different kinds of content to different clients. If the client is a conventional web browser, the server will generate HTML descriptions of resources; on the other hand if the client is an annotation tool, it can request data, for example, in ELAN eaf format.

Content negotiation will allow us to serve different representations of each of the resources to different kinds of client. We can, for example, return a version of an annotation set in the format required by an annotation tool such as ELAN or Transcriber. In this way, the interface can realise the format conversion functionality that is at the core of standards such as LAF (Ide and Romary, 2007) or AG (Bird and Liberman, 2001) transparently. The same annotation could then be accessed by an ELAN user and a Transcriber user without having to distribute two distinct versions of the annotation or go through any explicit conversion process.

In some situations, content negotiation is not possible - for example when including links in a web page or when dealing with older HTTP client software. In these cases it is possible to achieve the same end by augmenting the URI of a resource with a query string indicating the type of representation required. So, to retrieve an ELAN format representation of an annotation set one could retrieve `http://example.org/corpora/andosl/foobar?format=application/xml+eaf` (the exact keyword and format indicator needs standardisation, this example is included to illustrate this capability).

3.2.2. Low Level Access

There is a third possibility though that offers to realise the full potential of the web based annotation store. That is to return a form of each resource that can form the basis of a read/write interface to the store. The idea here is that instead of reading and writing annotation files in an XML format, a tool could query the server directly for information about the annotations on a document or media file. To support this, the response to a request for an annotation set could be a simple XML list of the URIs of the annotations, perhaps with a small amount of data from each such as a label or start and end times. Using this, an annotation tool could determine which annotations are of interest and query the server for more information about each. The response to a request for an annotation could be a simple XML representation of the annotation as a feature structure.

This kind of server would allow updates to be made to annotations using the same kind of messages sent from the client to the server. To add a new annotation to an annotation set the client would make a POST request to the annotation set URI `http://example.org/corpora/as123/` with a request body containing the feature structure for the new annotation. In response to the POST request, the server creates a new annotation and returns a HTTP response confirming that it was created with the URI of the new annotation. Similarly, a POST request to an existing annotation URI has the effect of updating the annotation. Finally, the DELETE request to an annotation or annotation set URI can be used to remove the corresponding resource. These requests can be used by an annotation tool to directly manipulate the annotations stored on the server rather than working through any kind of file format.

4. Building Upon the Interface

One of the primary advantages of defining a web based interface based on HTTP access to resources is that the existing infrastructure of the web can be leveraged to add new functionality with little extra effort. The web is a very mature family of technologies and many issues around effi-

cient, secure distribution of data have been addressed in general purpose technologies layered on top of HTTP. A few of the possibilities are outlined here.

4.1. Caching and Proxies

A significant problem with providing remote access to resources such as annotations or primary linguistic data is the time lag between a request and the response being delivered over the network. This would be an immediate barrier to adoption of this kind of technology in some applications which require very fast access to data. This is not a problem unique to annotation and since we have layered our interface on top of HTTP we can take advantage of HTTP caches to speed access to frequently accessed data.

An HTTP cache acts as a proxy between the client and server such that most transactions occur just as they would if no proxy were in place. The cache will however, remember the responses to some requests and, if configured appropriately, will return a local copy of the response if it is requested again. A cache can be run on an individual machine or within an organisation where the requests from all users within a research group would be cached together, speeding access to the resources being used by the group.

While a generic HTTP proxy cache such as Squid (`http://www.squid-cache.org/`) can be used in this way there is scope for writing a special purpose proxy cache that knows about usage patterns of annotation data. Such a proxy could pre-fetch annotations that might be used in the near future.

While caching files can be one important function of a proxy server, it can also fulfil another role in this context. A proxy acts as a mediator between the client and one or more servers and as such can federate access to multiple annotation servers. One could imagine a departmental or institutional proxy supporting access to many servers via a common cache while also serving local resources transparently to users. A network of such proxies could effectively provide distributed, redundant, storage of annotation data.

4.2. Authentication and Authorisation

As described so far, all resources are available to anyone on the internet to read and possibly update; this is clearly not what would be required by most researchers and for many language resources which must be restricted in some way. Again, we can make use of existing technology on the web to layer authentication and authorisation on top of the HTTP interface described above.

HTTP provides a simple authorisation scheme as part of the protocol which would allow resources to be password protected. Web servers such as the Apache server allow configuration settings that protect different URIs with different user names and passwords and this could be used to restrict access to distinct groups of users. Similarly, the operations that update an annotation (PUT, POST, DELETE) can be given different levels of password protection using standard server settings.

A more sophisticated solution has been developed for applications that require more complex authorisation rules to be enforced. The XACML (XML Access Control Markup Language, `http://www.oasis-open.org/committees/xacml/`) standard allows complex access control rules to be written which take into account external factors such as the date or file properties such as size or source of data. We are currently investigating the use of XACML in conjunction with our annotation server to provide fine grained access control to both annotations and primary data. For example, one might want to restrict access to part of a recording based on the identity of a speaker in that recording. XACML allows the rules to be written to express this restriction; we are now looking at how the server infrastructure needs to be configured to put this into practice.

Rather than require every server to maintain passwords and user credentials for authorised users, the Shibboleth system `http://shibboleth.internet2.edu/` implements a federation of identity providers such that a user can be authenticated against their home institution. An identity federation such as this would allow groups of researchers to be granted access to resources based on, for example, their host institution or membership of some project. We are currently working with the RAMS project at Macquarie `http://www.melcoe.mq.edu.au/projects/RAMP/` on integrating our server with the Muradora data repository `http://www.muradora.org/`, a version of the popular Fedora server that integrates Shibboleth and provides a web based interface to building XACML policy documents. Our work here aims to illustrate how access to source data, meta-data and annotations can be mediated by appropriate authentication and authorisation.

4.3. Version Management

Annotations are not often static; errors are found and corrected and new versions of corpora are published. Especially in the context of a collaborative annotation tool it must be possible to manage different versions of annotations and integrate version control operations such as roll-back of changes or generating patch sets to send to other users.

As part of our work on the back-end RDF annotation store we have developed a version control system for RDF triple stores that is designed to support these operations on annotation data (Cassidy and Ballantine, 2007).

If the URIs published for annotation sets and annotations are to be useful they must be constant over time. That is, I must be able to publish a reliable URI for the annotation set that I used for a given study, not one which points to the most recent version of that annotation. Hence we must be able to include revision information in the URI.

While we have not yet integrated our version control system with the HTTP interface, there are a number of possible ways in which one could refer to historical versions of data via a URI. One simple option is to prefix the corpus name with a revision identifier: `http://example.org/corpora/101029/andosl/msdjc001/ann0293` - where `101029` uniquely identifies the revision of the annotation that is being referred to. The most recent annotation could still be referenced with out the version identifier but the longer style could be used where longevity of reference is required.

4.4. Mashups of Data and Annotations

One of the defining features of the recent boom of applications on the web has been the growth of *mashups* built from data provided by different sources. A common component of these is Google Maps `http://maps.google.com/` which can be used to visualise geographic data available on the web. The open nature of the web and the fact that data is available in well defined formats using well defined interfaces means that data can be re-purposed into applications that might not have been conceived by the original authors.
In the annotation domain there are many possibilities for mashups that might combine annotation data with other widely available data sources such as WordNet, Wikipedia etc. Annotations might also be combined with each other; for example, merging different styles of annotation or augmenting annotations with data from lexical resources. The important point here is that this capability comes for free once we adopt an open, well defined interface using well understood technology.

5. Conclusion

This paper has given a brief overview of the design of a web based interface to an annotation store. The design uses the REST approach to make corpora, annotation sets and annotations available as first class resources on the web.
This approach changes the way that annotation tools work with annotation data. Instead of relying on local storage of data in files, tools can work with an annotation store through an abstract interface. The fact that this interface uses the HTTP protocol of the web means that the store can be remote and shared between users. By layering authentication, authorisation, caching and other standard HTTP technologies on top of the interface we can add additional functionality to the interface.

6. References

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC),*, pages 1373–1376, Granada, Spain, May.

S. Bird and M. Liberman. 2001. A Formal Framework for Linguistics Annotation. *Speech Communication*.

Steve Cassidy and James Ballantine. 2007. Version control for rdf triple stores. In *ICSOFT 2007*, Barcelona, Spain, July.

Roy Thomas Fielding. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. University of California, Irvine,.

N. Ide and L. Romary. 2007. Towards International Standards for Language Resources. In L. Dybkjaer, H. Hemsen, and W. Minker, editors, *Evaluation of Text and Speech Systems*, pages 263–84. Springer.

M. Laurent Romary and TC 37/SC 4/WG 2. 2006. Language resource management - Feature structures - Part 1: Feature structure representation. In ISO 24610–1.

Nicholas Thieberger and Ronald Schroeter. 2006. EOPAS, the EthnoER online representation of interlinear text. In Linda Barwick and Nicholas Thieberger, editors, *Sustainable Data from Digital Fieldwork*, pages 99–124, University of Sydney, December.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN : a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Multiple Purpose Annotation using SLAT — Segment and Link-based Annotation Tool —

Masaki Noguchi†, Kenta Miyoshi†, Takenobu Tokunaga†
Ryu Iida‡, Mamoru Komachi‡, Kentaro Inui‡

† Department of Computer Science, Tokyo Institute of Technology
Tokyo Meguro Ôokayama, 152-8552, Japan
{mnoguchi,kmiyoshi,take}@cl.cs.titech.ac.jp

‡ Graduate School of Information Science, Nara Institute of Science and Technology
Nara Ikoma Takayama 8916-5, 630-0192, Japan
{ryu-i,mamoru-k,inui}@is.naist.jp

Abstract

In recent years, the use of large scale corpora in NLP applications, such as statistical parsing, has become prominent. As their use gained credibility, naturally so did the types of information they provided. There exist today many groups that create corpora: ANC, SFB at the University of Potsdam, just to name a few. In many cases these groups also provide specialized annotation tools for their corpora. However, these tools are just that: specialized, i.e. designed to work with a very specific annotation definition, without flexibility in mind. In the early stages of a project, often times the specification for annotating changes. This makes it difficult to use a tool with such rigid boundaries. In this paper, we propose a browser-based annotation tool SLAT, which allows for easily adding and customizing annotations. We also explain the steps involved in customizing SLAT to meet a user's project needs.

1. Introduction

In recent years, the use of large scale corpora in NLP applications, such as statistical parsing, has become prominent. As their use gained credibility, naturally so did the types of information they provided.

There are many projects which construct corpora, such as ANC[1] and Sonderforschungsbereich (SFB) on information structure at the University of Potsdam[2] just to name a few. The annotation of sentences by hand is not only extremely time consuming, but also leads to various kinds of errors. These errors combined with other user-entered biases have a large effect on the performance (and subsequent evaluation) of systems trained on these corpora. Thus, the information provided by corpora must be both accurate and consistent. To this end, annotation tools for simplifying and constraining human input have been developed in various projects, and have decreased the costs of constructing corpora. These tools are developed to work with a very well defined annotation specification. In the early stages of a project often times the specification for annotating will change, making it difficult to use a tool with such rigid boundaries. The format for storing information also differs by tool, so their data is not immediately interoperable. The conversion of one format to another is required each time an experiment is conducted or a method evaluated.

In the next section, we briefly review some existing annotation tools and then describe our motivations for developing a new annotation tool. We introduce SLAT [sléit] (Segment and Link-based Annotation Tool), aimed at satisfying these motivations and briefly explain its features. Lastly, we summarize this paper and describe future work.

[1] http://www.americannationalcorpus.org
[2] http://www.sfb632.uni-potsdam.de/

2. Requirements for Annotation Tools

Stefanie Dipper et al.(Dipper et al., 2004) compared existing tools that use XML as their data storage format. They compared twelve individual research projects from several disciplines, having corpora that mostly consisted of 5 types of annotations: semantic, discourse and focus annotations, as well as diachronic data and typology. To manage these types of annotations, they described seven requirements for annotation tools: diversity of data, multi-level annotation, diversity of annotation, simplicity, customizability, quality assurance and convertibility. First three relate to data annotation while the latter four relate to the usability of the annotation tool. They compared five annotation tools to test the validity of these criteria: TASX Annotator[3], EXMARaLDA[4](Thomas, 2001), MMAX[5](Müller, 2006), PALinkA[6](Orăsan, 2003) and Systemic Coder[7].

3. Requirements during the Early Stages of a Project

As presented in the previous section, an annotation tool must satisfy these requirements to be successful in corpus annotation. Previously developed annotation tools have mostly focused on the usability of the system regarding the annotation task itself, i.e. how easy/difficult it is to add/remove annotations. Usability is clearly important. In designing an annotation tool, however, it is also crucially important to take the while demands of a corpus project, which typically not only annotate text but also designing the tag set and evaluate and maintain the resultant corpus,

[3] http://tasxforce.lili.uni-bielefeld.de/
[4] http://www.exmaralda.org/
[5] http://www.eml-research.de/english/research/nlp/download/
[6] http://clg.wlv.ac.uk/projects/PALinkA/
[7] http://www.wagsoft.com/Coder/

into account as design issues. More specifically, at least the following three issues should be addressed so that the tool can effectively support a project even during its initial unstable stages:

1. **Cost to install an annotation tool**
   Creating a corpus involves a large number of hands engaging in the task of annotation. It is particularly the case for those unfamiliar with computers that merely installing an annotation tool can become a burden.

2. **Variation of data schemes for each annotation task**
   Past annotation tools have been developed with a specific annotation scheme in mind, making it unsuitable for other types of annotation. A multipurpose annotation tool must use a flexible data scheme that can incorporate various types of annotation, and must have an interface adaptable to various annotation tasks.

3. **Quality of the corpus**
   As previously mentioned, the initial phases of a project are often filled with adjustments to how a corpus will be annotated. Since typical annotators work individually while referring to a specification, this period can result in poor consistency. These errors affect the quality of a corpus which in turn affects the performance and subsequent evaluation of a system.

We introduce SLAT (Segment and Link-based Annotation Tool), in the next section. For tackling the first issue, we adopt a client/server architecture. We present annotation abstraction for resolving the second issue and discuss some already-developed annotation tools and their own implementations. Finally, we summarize our findings and briefly touch upon the third issue enumerated above.

4. SLAT

SLAT is a web-based annotation tool that employs a client/server architecture. With the ubiquitousness of the internet, this means that SLAT can be accessed almost anywhere; the only prerequisite for beginning annotation is having access to the URL via a browser. This also serves to reduce the cost and time of installation on an annotator's machine. The server-end of SLAT is composed of a computer running a database and a PHP-enabled web-server. The SLAT server stores all documents to be annotated, annotation information and customized user configurations.
In this section, we first propose an abstraction of annotations using segments and links, which allows SLAT to adapt to many different annotation tasks. We then address the interface issues, detailing the components of the current SLAT interface, and finally demonstrate how SLAT can be easily customized.

4.1. Abstraction of Annotations

To explore a universal data scheme applicable to various types of annotations, we discuss the abstraction of annotations using a simple POS annotation example shown in Figure 1. In this example, annotation is carried out by affixing POS and named entity tags to specific regions of text, called *segments*. Thus, "John" is annotated as N and N-PER and "New York" as N and N-LOC etc. Relations between segments are then identified, such as coreference or a certain semantic role. This is called *linking*. Using this abstraction, almost any annotation can be represented. SLAT adopts stand-off annotation, i.e. all annotated data is stored separately from the original data.

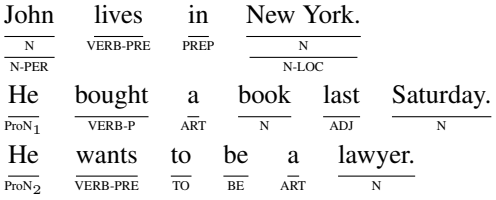

Figure 1: An example of POS annotation

4.1.1. Segments

When annotating a text, it is important to both indicate the particulars of a region as well as its relation to other parts of the text. A segment is indicated by marking the starting and ending offsets of a region. For representing this information, tags are inserted into the text. A fragment of the text can be multiple segments such as "John" and "New York" in Figure 1. Furthermore, segments can be nested and overlap, such as 'XXX YYY ZZZ'.

4.1.2. Links

As mentioned above, segments may have several types of relations to one another, e.g. "John" and "he" (coreference), or "bought" and "a book" (semantic role). All relations have at least two properties: transitivity and directionality. By combining these two properties, we can divide relations into four general groups:

1. **transitive and directed** E.g. "car"→"door"→"glass", *part-of* relations belong to this group. Temporal relations between events also belong to this group.
2. **transitive and undirected** Coordination and coreference, such as the relations between "John (N-PER)", "He (ProN$_1$)" and "He (ProN$_2$)"in Figure 1.
3. **non-transitive and directed** Semantic role labeling, e.g. the relation between "bought (VERB-P)" and "book (N)" belongs to this group.
4. **non-transitive and undirected** Relations in this group represent a special case only, and consist of only a pair.

4.2. Interface

SLAT's interface has been designed to allow for intuitive, visual annotation. It has two main panes in the center of the screen, as shown in Figure 2. The left pane, an editor pane, displays the text to be annotated while the right pane displays a list of all current segments and links. Annotating a segment is as easy as marking a region of text with the mouse.
The upper pane shows information of *selected* and *focused* segments. In Figure 2, "support systems" is *selected* and "adopt" is *focused*. The notion of the selected and forcued segments roughly corresponds to the source and destination segments of a link. A new link is annotated by regarding

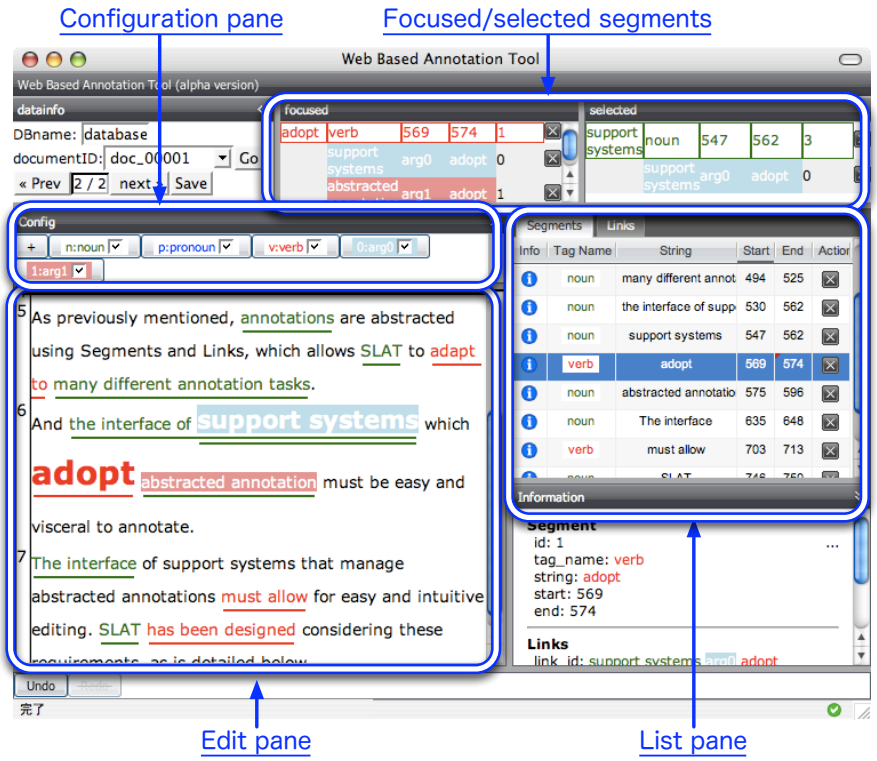

Figure 2: Snapshot of SLAT

selected segment as the destination and focused segment as the source of that link. And these segments have a difference in an operational respect. That is, the system allows users to move around focused segments by using arrow keys, while selected segments are determined by clicking the segments with the mouse. This operational distinction is useful for annotations where multiple links extend from a single segment, such as with predicate-argument annotation. The focusable segments are defined in the configuration as described below.

In the editor pane a segment is displayed as colored and underlined strings. Strings that are comprised of more than one annotation will have multiple underlines. A segment may be selected by clicking on an underlined region. When a segment is selected, links attached to that segment are presented by highlighting the counterpart segments with colors and underlines. In Figure 2, there are two links displayed: one is a link between "adopt" and "support systems" and the other is a link between "adopt" and "abstracted annotation".

The right list pane contains a table-view list of segments and links. Clicking a column header allows for sorting by properties such as *offsets*, *segment/link names* and so on. By clicking on a segment within this list, the left editor pane will scroll to display the selected item. Selecting a link item will identify both the destination and source segment within the editor pane.

4.2.1. Interface Design

Research shows that there are essentially two ways of representing relations: one using edges and the other table-based. In an interface that displays links using edges, identifying a link can become difficult if there is a large number of annotated links. However, a table-based interface has the obvious shortcoming of lacking good visual representation of source/destination. SLAT's interface was designed with both these points in mind. Relations with focused segments are highlighted by underlined and colored strings to avoid congestion in the editor pane. Highlighting can be toggled by a check-box in order to allow annotators concentrate on specific tags during annotation.

Many treebank projects represent the phrase structure of sentences using a tree representation. Phrase structures can be represented in terms of segments and links though the interface today is less than ideal for displaying its hierarchical structure. We designed our interface to be as adaptable to various annotation tasks as possible; segments and links are more versatile than tree representations, and in particular allow for overlapped segments which are troublesome to deal with using trees. That being said, a tree representation might be more suitable when annotating phrase structures and we have plans to incorporate another type of view pane for displaying trees, based on a user's configuration options.

4.3. Customization

SLAT allows users to customize tag-sets in two ways, (1) by using the GUI directly, and (2) by uploading a file containing tag-set definitions. Figure 3 shows a snapshot of the configuration interface, through which the user can create segment and link definitions.

A SLAT configuration can define different types of annota-

tions simultaneously e.g. coreference, predicate-argument structure and syntactic structure and whatsoever. Users can toggle the visibility of each tag by using the configuration pane just above the edit pane.

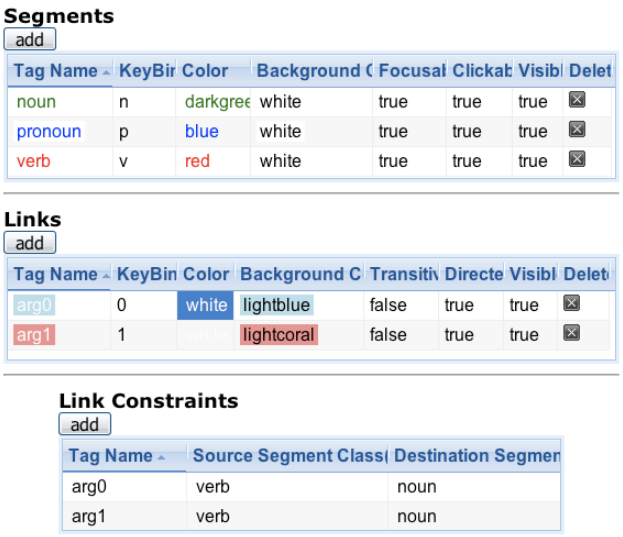

Figure 3: Snapshot of configuration pane

4.3.1. Segments

Tag-name defines the name of the segment, *key-bind* is an optional keyboard shortcut for creating a new segment while annotating a text; *color* and *background-color* define display colors, and *focusable* toggles whether or not a segment can be focused using arrow keys; *clickable* and *visible* each define whether a segment is selectable by clicking and if it is visible, respectfully. Sample definitions are shown in the upper table of Figure 3.

4.3.2. Links

Tag-name defines the name of the link, *key-bind* is the same as explained above, only for links; *transitivity* and *directed* define whether a link has each attribute as defined earlier. Based on these settings, SLAT can constrain the selection and pairing of source/destination tags. For allowing several source/destination combinations, they should all be defined here. Sample definitions are shown in the lower tables of Figure 3.

4.4. Other Features

When a segment is selected, the user's selection can be limited to only the focused/selected segment's tag name. This greatly decreases annotation errors related to accidentally selecting wrong segments. After annotation, a user may easily retrieve annotated text from SLAT via the web browser. SLAT supports undo/redo as well as customization and configuration of tag-sets. SLAT supports any language that can be encoded using UTF-8.

5. Summary and Future Work

With the goal of covering a broad range of annotation tasks, we have proposed a data scheme that is easier to understand and to use. In addition, we have introduced a tool SLAT, which implements many features, including several requirements designated especially important during the early stages of a project. SLAT's use of abstracted annotations, i.e. segments and links resolves many of the challenges presented in this paper, though there are still some issues to be solved.

Supporting annotators in assuring the consistency and quality of a corpus is a remaining challenge. The following is our reseach agenda for achieving this goal.

- Introduction of batch operations for keeping consistency
- Annotation help based on the workflow context
- Retrieval of cases similar to the current annotation target
- Visual methods for reporting errors
- Mining annotation data by multiple annotators to find annotation tips

6. Acknowledgment

This work is partially supported by the Grant-in-Aid for Scientific Research in Priority Areas JAPANESE CORPUS[8].

7. References

Stefanie Dipper, Michael Götze and Manfred Stede. (2004). Simple Annotation Tools for Complex Annotation Tasks: an Evaluation. In Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora. pp.54-62. Lisbon. Portugal.

Coreference Task Definition. (1995). The sixth in a series of Message Understanding Conferences (MUC-6). http://cs.nyu.edu/cs/faculty/grishman/COtask21.book_5.html

Olga Babko-Malaya. (2005). PROPBANK ANNOTATION GUIDELINES. http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf

Takahashi Tetsuro, Inui Kentaro. (2006). A multi-purpose corpus annotation tool: Tagrin. Proceedings of the 12th Annual Conference on Natural Language Processing. pp.228-231. Yokohama. Japan.

Christoph Müller. (2006). Representing and Accessing Multi-Level Annotations in MMAX2. In Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-dimensional Markup in Natural Language Processing. pp.73-76. Trento. Italy.

Constantin Orăsan. (2003). PALinkA: A highly customisable tool for discourse annotation. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue. Sapporo. Japan.

Schmidt Thomas. (2001). The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse. In Proceedings of the IRCS Workshop On Linguistic Databases, 11-13. Philadelphia. USA.

[8] http://www.tokuteicorpus.jp