Proceedings 10th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation

May 26, 2014

Reykjavik, Iceland

Harry Bunt, editor

isa-10:
10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation

Workshop Programme

08.30 – 08:50 Registration
08:50 -- 09:00 Opening by Workshop Chair

09:00 -- 10:30 Session A
09:00 -- 09:30 Hans-Ulrich Krieger, *A Detailed Comparison of Seven Approaches for the Annotation of Time-Dependent Factual Knowledge in RDF and OWL*

09:30 -- 10:00 Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, Piek Vossen, German Rigau and Willem-Robert van Hage, *NAF and GAF: Linking Linguistic Annotations*

10:00 -- 10:15 Johan Bos, *Semantic Annotation Issues in Parallel Meaning Banking*

10:15 --10:30 Assaf Toledo, Stavroula Alexandropoulou, Sophie Chesney, Robert Grimm, Pepijn Kokke, Benno Kruit, Kyriaki Neophytou, Antony Nguyen and Yoad Winter, *A Proof-Based Annotation Platform of Textual Entailment*

10:30 – 11:00 Coffee break

11:00 -- 13:00 Session B
11:00 -- 11:15 Bolette Pedersen, Sanni Nimb, Sussi Olsen, Anders Soegaard and Nnicola Soerensen, *Semantic Annotation of the Danish CLARIN Reference Corpus*
11:15 -- 11:45 Kiyong Lee, *Semantic Annotation of Anaphoric Links in Language*
11:45 -- 12:00 Laurette Pretorius and Sonja Bosch, *Towards extending the ISOcat Data Category Registry with Zulu Morphosyntax*
12:00 -- 13:00 Harry Bunt, Kiyong Lee, Martha Palmer, Rashmi Prasad, James Pustejovsky and Annie Zaenen, *ISO Projects on the development of international standards for the annotation of various types of semantic information*

13:00 – 14:00 Lunch break

14:00 -- 16:00 Session C
14:00 -- 14:30 Volha Petukhova, *Understanding Questions and Finding Answers: Semantic Relation Annotation to Compute the Expected Answer Type*
14:30 -- 14:45 Susan Windisch Brown, *From Visual Prototypes of Action to Metaphors: Extending the IMAGACT Ontology of Action to Secondary Meanings*

14:45 -- 15:15 Ekaterina Lapshinova-Koltunski and Kerstin Anna Kunz, *Annotating Cohesion for Multillingual Analysis*
15:15 -- 16:00 Poster session: elevator pitches followed by poster visits
Leon Derczynski and Kalina Bontcheva: *Spatio-Temporal Grounding of Claims*

Made on the Web in Pheme

Mathieu Roche: *How to Exploit Paralinguistic Features to Identify Acronyms in Texts*

Sungho Shin, Hanmin Jung, Inga Hannemann and Mun Yong Yi: *Lessons Learned from Manual Evaluation of NER Results by Domain Experts*

Milan Tofiloski, Fred Popowich and Evan Zhang: *Annotating Discourse Zones in Medical Encounters*

Yu Jie Seah and Francis Bond: *Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanes*e

16:00 --16:30 Coffee break

16:30 -- 18:00 Session D

16:30 -- 17:00 Elisabetta Jezek, Laure Vieu, Fabio Massimo Zanzotto, Guido Vetere, Alessandro Oltramari, Aldo Gangemi and Rossella Vanvara, *Extending `Senso Comune' with Semantic Role Sets*

17:00 -- 17:30 Paulo Quaresma, Amália Mendes, Iris Hendrickx and Teresa Gonçalves *Automatic Tagging of Modality: Identifying Triggers and Modal Values*

17:30 -- 18:00 Rui Correia, Nuno Mamede, Jorge Baptista and Maxine Eskenazi, *Using the Crowd to Annotate Metadiscursive Acts*

18:00 Workshop Closing

Editor

Harry Bunt	Tilburg University

Workshop Organizers/Organizing Committee

Harry Bunt	Tilburg University
Nancy Ide	Vassar College, Poughkeepsie, NY
Kiyong Lee	Korea University, Seoul
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin

Workshop Programme Committee

Jan Alexandersson	DFKI, Saarbrücken
Paul Buitelaar	National University of Ireland, Galway
Harry Bunt	Tilburg University
Thierry Declerck	DFKI, Saarbrücken
Liesbeth Degand	Université Catholique de Louvain
Alex Chengyu Fang	City University Hong Kong
Anette Frank	Universität Heidelberg
Robert Gaizauskas	University of Sheffield
Koiti Hasida	Tokyo University
Nancy Ide	Vassar College
Elisabetta Jezek	Università degli Studi di Pavia
Michael Kipp	University of Applied Sciences, Augsburg
Inderjeet Mani	Yahoo, Sunnyvale
Martha Palmer	University of Colorado, Boulder
Volha Petukhova	Universität des Saarlandes, Saarbrücken
Andrei Popescu-Belis	Idiap, Martigny, Switzerland
Rarhmi Prasad	University of Wisconsin, Milwaukee
James Pustejovsky	Brandeis University
Laurent Romary	INRIA/Humboldt Universität Berlin
Ted Sanders	Universiteit Utrecht
Thorsten Trippel	University of Bielefeld
Zdenka Uresova	Charles University, Prague
Piek Vossen	Vrije Universiteit Amsterdam
Annie Zaenen	Stanford University

Table of contents

Author Index

A Detailed Comparison of Seven Approaches for the Annotation of Time-Dependent Factual Knowledge in RDF and OWL

Hans-Ulrich Krieger

German Research Center for AI (DFKI GmbH)
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
krieger@dfki.de

Abstract

Representing time-dependent factual knowledge in RDF and OWL has become increasingly important in recent times. Extending OWL relation instances or RDF triples with further temporal arguments is usually realized through new individuals that hide the range arguments of the extended relation. As a result, reasoning and querying with such representations is extremely complex, expensive, and error-prone. In this paper, we discuss several well-known approaches to this problem and present their pros and cons. Three of them are compared in more detail, both on a theoretical and on a practical level. We also present schemata for translating triple-based encodings into general tuples, and vice versa. Concerning query time, our preliminary measurements have shown that a general tuple-based approach can easily outperform triple-based encodings by several orders of magnitude.

Keywords: temporal annotation; synchronic & diachronic relations; binary vs. N-ary representation schemata for factual statements.

1. Introduction

Representing temporally-changing information becomes increasingly important for reasoning and query services defined on top of RDF and OWL, for practical applications such as business intelligence in particular, and for the Semantic Web/Web 2.0 in general. Extending *binary* OWL ABox relation instances or RDF triples with further temporal arguments translates into a massive proliferation of useless "container" objects. Reasoning and querying with such representations is extremely complex, expensive, and error-prone.

In this paper, we critically discuss several well-known approaches to the encoding of time-dependent information in RDF and OWL. We present seven approaches and explain their pros and cons. Three of them are then compared in more detail, both theoretically and practically w.r.t. space consumption and answer time for simple queries. Two of the three approaches stay within the existing RDF paradigm, whereas the third proposal argues for replacing the RDF triple by a more general tuple in order to ease reasoning and querying, but also to come up with ontologies that have a smaller memory footprint when compared to semantically equivalent triple-based encodings.

In order to make the measurements for the three approaches comparable, we have used the rule-based semantic repository *HFC* (Krieger, 2013) that we have developed over the last years and which is comparable to popular engines, such as Jena, OWLIM, or Virtuoso. We also present schemata for translating temporal triple-based encodings into general tuples, and vice versa. Concerning query time, our preliminary measurements have shown that a general tuple-based approach can easily outperform a triple-based encoding by 1 to 5 orders of magnitude.

2. Synchronic and Diachronic Relations

Linguistics and philosophy make a distinction between synchronic and diachronic relations in order to characterize statements whose truth value do (or do not) change over time. *Synchronic* relations, such as dateOfBirth, are relations whose instances do not change over time, thus there is no direct need to attach a temporal extent to them. Consider, e.g., the natural language sentence

> *Tony Blair was born on May 6, 1953.*

Assuming a RDF-based N-triple representation (Grant and Beckett, 2004), an information extraction (IE) system might yield the following set of triples:

```
tb rdf:type Person
tb hasName "Tony Blair"
tb dateOfBirth "1953-05-06"^^xsd:date
```

Since there is only *one unique* date of birth, this works perfectly well and properly capture the intended meaning.

Diachronic relationships, however, vary with time, i.e., their truth value do change over time. Representation frameworks such as OWL that are geared towards unary and binary relations can not directly be extended by a further (temporal) argument. Consider the following sentence:

> *Christopher Gent was Vodafone's chairman until July 2003. Later, Chris became the chairman of GlaxoSmithKline with effect from 1st January 2005.*

Given this, an IE system might discover the following time-dependent facts:

```
[????-??-??,2003-07-??]: cg isChairman vf
[2005-01-01,????-??-??]: cg isChairman gsk
```

Applying the synchronic temporal representation schema from above gives us

```
cg isChairman vf
cg hasTime [????-??-??,2003-07-??]
cg isChairman gsk
cg hasTime [2005-01-01,????-??-??]
```

However, the resulting RDF graph mixes up the association between the original statements and their temporal extent

```
 [????-??-??,2003-07-??]: cg isChairman vf
*[2005-01-01,????-??-??]: cg isChairman vf
*[????-??-??,2003-07-??]: cg isChairman gsk
 [2005-01-01,????-??-??]: cg isChairman gsk
```

as the second and third association is not supported by the above natural language quotation.

3. Approaches to Diachronic Representation

Several well-known techniques of extending *binary* relations with additional arguments have been proposed in the literature.

3.1. Equip Relation With Temporal Arguments

This approach has been pursued in temporal databases (called *valid time*) and the logic programming community. For instance, a binary relation, such as worksFor between a person p of type Person and a company c of type Company becomes a quaternary relation with two further temporal arguments s and e, expressing the temporal interval $[s, e]$ in which the atemporal statement $\mathsf{worksFor}(p, c)$ is true (instants are represented by stating that $s = e$):

$$\mathsf{worksFor}(p, c) \longmapsto \mathsf{worksFor}(p, c, \underline{s, e})$$

Unfortunately, OWL and description logic (DL) in general only support unary (classes) and binary (properties) relations in order to guarantee decidability of the usual inference problems. Thus forward chaining engines (such as OWLIM and Jena) as well as tableaux-based reasoners (e.g., Racer or Pellet) are unable to handle such descriptions.

We note here that this approach is clearly the *silver bullet* of representing binary factual statements, since it is the easiest and most natural one, although a direct interpretation is incompatible with RDF and almost all currently available reasoners. We will favor this kind of representation in the second part of the paper when presenting the measurements, using *HFC* (Krieger, 2013).

3.2. Apply a Meta-Logical Predicate

McCarthy & Hayes' situation calculus, James Allen's interval logic, and the knowledge representation formalism KIF use variants of the meta-logical predicate holds. Hence, our $\mathsf{worksFor}(p, c)$ relation instance becomes $\mathsf{holds}(\mathsf{worksFor}(p, c), t)$. McCarthy & Hayes call a statement whose truth value changes over time a *fluent* (McCarthy and Hayes, 1969). The extended quaternary relation from the previous subsection can be seen as a *relational* fluent, whereas the holds expression here, however, embodies a *functional* fluent, meaning that $\mathsf{worksFor}(p, c)$ is assumed to yield a situation-dependent value.

Such kinds of relations are *not* possible in OWL, since description logics limit themselves to subsets of *function-free* first-order logic and because only a weak form of relation composition is possible in OWL. However, we can reify the atemporal fact $\mathsf{worksFor}(p, c)$ in RDF, so that the above holds relation instance can at least be *encoded* by introducing a new individual o, represented as an RDF blank node. We note that in the original calculus, situations were defined at an *instant* of time, thus we use only a single temporal argument t here.

$$\underline{\mathsf{holds}}(\mathsf{worksFor}(p, c), \underline{t}) \longmapsto \exists o \,.\, \mathsf{holds}(o, t) \wedge \\ \mathsf{type}(o, \mathsf{AtemporalFact}) \wedge \mathsf{subject}(o, p) \wedge \\ \mathsf{predicate}(o, \mathsf{worksFor}) \wedge \mathsf{object}(o, c)$$

As an alternative, we might turn the worksFor relation into a class:

$$\underline{\mathsf{holds}}(\mathsf{worksFor}(p, c), \underline{t}) \longmapsto \exists o \,.\, \mathsf{holds}(o, t) \wedge \\ \mathsf{type}(o, \mathsf{WorksFor}) \wedge \mathsf{subject}(o, p) \wedge \mathsf{object}(o, c)$$

However, this would require to always introduce a new class for the representation of each diachronic relation.

3.3. Reify the Original Relation

Reifying a relation instance again leads to the introduction of a new object and five additional new relationships. In addition, a new class needs to be introduced for *each* reified relation, plus accessors to the original arguments, very similar to the approach directly above. Furthermore, and very important, relation reification loses the original relation name, thus requiring a massive modification of the original ontology.

Coming back to our worksFor example, we obtain (WorksFor is the newly introduced class)

$$\underline{\mathsf{worksFor}(p, c, s, e)} \longmapsto \exists o \,.\, \mathsf{type}(o, \mathsf{WorksFor}) \wedge \\ \mathsf{person}(o, p) \wedge \mathsf{company}(o, c) \wedge \\ \mathsf{starts}(o, s) \wedge \mathsf{ends}(o, e)$$

It is worth noting that this encoding can be seen as a kind of "owlfication" of *Neo-Davidsonian* semantics (Parsons, 1990), as the original relation is turned into an *event*.

3.4. YAGO's Fact Identifier

The approach YAGO (Hoffart et al., 2011) takes is related to Approach 2 and 3 directly above, as it is a kind of *external* reification. YAGO uses its own extension of the N3 plain triple format, called N4, which associate unique identifiers i with each time-dependent fact.

The above quaternary relation instance then is represented as follows:

$$\underline{\mathsf{worksFor}(p, c, s, e)} \longmapsto \exists i \,.\, i : \mathsf{worksFor}(p, c) \wedge \\ \mathsf{occursSince}(i, s) \wedge \mathsf{occursUntil}(i, e)$$

Note that the association $i : \mathsf{worksFor}(p, c)$ has the disadvantage of *not* being part of the triple repository (as it is a quadruple technically; we guess that there exists a separate extendable mapping table). Thus, entailment rules and queries will never have access to these quadruples, unless some custom functionality has been implemented in the semantic repository. Nevertheless, this is a valid and proper *annotation* schema, however *not* expressible in OWL.

Rather, such a kind of association can be seen as an extension of the idea behind *annotation properties* in OWL in that not only classes, properties, and individuals can be annotated with information, but also binary relation instances (= triples), thus occursSince and occursUntil from above can be regarded as relation instance annotation properties. Unfortunately, we are not aware of such an extension.

3.5. Wrap Range Arguments

Wrapping the range arguments of a relation instance, i.e., grouping them in a new object, allows us to keep the original relation name, although the approach still requires to rewrite the original ontology:

$$\begin{array}{l}\mathsf{worksFor}(p, \underline{c, s, e}) \;\longmapsto\; \exists o\,.\,\mathsf{worksFor}(p, o) \wedge \\ \quad \mathsf{type}(o, \mathsf{CompanyTime}) \wedge \mathsf{company}(o, c) \wedge \\ \quad \mathsf{starts}(o, s) \wedge \mathsf{ends}(o, e)\end{array}$$

Again, a new object (o), a new class (CompanyTime), and new accessors (company, starts, ends) need to be introduced. W3C suggests this obvious pattern to be used to encode arbitrary *N-ary relations* (Hayes and Welty, 2006). Alternatively, instead of defining a new class for each range type of the original relation, one might define a general class, say RangePlusTime, together with three accessors value, starts, and ends, in order to avoid a *reduplication* of the original class hierarchy on the property level. We use the latter refinement in our measurements below.

3.6. Encode the 4D View in OWL

(Welty and Fikes, 2006) have presented an implementation of the 4D or perdurantist view in OWL, using so-called time slices (Sider, 2001). Relations from the original ontology no longer connect the original entities, but instead connect time slices that belong to those entities. A time slice here is merely a container for storing the time dimension of space-time. At least, the original relation name is kept, although such a representation requires a lot of rewriting and even introduces *two* new container objects:

$$\begin{array}{l}\underline{\mathsf{worksFor}(p, c, s, e)} \;\longmapsto\; \exists t, t'\,.\,\mathsf{worksFor}(t, t') \wedge \\ \quad \mathsf{type}(t, \mathsf{TimeSlice}) \wedge \mathsf{hasTimeSlice}(p, t) \\ \quad \mathsf{type}(t', \mathsf{TimeSlice}) \wedge \mathsf{hasTimeSlice}(c, t') \\ \quad \mathsf{starts}(t, s) \wedge \mathsf{ends}(t, e) \wedge \\ \quad \mathsf{starts}(t', s) \wedge \mathsf{ends}(t', e)\end{array}$$

We note here that *this* approach and the approach *below* only work for binary relations. This restriction, however, do no harm to RDF-encoded OWL ontologies, since an RDF triple encodes a binary relation.

3.7. Interpret Original Entities as Time Slices

In (Krieger et al., 2008), we have slightly extended and at the same time simplified the perdurantist/4D view from directly above. p and c from the example above are still first-class citizens, now called *perdurants* which possess time slices, *explaining* the behavior of an entity within a certain temporal extent (e.g., being a Person or a Company) and are able to group multiple facts that stay *constant* within the same period of time. In the extended relation instance, p and c are then replaced by new IDs p' and c' (similar to the approach above), but these new individuals are still typed to the original classes, here: Person and Company, resp. Keeping the original typing thus allows us to superimpose the original class hierarchy with the notion of a time slice.

$$\begin{array}{l}\underline{\mathsf{worksFor}(p, c, s, e)} \;\longmapsto\; \exists p', c'\,.\,\mathsf{worksFor}(p', c') \wedge \\ \quad \mathsf{type}(p', \mathsf{Person}) \wedge \mathsf{hasTimeSlice}(p, p') \\ \quad \mathsf{type}(c', \mathsf{Company}) \wedge \mathsf{hasTimeSlice}(c, c') \\ \quad \mathsf{starts}(p', s) \wedge \mathsf{ends}(p', e) \wedge \\ \quad \mathsf{starts}(c', s) \wedge \mathsf{ends}(c', e)\end{array}$$

The nice thing with this reinterpretation is that it does not require any rewriting of the TBox and RBox of an ontology and makes it easy to equip arbitrary upper and domain ontologies with a concept of time, supplied by an independent time ontology (e.g., OWL-Time) that only needs to talk about instants and/or intervals; see (Krieger, 2010).

Perdurants p and c above only need to be introduced once, independent of which time slice they are linked to. For example, assuming perdurant p possesses three time slices for $\mathsf{worksFor}(p', c', s, e)$, $\mathsf{worksFor}(p'', c'', s, e)$, and $\mathsf{hasWorkAddress}(p''', a', s, e)$. Since the starting and ending time coincide in the three statements, p', p'', and p''' can be identified, and the temporal extent needs to be specified only once (and not three times).

4. Theoretical Considerations

Within this section, we will consider three of the above seven approaches (Sections 3.1.–3.7.) which we find to be the most promising ones. On a theoretical level, we will count how many bytes, tuple elements, and triples/tuples overall are needed to represent a diachronic relation instance, using approaches **1**, **5**, and **7**.

During the last years, we have gained some experience with all three formats in several German and European projects. In the European project *NIFTi* and *TrendMiner*, we have applied Approach 1 (Krieger and Kruijff, 2011; Krieger and Declerck, 2014). The German *TAKE* project has used Approach 5 to store biographical knowledge. The ontology which backs up the LT-World language portal had been rewritten to adhere to Approach 5, as it lacked an explicit treatment of time. In *MUSING*, we have used Approach 7 to equip the PROTON upper ontology with a notion of time (Leibold et al., 2010). For the *MONNET* project, we have also chosen Approach 7 to represent the Web content of companies, listed on Deutsche Börse's DAX and NYSE's Euronext.

In the following, we will restrict ourself to quaternary relations $p \subseteq D \times R \times T \times T$, where T is used to describe the starting and ending point of a fluent. The reason for this is that approach 7 (and 6) only works for binary relations that are extended by one or two further temporal arguments. Thus a quaternary diachronic relation instance $p(d, r, s, e)$ encodes a truth value for $p(d, r)$ within interval $[s, e]$. We are neutral as to whether temporal intervals are convex (i.e., contain "holes") or whether the temporal metric utilizes $\mathbb{N}$, $\mathbb{Q}$, or $\mathbb{R}$ for T—this is unimportant for the presentation above and the measurements below. We finally note that T can be easily extended by a further disjoint element, say ?, in order to permit left-open or right-open temporal intervals. Given this, comparison operators over time instants or the Allen relations over intervals, however, no longer will be Boolean, but instead become three-valued relations.

4.1. Approach 1: Quintuples

The quaternary relation instance $p(d, r, s, e)$ is represented as a tuple in *HFC* by an extension of the plain N-triple format (Grant and Beckett, 2004):

```
d p r s e
```

This tuple consists of 5 elements/arguments and requires (at least) 20 (= $5 * 4$) bytes, assuming an (internal) `int[]` representation with 4 byte integers (which is the case in *HFC*). Using integer arrays is a common way to represent triples/tuples internally, since the external representation of URIs and XSD atoms needs to be addressed only during input and output. Overall, we obtain 1 object (the integer array) to represent the whole tuple. This last number is very important, since it is desirable to access information directly in a semantic repository, instead of "fiddling" around with helper structures (container objects) that blow up the memory. In addition, the overall number of elements is equally important, since triple repositories usually build up large index structures to efficiently access all those triples that match a specific element at a certain position in a triple.

4.2. Approach 5: W3C's N-ary Relations

As we have indicated in Section 3.5., the triple representation of the quaternary relation instance results in 5 triples/complex objects:

```
d p o
o rdf:type nary:RangePlusTime
o nary:value r
o nary:starts s
o nary:ends e
```

Overall, 5 triples translate into 15 (= $5 * 3$) elements or 60 (= $5 * 12$) bytes. Furthermore, for each `p`, we might need an additional class for the type of `o`, as well as accessors `value`, `starts`, and `ends`. Since these tuples need to be specified only once, we do not count them here. This approach introduces **one** brand-new individual `o` (a blank node) which turns out to be *problematic*, since it might lead to a *non-terminating closure computation* during the application of entailment rules; not covered here, see (Krieger, 2012).

4.3. Approach 7: Time Slices

As described in Section 3.7., perdurants `d` and `r` need only be introduced once, so we do not take them into account. As is the case for approach 5 above, new individuals `d′` and `r′` are introduced here; in fact, **two** for each fluent we like to represent:

```
d′ p r′
d′ rdf:type ...  ;; domain/range of the
r′ rdf:type ...  ;; original relation p
d′ fourd:starts s
d′ fourd:ends e
r′ fourd:starts s
r′ fourd:ends e
d fourd:hasTimeSlice d′
r fourd:hasTimeSlice r′
```

This representation utilizes 9 triples, leading to 27 elements or 108 bytes per fluent in the worst case. We note here that `r′` only needs to be equipped with a temporal extent and linked to perdurant `r` iff p is an OWL *object* property, i.e., *not* mapping to XSD atoms (best case: 5 triples). The below measurements assume the *worst* case.

4.4. Comparison: When to Apply Which Approach

Let us now summarize the pros and cons of the three approaches.

Approach 1. This is—for us—the most intuitive approach: ABox relation instances are simply extended by two further temporal arguments. Existing ontologies (TBox and RBox) can be easily equipped with a treatment of time. RDFS/OWL entailment rules as well as custom rules are more intuitive, easier to formulate, and less error-prone when compared to approach 5 and 7. Approach 1 performs best in terms of memory consumption and querying/reasoning time. Contrary to approach 5 and 7, it does *not* introduces new individuals, a precondition for guaranteeing the termination of the materialization process; see (Krieger, 2012).

Approach 5. This approach, recommended by the best practice group of W3C, is able to encode arbitrary n-ary relations (as is trivially the case for approach 1). The encoding is worth to consider if ontologies are defined from scratch and require time-dependent relations. Contrary to approach 1, approach 5 is compliant with the triple model of RDF. Unfortunately, standard RDFS and OWL reasoning is no longer possible which is also the case for approach 7. This approach introduces a new blank node for each ABox relation instance.

Approach 7. This treatment is great if an ontology is already given, but misses a notion of time. The approach does not require to rewrite the TBox and the RBox of an ontology (contrary to approach 5) and also stays inside RDF. The *time slices are possessed by perdurants* view is attractive, but is the worst of the three approaches in terms of memory consumption. Two further individuals are introduced here.

5. Practical Measurements

In order to compare the three approaches on a practical level, we need a semantic repository that is able to *directly* encode arbitrary n-ary relations (in our special case: quintuples). Popular engines, such as RACER, Pellet, Jena, OWLIM, or Virtuoso which are geared towards binary relations/RDF triples can thus *not* be applied here. As mentioned in Section 1., the experiments were performed using *HFC*, a forward chaining engine and semantic repository that we have developed over the last years and that is used in our lab.

5.1. Initial Numbers

The numbers below are computed against the mid-size ontology that backs up an earlier version of the LT-World

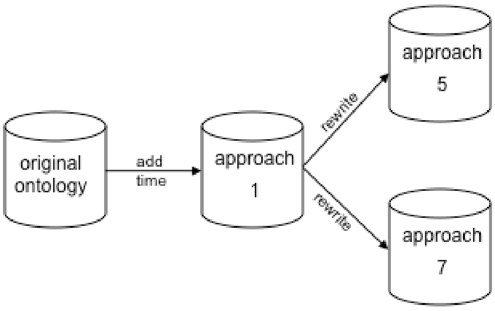


Figure 1: Rewrite schema for obtaining data sets for approaches 1, 5, and 7.

	size [MB]	#tuples	RAM [GB]	time [s]
1	53	548,132	0.42	4.3
5	129	2,740,660	1.67	14.3
7	273	4,360,428	2.15	25.9

Figure 2: Initial numbers for approaches 1, 5, and 7.

language portal (`www.lt-world.org`). The measurements are obtained on a 64bit Intel Core i7 (2.8 GHz), using Java 1.6 with an initial heap of 4GB. The unexpanded ABox consists of 204,959 RDF triples. Fully materialized, 548,132 triples are obtained. Since temporal information is missing, we randomly attach temporal starting and ending points to ABox relation instances through XSD `int` atoms which we let vary between 0 and 1,000 using a random generator (implemented by `java.lang.Math.random()`). This synthetical data (*without* the original triples) is used for approach 1.

We have then produced two further meaning-preserving data sets by rewriting the quintuples to RDF triples, compliant with the formats that are used in approach 5 and 7 (see Figure 1).

For approach 5, we have used blank nodes of type RangePlusTime to group the original value and the starting and ending time of each ABox relation instance. To address approach 7 properly, we have chosen the subject and object URIs of the original triples as names for the perdurants and have attached ascending integers to the original names in order to generate new URIs for the time slices themselves.

Given apporach 1, 5, and 7, Figure 2 then describes the three ontologies in terms of space (file size, number of triples/quintuples, main memory requirement) and loading time in order to set up *HFC* as a repository on which queries are carried out, as described in the next section.

Given these "offline" numbers, approach 1 seems to be far superior. The next section amplifies this judgment through further numbers obtained from "online" measurements for relatively easy queries.

5.2. Querying the Ontologies

This section presents measurements for six SPARQL-like queries posted in *HFC*, given approach 1, 5, and 7. The queries were originally written for approach 1 (see Figure 3) and were transformed manually to the format required by approach 5 (see Figure 4) and 7. **No** translation is depicted here for approach 7 (this would require a further half page).

The first and second query obtains the starting as well as the starting and ending times over all fluents. Query three selects those objects whose fluents are true intervals (filter: start $\neq$ end). The next query searches for subjects in symmetric relation instances that might differ in their starting and ending time. Query five simply accesses all time-stamped information for a specific individual (here: `ltw:obj_68081`). Finally, query six finds those subjects that have an ending time equal to a specific instant (here: `936`).

As can be seen in Figure 4, the queries for approach 5 (as is the case for approach 7) are no longer easy to read and take much longer to complete; in some cases this divergency *can make a difference* between doable and intractable applications which employ such kind of queries.

5.3. Comparison

As can be easily recognized from the measurements depicted in Figure 5, approach 1 easily outperforms approach 5 and 7 by **1 to 5 orders of magnitude**.

We are not only convinced that querying is faster, intuitive and less error-prone for approach 1, but have shown in (Krieger, 2012) that the same happens, even drastically for a more complex case, viz., reasoning over a temporal extension of the RDFS and OWL entailment rules (Hayes, 2004; ter Horst, 2005).

6. Summary

We hope to have shown that a general tuple-based approach for annotating time-dependent factual knowledge on the Web is far superior to triple-based approaches. We are convinced that the time is ripe to move towards this conservative extension of the RDF data model. We note here that even ontologies that utilize approaches 2 to 7 can be easily rewritten to format 1. Due to space requirements, neither are we able to depict and explain any temporal RDFS and OWL entailment rules (Krieger, 2012), nor complex custom rules in the different formats. We are certain that a closer comparison of such rules would even amplify our position, since Semantic Technologies not only are interested in accessing already externalized information (this paper), but also require inferential capabilities to make implicit knowledge explicit.

The attentive reader of this paper might ask him-/herself how we address instantiations of the above schemata in a different *external* representation format, such as XML, and how we handle relations with more than two arguments. We will speculate about this in the next two addenda.

7. Addendum 1: XML Representation

In order to use harvested data from the Web outside the RDF universe and a specific reasoner (in our case: *HFC*), it might be interesting to have an XML exchange representation for the above approaches. Unfortunately, due to the additional degree of freedom in XML to specify a value,

```
(1) SELECT DISTINCT ?start
      WHERE ?subj ?pred ?obj ?start ?end
(2) SELECT DISTINCT ?start ?end
      WHERE ?subj ?pred ?obj ?start ?end
(3) SELECT ?obj
      WHERE ?subj ?pred ?obj ?start ?end
      FILTER ?start != ?end
(4) SELECT DISTINCT ?subj
      WHERE ?subj ?pred ?obj ?start1 ?end1 &
            ?obj ?pred ?subj ?start2 ?end2
(5) SELECT *
      WHERE ltw:obj_68081 ?pred ?obj ?start ?end
(6) SELECT DISTINCT ?subj
      WHERE ?subj ?pred ?obj ?start "936"^^xsd:int
```

Figure 3: Queries for approach 1 (quintuples).

```
(1) SELECT DISTINCT ?start
      WHERE ?blank rdf:type nary:RangePlusTime &
            ?blank nary:starts ?start
(2) SELECT DISTINCT ?start ?end
      WHERE ?blank rdf:type nary:RangePlusTime &
            ?blank nary:starts ?start &
            ?blank nary:ends ?end
(3) SELECT ?obj
      WHERE ?subj ?pred ?blank &
            ?blank rdf:type nary:RangePlusTime &
            ?blank nary:value ?obj &
            ?blank nary:starts ?start &
            ?blank nary:ends ?end
      FILTER ?start != ?end
(4) SELECT DISTINCT ?subj
      WHERE ?subj ?pred ?blank1 &
            ?blank1 rdf:type nary:RangePlusTime &
            ?blank1 nary:value ?obj &
            ?obj ?pred ?blank2 &
            ?blank2 rdf:type nary:RangePlusTime &
            ?blank2 nary:value ?subj
(5) SELECT ?pred ?obj ?start ?end     ;; '*' would also show up ?blank
      WHERE ltw:obj_68081 ?pred ?blank &
            ?blank rdf:type nary:RangePlusTime &
            ?blank nary:value ?obj &
            ?blank nary:starts ?start &
            ?blank nary:ends ?end
(6) SELECT DISTINCT ?subj
      WHERE ?subj ?pred ?blank &
            ?blank rdf:type nary:RangePlusTime &
            ?blank nary:ends "936"^^xsd:int
```

Figure 4: Queries for approach 5 (W3C's N-ary relation encoding).

query [sec]	**1** (1,001)	**2** (293,880)	**3** (544,115)	**4** (1,585)	**5** (37)	**6** (1,398)
1	0.332	0.470	0.440	1.993	0.011	0.037
5	1.975	2.324	5.977	11.066	168.814	329.980
7	3.306	4.076	10.052	——	728.242	284.730

Figure 5: Processing time for the three approaches w.r.t. queries 1–6. The numbers in parentheses at the head of the table list how many results are returned by each query. Query 4 for approach 7 runs out of memory (4GB) after 96 seconds. Queries 5 and 6 are performed 100 times to measure total time.

even more kinds of representations are possible here (examples are related to approach 1 and 3, given our running worksFor example):

```
(1) <worksFor person="p" company="c" ...>
    </worksFor>

(2) <worksFor>p c s e</worksFor>

(3) <RelationInstance pred="worksFor">
      p c s e
    </RelationInstance>

(4) <Event type="worksFor">
      <person>p</person>
      <company>c</company>
      ...
    </Event>

(5) <WorksFor>
      <person>p</person>
      <company>c</company>
      ...
    </WorksFor>
```

We take a liberal stance here as our interest is not in defining an "external" exchange format, but in deciding which "internal" format performs best in terms of (i) *memory consumption*, (ii) *running time* (querying and reasoning), and (iii) *human readability*. Nevertheless, we would probably opt for either the "external" solution (4) or (5) which are related to the "internal" approach (3).

8. Addendum 2: Beyond Binary Relations

The approaches above were investigated on how well they perform w.r.t. *binary* relations whose two arguments can be considered to be *obligatory*. Such kind of relations are the default case in today's popular knowledge resources, such as YAGO, DBpedia, BabelNet, or Google's Knowledge Graph.

In case more and especially *optional* arguments are investigated, our verdict concerning the different approaches will probably turn into a different direction, so the representation format needs to be updated (in the best case) or changed (in the worst case). Consider the following example, taken from (Davidson, 1967, p. 83)

> *Jones buttered the toast in the bathroom with a knife at midnight.*

The binary base relation butter (we assume a direct mapping of the transitive verb to the relation name here) now needs to be split and/or extended by further optional arguments, as the following sentences are perfectly legal:

> *Jones buttered the toast.*
> *Jones buttered the toast in the bathroom.*
> *Jones buttered the toast with a knife.*
> *Jones buttered the toast at midnight.*
> *Jones buttered the toast in the bathroom with a knife.*
> *Jones buttered the toast with a knife in the bathroom.*
> *Jones buttered the toast in the bathroom at midnight.*
> .....

In principle, the number of adjuncts is not bounded, thus adding a large number of potentially underspecified direct relation arguments is probably a bad solution. Today's technologies often address such "hidden" arguments through a kind of *relation composition*, viz., defining further properties such as instrument (to access *knife*) or location (to access *bathroom*) on the object (*toast*) of the relation instance:

$$\mathsf{instrument} \circ \mathsf{butter}$$
$$\mathsf{location} \circ \mathsf{butter}$$

We think that modeling the optional arguments in such a way is *unsatisfactory* as instrument or location "operate" on the object of the binary relation instance and *not* on the relation instance itself!

Our *personal* solution would model the *obligatory* arguments, including (under- or unspecified) time and perhaps space, as *direct* arguments of the corresponding relation instance or tuple. A further argument, an *event* identifier, also takes part in the relation. Optional arguments, however, would be addressed through binary relations, now working on the event argument. Applying this kind of *Davidsonian* or *event* representation to the above example gives us (informal relational notation)

$$\begin{aligned}\exists e\,.\, &\mathsf{butter}(e, \mathit{Jones}, \mathit{toast}, \mathit{at\ midnight}) \land \\ &\mathsf{location}(e, \mathit{bathroom}) \land \\ &\mathsf{instrument}(e, \mathit{knife})\end{aligned}$$

9. Acknowledgements

The research described in this paper has been financed by the European Integrated project *TrendMiner* under contract number FP7 ICT 287863. The author has profited from discussions with Thierry Declerck and Bernd Kiefer—thank you guys! Finally, I would like to thank the three reviewers for suggestions and support.

10. References

Davidson, Donald. (1967). The logical form of action sentences. In Rescher, Nicholas, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

Grant, Jan and Beckett, Dave. (2004). RDF test cases. Technical report, W3C, 10 February.

Hayes, Patrick and Welty, Chris. (2006). Defining N-ary relations on the semantic web. Technical report, W3C.

Hayes, Patrick. (2004). RDF semantics. Technical report, W3C.

Hoffart, Johannes, Suchanek, Fabian M. Berberich, Klaus, Kelham, Edwin Lewis, de Melo, Gerard, and Weikum, Gerhard. (2011). YAGO2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*, pages 229–232.

Krieger, Hans-Ulrich and Declerck, Thierry. (2014). TMO—the federated ontology of the TrendMiner project. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*.

Krieger, Hans-Ulrich and Kruijff, Geert-Jan M.˙ (2011). Combining uncertainty and description logic rule-based reasoning in situation-aware robots. In *Proceedings of the AAAI 2011 Spring Symposium "Logical Formalizations of Commonsense Reasoning"*.

Krieger, Hans-Ulrich, Kiefer, Bernd, and Declerck, Thierry. (2008). A framework for temporal representation and reasoning in business intelligence applications. In *AAAI 2008 Spring Symposium on* AI Meets Business Rules and Process Management, pages 59–70. AAAI.

Krieger, Hans-Ulrich. (2010). A general methodology for equipping ontologies with time. In *Proceedings LREC 2010*.

Krieger, Hans-Ulrich. (2012). A temporal extension of the Hayes/ter Horst entailment rules and an alternative to W3C's n-ary relations. In *Proceedings of the 7th International Conference on Formal Ontology in Information Systems (FOIS 2012)*, pages 323–336.

Krieger, Hans-Ulrich. (2013). An efficient implementation of equivalence relations in OWL via rule and query rewriting. In *Proceedings of the 7th IEEE International Conference on Semantic Computing (ICSC)*, pages 260–263.

Leibold, Christian, Krieger, Hans-Ulrich, and Spies, Marcus. (2010). Ontology-based modelling and reasoning in operational risks. In Kenett, Ron S. and Raanan, Yossi, editors, *Operational Risk Management: A Practical Approach to Intelligent Data Analysis*, chapter 3, pages 41–59. Wiley.

McCarthy, John and Hayes, Patrick J.˙ (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D. editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press.

Parsons, Terence. (1990). *Events in the Semantics of English. A Study in Subatomic Semantics*. MIT Press, Cambridge, MA.

Sider, Theodore. (2001). *Four Dimensionalism. An Ontology of Persistence and Time*. Oxford University Press.

ter Horst, Herman J.˙ (2005). Combining RDF and part of OWL with rules: Semantics, decidability, complexity. In *Proceedings of the International Semantic Web Conference*, pages 668–684.

Welty, Christopher and Fikes, Richard. (2006). A reusable ontology for fluents in OWL. In *Proceedings of 4th FOIS*, pages 226–236.

NAF and GAF: Linking Linguistic Annotations

Antske Fokkens♣,Aitor Soroa♢, Zuhaitz Beloki♢, Niels Ockeloen♣, German Rigau♢, Willem Robert van Hage♠ and Piek Vossen♣

♣Network Institute, VU University Amsterdam, The Netherlands.
♢IXA NLP Group, University of the Basque Country, Donostia, Spain.
♠Innovation Lab SynerScope B.V., TU Eindhoven, The Netherlands
{antske.fokkens, niels.ockeloen, piek.vossen}@vu.nl,
{a.soroa,zuhaitz.beloki,german.rigau}@ehu.es,
willem.van.hage@synerscope.com

Abstract

Interdisciplinary research between computational linguistics and the Semantic Web is increasing. The NLP community makes more and more use of information presented as Linked Data. At the same time, an increasing interest in representing information from text as Linked Data can be observed in the Semantic Web community. It is however not necessarily straightforward to adapt existing NLP modules so that they can read in and produce linguistic annotations in RDF. This paper presents the representations we use in two projects that involve both directions of interaction between NLP and the Semantic Web. In previous work, we have shown how instances represented in RDF can be linked to text and linguistic annotations using GAF. In this paper, we address how we can make further use of Linked Data by using its principles in linguistic annotations.

1. Introduction

Research involving computational linguistics and Linked Data is increasing. The Semantic Web community is looking into Natural Language Processing (NLP) to include information from text to the Semantic Web. At the same time, more and more NLP applications make use of Linked (Open) Data. These research directions call for representations that facilitate interaction between Resource Description Framework (RDF) and linguistic annotations. The idea of using Linked Data in linguistic representations has already been suggested by Ide et al. (2003) for the Linguistic Annotation Framework. Several terminology repositories for NLP have been developed such as the ISO TC37/SC4 Data Category Registry,[1] or the Ontologies for Linguistic Annotation, OLiA (Chiarcos, 2008). It is however not necessarily straightforward to adapt existing linguistic pipelines so that they represent the information they generate in RDF.

In this paper, we describe our approach to facilitate communication between the linguistic annotations produced by our NLP tools and representations in RDF. We describe our framework developed in previous work which allows us to link RDF statements that describe interpretations of text to linguistic annotations. We then go beyond this basic link between linguistic annotations and semantic interpretation and introduce an approach for representing the linguistic annotations themselves in RDF in such a way that does not require complete revisions of our NLP tools or the development of complex conversion wrappers.

Statements about the world presented as Linked Data are linked to linguistic analyses of text using the Grounded Annotation Framework (Fokkens et al., 2013, GAF). As Fokkens et al. (2013) explain, GAF provides a natural way to represent (cross-document) coreference, possibly grounded in the Semantic Web. Together with possibilities of modeling provenance provided by the PROV-O (Moreau et al., 2012), it indicates the source of information making it particularly suitable to model alternative perspectives.

GAF links RDF statements to linguistic annotations represented in any format, as long as they have unique identifiers. Representing linguistic annotations in RDF facilitates this and has the additional advantage that we can define links between linguistic annotations. These links can help us to combine evidence from different modules and hence improve our semantic interpretation.

We describe our ongoing work on making linguistic annotations RDF-based through revisions of the KYOTO Annotation Format (Bosma et al., 2009, KAF), while we continue to use a wide range of NLP modules including 3rd party software. The revised version of KAF, the so-called NLP Annotation Format (NAF), can easily be converted to RDF by assigning Internationalized Unique Identifiers (IRIs)[2] to each annotation and by providing a uniform approach to include provenance information and confidence scores.

The rest of this paper is structured as follows. In Section 2., we provide background information on NewsReader and BiographyNet, the two projects that provided the main requirements for our representation. This is followed by an overview of related work in Section 3. Section 4. provides a brief introduction to GAF. This is followed by an explanation of advantages and challenges in using RDF for linguistic annotations in Section 5. Section 6. describes NAF and is followed by our conclusion in Section 7.

2. Background and Motivation

NAF and GAF were developed as part of two interdisciplinary projects involving NLP and the Semantic Web: NewsReader[3] and BiographyNet.[4] These projects involve both information extraction and the use of Semantic Web technologies for NLP analyses. The requirements set out

[1] http://www.isocat.org/

[2] The use of IRIs rather than URIs is introduced in RDF 1.1. IRIs accept a wider range of unicode characters than URIs.

[3] http://www.newsreader-project.eu

[4] http://www.biographynet.nl

for NAF and GAF are mainly defined by these two projects. They are described in Section 2.1. We then present the main requirements these projects impose in Section 2.2.

2.1. NewsReader and BiographyNet

NewsReader develops technology to process daily news streams in four languages. A range of modules extract *what* happened to *whom*, *when* and *where*, removing duplication, complementing information, registering inconsistencies and keeping track of original sources. Incoming information is integrated with the past, distinguishing new information from old and storylines are unfolded. Output is stored as RDF triples in a central repository called KnowledgeStore (Corcoglioniti et al., 2013) that is also used for reasoning over knowledge.

BiographyNet is centered around the Biography Portal of the Netherlands,[5] a collection of Dutch biographical dictionaries. It is an interdisciplinary project where NLP and Semantic Web technologies are used to support historic research on biographical data. One of the roles of NLP in this project is to interpret text from the biographies automatically and translate it to RDF triples.

These projects have several goals in common that influence the requirements of our representation. First, we create RDF representations of information expressed in natural language in both projects. Second, both projects combine information coming from several sources which partially cover the same topics. Different sources may confirm information, but they can also contradict each other and provide different perspectives on the same topic. We attempt to reveal such differences in perspective in both NewsReader and BiographyNet. Third, NewsReader and BiographyNet both involve several highly challenging tasks that involve multiple NLP components (event detection, cross-document coreference, opinion mining, etc.). Therefore, these projects make use of existing state of the art tools as much as possible.

2.2. Representation requirements

When representing different perspectives, it is essential for the representation schema to allow us to keep track of the **provenance** of all annotations. Provenance information provides insight into where the data came from, what was done with it, what sources and tools were used in the process of creating annotations for the data and who was responsible for the data, tools and execution of the process.

Knowing the source of annotations is particularly important when dealing with contradictory or conflicting information. Because information may be used for historic research (BiographyNet) or decision makers monitoring the news (NewsReader), users need to have a general indication of the reliability of information. This includes the paper, person or publisher that provided information, but also information on the NLP modules that were involved in extracting the information. Provenance information should thus, whenever possible, be accompanied by **confidence scores**.[6]

The connection between information in data and the original source forms an essential part of indicating the provenance. We establish this link through GAF. Furthermore, we want to use information represented as Linked Data to support disambiguation. Our representation format should thus be **conform to RDF** principles as much as possible.

The tasks we set out to do within the projects involve both new representations and existing ones. We are exploring several new challenging topics including complex relations between events, (changing) perspectives and storylines. For several of these topics, there are no existing standard representations. This means that it should be easy to integrate new representations in our format but also that new layers are built on top of previous annotations, resulting in deeper hierarchical representations.

The format should thus be **simple** and **flexible** to allow for new additions. On the other hand, we have more than one tool available for some of the tasks we are carrying out. We want to investigate if we can improve our results by combining the output of different tools. This means that it must be possible to include alternative analyses on the same object next to each other and we need a method to link similar information through appropriate relations (in case the tools are not based on the same theoretical framework). Finally, we are dealing with a massive amount of data in NewsReader.[7] The format should thus be as compact as possible, and has to allow for parallel execution of the NLP modules. It should be noted, however, that the formalism should first and foremost include all required information and be practical. Structure and content will thus not be compromised for the sake of compactness when including essential information or practicality is at stake.

3. Related Work

During the last two decades, several proposals have been made for representing linguistic annotations in such a way that they can be processed by a variety of NLP tools. Differences in theoretical insights and assumptions make standardization challenging. Recent efforts therefore mainly aim for interoperability among formats (Ide and Suderman, 2012). In this section, we will describe several formats that serve this purpose. We then discuss efforts of representing linguistic information in RDF.

3.1. Linguistic Annotations

The General Architecture for Text Engineering (Cunningham, 2002, GATE) provides an infrastructure for integrating NLP tools. The architecture aims at providing an environment for building robust NLP tools and resources. It supports creating NLP pipelines by providing a basic set of NLP tools that can easily be extended and an environment that makes it relatively easy to integrate new components.

Internally, GATE uses a unified format that is based on TIPSTER format (Grishman, 1997), the Atlas format (Bird et al., 2000) and uses Thompson and McKelvie (1997)'s proposal for stand-off markup. Information is represented in Annotation Graphs (AGs). Annotations form the labels of

[5] `http://www.biografischportaal.nl/en/zoek`

[6] Several of our NLP modules assign confidence scores. Only scores directly assigned by our modules are represented.

[7] LexisNexis estimates that each working day around 1 million news articles are published.

```
<role id="rl109" semRole="A0">
<!--Daimler-->
<externalReferences>
<externalRef reference="bring-11.3#Instrument"
resource="VerbNet"/>
<externalRef reference="steal-10.5#Agent"
resource="VerbNet"/>
<externalRef reference="Removing#Agent"
resource="FrameNet"/>
<externalRef reference="Removing#Cause"
resource="FrameNet"/>
</externalReferences>
<span><target head="yes" id="t312"/></span>
</role>
```

Figure 1: Partial Semantic Role Analysis

the edges in the graph that go from one node to another. These nodes have pointers to locations in the annotated text. Annotations furthermore consist of an identifier, a type and additional feature-value pairs. Because nodes can only point to locations in the text and not to other annotations, the annotation does not form a true graph. It is difficult to represent hierarchical annotations (Ide and Suderman, 2012) making it less suitable for our purposes.

The Unstructured Information Management Architecture (Ferrucci and Lally, 2004, UIMA) provides data representations and interfaces that are platform independent. Its main purpose is to provide interoperability. Information is represented in the Common Analysis Structure (CAS). In CAS, annotations are defined as typed objects. For each type, one supertype and a set of features associated with the type are defined. Types have a *is-a* relation with their supertype and inherit the supertype's features. Annotations are associated with a "subject of analysis" (sofa), which corresponds to the annotated data. In the case of NLP, this is usually the text. Annotations are identified by their start and end position in the annotated data.

Compared to NAF, UIMA seems less flexible. For instance, when running a pipeline that uses multiple modules for semantic representations, we postpone the decision on what is likely to be the best interpretation until we have collected as much evidence as possible. This includes the relations between alternative analyses. It is not straightforward to model these relations, which might be fuzzy, in a type hierarchy where relations between types and their supertypes are *is-a* relations and no multiple inheritance is allowed.

For instance, Figure 1 illustrates an analysis of the semantic roles of *Daimler* in the sentence *Daimler takes 40%*. There is overlap between the role "steal-10.5#Agent" and combination of "Removing#Agent" and "Removing#Cause", but none of these roles is a more general or more specific type than the others. The role "bring-11.3#Instrument" contradicts the other outputs. The three similar roles are, in this case, closer to the correct interpretation than the contradictory role. Formally defined relations between these roles would reveal that the semantic role analyses provide more evidence for the interpretation where Daimler ends up with the 40% than the one where Daimler is the instrument for bringing it. However, the relations between these analyses cannot be expressed by subtyping.[8]

Bosma et al. (2009) followed the principles defined as part of LAF. The basic idea is that linguistic annotations are stand-off annotations represented in XML. The representation is layered: different linguistic entities have their own layer. Annotations can assign properties to these entities, including links between entities in a different layer. Information can be added incrementally by introducing new layers. KAF provided hierarchical annotations (not provided by GATE) and the flexibility to provide different and possibly conflicting annotations (not provided by UIMA).

KAF was used successfully to glue NLP tools together in KYOTO[9] and subsequent projects such as OpeNER.[10] It still has some limitations that needed to be addressed. First and foremost, it is not RDF compatible, nor designed in a way that it is easy to convert to RDF. In some layers information is lumped together in a way that makes it difficult to add provenance and confidence scores to individual annotations. Finally, information is sometimes repeated several times in the same representation leading to unnecessary increase in space. NAF, the sequal of KAF, was designed to address these limitations.

The Graph Annotation Format (Ide and Suderman, 2007, GrAF) is a serilizaion of LAF that can represent merged annotations in a single graph. Its interoperability is demonstrated by Ide and Suderman (2012) who show how GrAF representations can be converted to GATE and UIMA and vice versa. The fact that this is possible with other LAF-based formats indicates that it is also likely to be feasible to integrate GATE and UIMA representations in NAF.

3.2. RDF in Linguistic Annotations

The idea of using Linked Data and RDF to represent linguistic annotations for achieving interoperability among linguistic resources has been discussed for several years (Chiarcos et al., 2012). Following Linked Data and RDF principles provide a way to address the so-called conceptual interoperability among resources, i.e. the ability of heterogeneous NLP resources and tools to talk and understand each other.

Ide et al. (2003) explicitly mention RDF as a possible format to provide semantic coherence in representations. Furthermore, linking annotation categories to URIs belonging to a shared terminology is a fundamental part of LAF. ISOcat is completely compatible with RDF (Kemps-Snijders et al., 2008). The NLP2RDF initiative collects a number of efforts for representing NLP related information in RDF, including notable efforts such as OLiA (Ontologies for Linguistic Annotation (Chiarcos, 2008)).

Still, to our knowledge, there are relatively few implementations of RDF-compatible annotation formats that are actively used or produced by NLP modules. Notable exceptions are the NLP Interchange Format (Hellmann et al., 2013, NIF), which is tightly linked to OLiA, UIMA

[8] This example deals with the representation of a single mention in the text, however, other mentions expressing the same statement may add further evidence and/or futher contradictions. This becomes apparent when representing the instances in a semantic layer that collects evidence form all mentions. It is at this level, where we will ultimately have to resolve conflicting information from mentions. The mentions in the text layer often remain undecisive about these interpretations.

[9] http://www.kyoto-project.eu

[10] http://www.opener-project.org

Clerezza,[11] and Cassidy (2010)'s conversion of GrAF to RDF.

Hellmann et al. (2013) provide an elaborate description of NIF and a user evaluation. NIF uses RDF to represent linguistic annotations. Annotations are related to strings which are defined by their start and end offsets in the text. These representations are **simple** and **compact** and it is easy to represent information from different tools. It is straightforward to include information on provenance using PROV-O (Moreau et al., 2012) and confidence. Many of these advantages are the result of NIF's RDF compatibility. We will elaborate on the advantages of using RDF in linguistic representations in Section 5.

NIF has the disadvantage that it is not easy to integrate its representations in NLP tools, as shown by Hellmann et al. (2013)'s user evaluation. Because linguistic annotations are linked to strings it is furthermore not practical for representing hierarchical structures. NIF Stanbol[12] addresses this problem by assigning an identifier to annotations, but this variation of NIF is still in its initial stages of development and is not ready to be used in a complex NLP architecture.

UIMA Clerezza provides a basic mapping mechanism to convert CAS to RDF. We are not aware of a publication that provides in-depth information on this mapping or on how these representations are used. It is therefore not clear whether representations in CAS can easily be represented in RDF or whether such representations are practical to use. It seems, nevertheless, that UIMA together with UIMA Clerezza offers a functionality similar to NAF. Apart from the restrictions of CAS outlined above, it would however been significantly more time consuming to adapt our current NLP modules to UIMA than revising KAF. Furthermore as we pointed out, we still need to deal with conflicting annotations.

Cassidy (2010) describes the process of converting GrAF to a representation in RDF.[13] His motivation is similar to ours. He addresses the advantage of using URIs for linguistic annotations which are defined in ontologies. The implementation is a direct mapping from GrAF's XML representation to XML. Cassidy (2010) shows that GrAF can be converted to RDF, but also points out that a data model that defines information captured by the GrAF's XML schema in a format-neutral way would be preferable, but this had not been developed at the time. To our knowledge, this has not changed. Cassidy (2010)'s work is similar to the work presented in our paper, because he also converts a LAF-based format to RDF. However, he does not address what the resulting data model should look like and how this relates to the original GrAF representation.

NAF is a revision of KAF that addresses several of KAF's limitations by improving its compatibility with RDF. This step is in line with the vision of LAF presented by Ide et al. (2003), who already suggest RDF as a XML compatible format that can be used for semantic coherence. We adapt properties from NIF where possible to stimulate interoperability between tools that work with NIF representations. However, we avoid the challenges related to integrating NIF in our own tools or building NIF wrappers, since our representation maintains a large part of the XML schema that was used in KAF. We thus continue to use a LAF-based format, but have structured it in a way that it can be converted to RDF by simple generic rules resulting in a data model that is particularly suitable for representing provenance and confidence scores.

The following section describes GAF, a framework that can link annotations in any of the formats described above to instances in RDF.

4. Linking Linguistic Annotations to the Semantic Web

In this section, we provide a brief introduction to the Grounded Annotation Framework (GAF). A more elaborate description and motivation can be found in Fokkens et al. (2013).

As mentioned in Section 2.1., we aim to extract what happened to whom, when and where. The information we seek is thus centered around events. We use the Simple Event Model (Hage et al., 2011, SEM) to represent this information at the instance level as opposed to the mention level in text. There are several RDF schemas and OWL ontologies for representing events, but SEM is among the most flexible. In particular, it can contain contradictory information as required by our goal to model different perspectives.

Events are formally represented as *instances* in a semantic layer, just like the participants, locations and times related to the events. GAF introduces the `gaf:denotes` and `gaf:denotedBy` relations. This allows us to link the instances represented in SEM to *mentions* of these instances in text. This approach has several advantages over other approaches to model events in NLP.

First, the approach provides a natural way to model coreference. A set of mentions in text that corefer all denote the same instance. This avoids the (arbitrary) selection of one specific mention as the "anchor", "trigger" or "main referent" to which other mentions corefer. This is particularly relevant for modelling cross-document coreference in NewsReader and BiographyNet where many different sources from different times may refer to the same event making it even more challenging to identify which mention should function as the "anchor".

Second, not all information on events comes from text. Videos, pictures, sensors, or data registration containing mobile phone data may also provide information on events. Because GAF can link SEM representations to any kind of mention, it provides a natural way to integrate information from various kinds of sources.

Third, the instance layer can combine information from many different mentions in a unified repersentation, resolving possible conflicts and complementing information that is lacking in individual representations of mentions. As such, it provides the possibility to override interpretations of individual mentions that lack the evidence for the correct interpretation. It therefore enables us to be more robust and

[11] `http://incubator.apache.org/clerezza/clerezza-uima/`

[12] `http://persistence.uni-leipzig.org/nlp2rdf/specification/stanbol.html`

[13] See also `http://web.science.mq.edu.au/~cassidy/wordpress/?p=330#more-330`

underspecified when representing semantic information for mentions in NAF.

Fourth, GAF can link an instance or RDF statement to any mention that has a unique identifier. We can thus link the statement that a specific person is an agent in an event to a semantic role or syntactic relation or combine information from different event models proposed by the NLP community.

In summary, GAF provides a straightforward way to link linguistic annotations to semantic representations in RDF. The only requirement is that these annotations have a unique identifier making it widely applicable. The next section will discuss advantages and challenges to make more extensive use of RDF in linking linguistic annotations.

5. Linguistic Annotations in RDF

RDF is a useful data model for linguistic representations for several reasons. However, RDF representations pose a challenge when these representations are used as input for NLP modules. This section addresses both sides of using RDF for representing linguistic information.

5.1. Advantages of RDF representations

RDF is by nature a graph model, which makes declarative specification of dependency patterns easy, for instance in SPARQL. Triple stores are typically optimized for queries that require multiple joins. That makes evaluation of dependency graph queries, which are typically long branched chains, efficient. This facilitates the communication between representations in RDF and linguistic processing tools.

Another advantage of RDF is that it uses IRIs[14] for identification and IRIs are not limited to the scope of a document, but have a global validity. This makes it easy to represent coreference relations across documents as done in GAF as explained in Section 4.

Furthermore, RDF forms the basis on which RDFS and OWL ontology reasoning is possible. This allows for some very useful operations, such as subclass, subproperty and property chain reasoning. We therefore propose to use IRIs more extensively than is currently done in NIF. NIF represents most linguistic attributes and values as strings. In NAF, we try to use IRIs as much as possible while representing linguistic information.

Schuurman and Windhouwer (2011) note the challenges involved in defining standardized sets of linguistic properties. ISOcat (Kemps-Snijders et al., 2008) provides standards with useful definitions, but because of differences in linguistic theories or cross-linguistic properties it is not always possible to use existing sets. New, sometimes closely related, categories will be introduced as linguistic annotations. If we can represent linguistic properties with ontologies, we can define how output of different tools relate to each other.

If there are differences in granularity between output of certain tools, reasoning can be used to generalize over linguistic information. It is also possible to define equivalence or near equivalence. The possibility of defining relations between linguistic classes increases the interoperability and comparability of tools (Hellmann et al., 2013). For instance, Agirre et al. (2009) define a basic set of nine Part of Speech (PoS) tags which are used in KAF. Several other modules that use PoS tags as their input assume that this set is used. If we include a PoS tagger that is trained on the Penn Treebank, this will assign tags according to the set defined by Santorini (1990). We can define that a common plural noun (NNS) and a common singular or mass noun (NN) from Santorini's set are both subtypes of the nominal class (N) used by Agirre et al. (2009). RELcat (Schuurman and Windhouwer, 2011; Windhouwer, 2012) provides a set of basic relations specically designed for this purpose. We can make use of these relations in NAF.

5.2. The challenge of using RDF

Several challenges exist when it comes to creating linguistic representations in RDF. In fact, Hellmann et al. (2013) state that "RDF can hardly be used efficiently for NLP in the internal structure of a framework". We will define those in two categories: Those caused by generic differences in structure and expressivity between RDF and XML representations, and those caused by practical use of these different interpretations in NLP tools and pipelines, such as the ability to read in and (re)use annotation information from other tools with relatively low cost.

Comparing XML and RDF is a bit like comparing apples and oranges; while XML in itself is a data format and serialization format, RDF is an abstract data *model* which can be serialized using several data formats and syntaxes. While RDF is meant to express semantic relations between objects, "XML is first and foremost a means to define grammars" (Decker et al., 2000).

Often, intrinsic properties of a defined XML grammar are used to express important information, e.g. the nesting of elements is used to denote a hierarchical relation within the data. Furthermore, concepts such as "document order" are intrinsic to XML related technologies including DOM, XPath and XSLT.

Since multiple serialization formats are available for RDF, syntactical grammar properties can not implicitly be used to encode information such as ordering, hierarchy, etc. Hence, the information encoded in such grammar based features needs to be modeled explicitly in the data model. Though this is very well possible and one could argue that this is a more sound solution to start with (Cassidy, 2010), it does not alter the fact that adopting current NLP tools and pipelines to use such a data model is a non trivial task. As mentioned above, NIF Stanbol offers the basic structure for such a model, but it is still in its initial stages of development. Current representations of linguistic annotations in RDF have a radically different structure from the one used in LAF-based models making it challenging to build wrapping tools around NLP modules that use LAF-based representations.

6. The NLP Annotation Format

The previous section showed why RDF can be useful for representing linguistic annotations, but also that there are

[14] Recall that IRIs are the new internationalized variant of URIs used in RDF 1.1.

```
<NAF>
 <!-- text layer -->
 <text>
  <wf id="w1" offset="0" length="4">John</wf>
  <wf id="w2" offset="5" length="6">taught</wf>
  <wf id="w3" offset="12" length="11">mathematics</wf>
  <wf id="w4" offset="24" length="2">in</wf>
  <wf id="w5" offset="27" length="3">New</wf>
  <wf id="w6" offset="31" length="4">York</wf>
 </text>
 <!-- term layer -->
 <terms>
  <term id="t1" lemma="John" pos="R">
   <span><target id="w1"/></span>
  </term>
  <term id="t2" type="open" lemma="teach" pos="V">
   <span><target id="w2"/></span>
  </term>
  ...
 </terms>
 <!-- entity layer -->
 <entities>
  <entity id="e1" type="person">
   <references>
     <!--John-->
    <span><target id="t7"/></span>
   </references>
  </entity>
  <entity id="e2" type="location">
   <!--New York-->
   <references>
    <span><target id="t5"/><target id="t6"/></span>
   </references>
   <externalReferences>
    <externalRef
      reference="http://dbpedia.org/page/New_York_City"
      confidence="0.8"/>
   </externalReferences>
  </entity>
 </entities>
</NAF>
```

Figure 2: Excerpt of a NAF document showing the text, term and entity layers.

some challenges involved in adapting tools to use RDF. We propose a solution where we maintain a LAF-based representation similar to KAF, but revise it so that it can easily be converted to RDF. The structure remains easy integratable in our existing tools, but also allows us to take advantage of possibilities offered by RDF. In this section, we describe the current status of the format.

6.1. Current Status of NAF

Like KAF, NAF comprises several annotations over a text at different linguistic levels (morphosyntactic, syntactic, semantic and pragmatic),[15] adopts a stand off strategy for annotating the source text and is XML based. The following general rules are followed in all layers:

- `<span>` elements are used to define the range of linguistic elements to which an annotation applies.
- Linguistic annotations of a particular level always span elements of previous levels.
- Linguistic annotations of different levels are not mixed.

The "levels" in the general rules refer to different types of linguistic information, which can be groupments of linguistic entities (e.g. tokens vs. terms vs. chunks), relations between linguistic entities (e.g. dependencies, semantic roles) or information about a linguistic entity (e.g. disambiguated word sense). Figure 2 shows an excerpt of a NAF document comprising three layers: text, terms and entities. The span elements for the entities point to identifiers in the term layer, while the span elements in the term layer point to the identifiers of the tokens in the text layer.

In order to reduce unnecessary duplication of information and facilitate conversion to NAF-RDF, the following additional rules were defined for NAF:

- No duplicate representations of fixed properties of a specific linguistic annotation
- Consistent structure of different linguistic layers
- Usage of IRIs whenever possible to refer to external entities and linguistic properties

Consistency is important so that generic rules can be used to convert standard NAF to NAF-RDF. For instance, in NAF we always use the `<naf:span>` element to point to lower linguistic entities and `<naf:from>` and `<naf:to>` attributes to define a relation from one linguistic entity (the source) to another (the target).

IRIs are used as much as possible so that we can make use of the advantages of RDF conform representations, as outlined above. In the example in Figure 2, the entity "New York" is recognized by the NER module and is linked to the appropriate DBpedia page. The association between the entity and the external reference is represented using an IRI (`http://dbpedia.org/page/New_York_City`).

We strive to indicate attributes and their values through IRIs as well. Representing information by IRIs has the advantage that we can define properties of linguistic values formally in RDF. This avoids repetitions of such properties as found in KAF. A requirement that all IRIs should also be represented in ontologies can however form a hindrance to quickly integrate new annotations. Creating ontologies and use their definitions in NAF is therefore optional, though highly recommended.

The principles behind NAF are mostly followed by the NLP modules that currently use NAF. There are however a few additional revisions needed to meet all our requirements. We will outline them in the next subsection.

6.2. Further simplifying RDF conversion

NAF layers can easily be converted to RDF, but it is currently not possible to do so with a generic script that applies to all layers. In NAF-RDF, all annotations are represented as triples. A typical triple would have the identifier of a linguistic object as subject, an attribute as predicate and the attribute's value as its object. Figure 3 provides an example of NAF annotations in RDF. Triples can be placed in named graphs. We can provide provenance information and confidence values for each named graph. Triples will thus be placed in the same named graph according to their provenance and confidence values. Note that this will often mean that a named graph contains only one triple. An XML element in NAF can be translated to RDF by taking the identifier as subjects, attributes as RDF predicates and values as objects. This means that an XML element in NAF should

[15] Currently, NewsReader uses 12 different layers for processing text ranging from low level analysis, such as tokenization, to high-level analysis such as semantic roles and factuality

```
@prefix docId: <http://iri/to/my/document#> .
@prefix naf: <http://wordpress.let.vupr.nl/naf/> .

:Terms { docId:t1 naf:hasSpan docId:w1 .
         docId:t2 naf:hasSpan docId:w2 . }

:T1   { docId:t1 naf:hasLemma "John" ;
                 naf:hasPos naf:R  . }

:T2a { docId:t1 naf:hasLemma "teach" ;
                      naf:hasPos naf:V   . }

:T2b { docId:t2 naf:isTermType naf:open . }

:PosConf { :T1  naf:confidenceScore 0.78 .
           :T2a naf:confidenceScore 0.64 . }

:typeConf { :T2b naf:confidenceScore 0.94 . }

:Prov  { :PosConf prov:wasGeneratedBy docId:Pos1 .
         :typeConf prov:wasGeneratedBy docId:Pos1 .
         docId:pos1 prov:used naf:IXAposTagger . }
```

Figure 3: Simplified term representation (in RDF TriG)

always provide a unique identifier and may only contain attributes that belong in the same named graph, i.e. that have the same provenance and confidence scores. If there are annotations associated with an object that have different provenance or confidence scores, multiple XML elements must be used to represent this information.

As can be seen in the term layer in Figure 2, this is currently not the case. The term element indicates the type, lemma and Part of Speech (PoS). Even though lemma and PoS are often determined by the same tool and have the same provenance, one could imagine that they do not always have the same confidence score. The type indicates whether the word is a member of a closed class and will definitely have different confidence scores from lemma and PoS. According to the requirement outlined above, the type should thus be indicated in its own element and the same may apply to PoS and lemma. Note that it is possible to assign more than one confidence score to the same named graph. In this case, however, all confidence scores apply to all triples in the graph. It is therefore not an option for the scenario outlined above.

In Figure 3, we represent the output of one tool which is a lemmatizer and PoS-tagger and can indicate whether a word is closed or open class. In this case, the tool assigns identical confidence scores to lemmas and PoS-tags and a different score to the type. Even if provenance can be provided for each linguistic object, it is typically provided for a set of annotations that are created by the same activity. Because PoS-tags, lemmas, types and their confidence scores are provided by running the same module, they have the same provenance.

Most annotations in NAF are represented as attributes in elements. There are two notable exceptions: `<span>` and `<externalReferences>`. A `<span>` is a child of a linguistic element and includes one or more `<target>` elements. These targets refer to linguistic elements from other layers. As their name indicates, `<externalReferences>` can link linguistic elements to annotations defined in external resources. This element can contain one or more `<externalRef>` elements that always consists of a reference and a resource. Because the structure of these two exceptions is consistent across layers, they too can be converted to RDF representations by generic rules. However, note that both a reference and a resource are indicated for `<externalRef>`. If we use a IRI to indicate the reference, we no longer need to provide the resource. The resource is an invariable property of the reference and need not be provided for individual elements. Resource attributes can be removed from external references as soon as we start making use of IRIs more extensively. Revisions concerning the attributes that may occur in the same element will be implemented as we start adding more modules to our architecture and experimenting with more than the top-ranking outcome of tools. Provenance and confidence information for individual annotations play a significant role in this step and may lead to further revisions of the structure. It should however be anticipated when new layers of information are added to NAF.

Modeling the provenance of information is essential for GAF. We can evaluate the value of informaton coming from many different layers and across different mentions (within the same document and accros documents) to be unified at the instance level in SEM. The ultimate goal is to come to an adequate representation in SEM which can be based on many different pieces of evidence. This also allows us to model interpretation of text given diffent background knowledge that may complement the partial and vague information in text, which is the rule rather than the exception.

7. Conclusion and Future work

In this paper, we presented ongoing work to link linguistic annotations using RDF and, vice versa, to convert textual information to RDF. We introduced NAF, a LAF-based representation format that is specifically structured in a way so that it can easily be converted to RDF. NAF aims to be a consistent and compact representation schema, easy to convert to RDF and it facilitates the integration of provenance and confidence scores into the model.

Our work on NAF can be seen as a continuation of our previous work on GAF which we also described in this paper. This generic framework can relate any instance to a mention of this instance and any RDF triple to a mention of the relation expressed by the triple. GAF provides a link between the Semantic Web and linguistic annotations and forms at the same time a natural way to model (cross-document) coreference. NAF forms the next step in facilitating the representation of linguistic annotations in RDF. Within GAF, NAF should provide the pieces of linguistically grounded evidence with their provenance. We argued that we need to allow for flexibility, redudancy and even conflicts at the level of linguistics annotation of mentions in NAF, to be resolved at the semantic level in SEM reasoning over the provenance of the evidence.

We outlined the general advantages of using RDF for linguistic representations. They include efficient graph search, straightforward coreference representations, and the possibility of using reasoning to link linguistic representations. This last property is particularly important since it supports interoperability. We also point out some challenges due to differences between RDF and XML representations typically used for representing linguistic annotations.

The solution NAF offers is to maintain properties of an existing linguistic annotation format that has proven its practicality for complex linguistic pipelines and adapt it so that it can easily be converted to RDF. We have outlined a set of principles for NAF that do not harm the flexibility, interoperability or practicality of KAF, but does facilitate conversion of NAF to RDF.

As pointed out in Section 6.2., a few more steps need to be made in NAF to make it fulfill all our requirements. The next steps will therefore mainly focus on replacing information by IRIs. While creating ontologies and IRIs for representing linguistic annotations, we aim to look at alternative representation formats as much as possible in order to improve interoperability of NAF. In particular, we will try to make use of NIF representations ideally by joining the NLP2RDF initiative as suggested by Hellmann (p.c.).

Acknowledgements

We thank Harry Bunt and three anonymous reviewers for valuable feedback. All remaining errors are our own. This work was supported by the European Union's 7th Framework Programme via the NewsReader (ICT-316404) project and the BiographyNet project (Nr. 660.011.308), funded by the Netherlands eScience Center (`http://esciencecenter.nl/`).

8. References

Agirre, E., Artola, X., de Ilarraza, A. D., Rigau, G., Soroa, A., and Bosma, W. (2009). KAF: Kyoto annotation framework.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., and Liberman, M. (2000). ATLAS: A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). KAF: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009*, Pisa, Italy.

Cassidy, S. (2010). An RDF realisation of LAF in the DADA annotation server. In *Proceedings of ISA-5*, Hong Kong.

Chiarcos, C., Nordhoff, S., and Hellmann, S. (2012). *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Corcoglioniti, F., Rospocher, M., Cattoni, R., Magnini, B., and Serafini, L. (2013). Interlinking unstructured and structured knowledge in an integrated framework. In *7th IEEE International Conference on Semantic Computing (ICSC), Irvine, CA, USA*.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

Decker, S., Melnik, S., Harmelen, F. V., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., and Horrocks, I. (2000). The semantic web: The roles of XML and RDF. *Internet Computing, IEEE*, 4(5):63–73.

Ferrucci, D. and Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Fokkens, A., van Erp, M., Vossen, P., Tonelli, S., van Hage, W. R., Serafini, L., Sprugnoli, R., and Hoeksema, J. (2013). GAF: A Grounded Annotation Framework for events. In *Proceedings of the first Workshop on Events: Definition, Dectection, Coreference and Representation*, Atlanta, USA.

Grishman, R. (1997). TIPSTER architecture design document version, 2.3. Technical report.

Hage, W. V., Malaisé, V., Segers, R., Hollink, L., and Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). *J. Web Sem.*, 9(2):128–136. `http://dx.doi.org/10.1016/j.websem.2011.03.003`.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using Linked Data. In *Proceedings of the 12th International Semantic Web Conference*.

Ide, N. and Suderman, K. (2007). GrAF: a graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8, Prague, Czech Republic.

Ide, N. and Suderman, K. (2012). Bridging the gaps: interoperability for language engineering architectures using GrAF. *Language Resources and Evaluation*, (46):75–89.

Ide, N., Romary, L., and Villemonte de La Clergerie, E. (2003). International standard for a linguistic annotation framework. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)*. Association for Computational Linguistics.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. (2008). Isocat: Corralling data categories in the wild. In *LREC*.

Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., and Tilmes, C. (2012). PROV-DM: The PROV data model. Technical report.

Santorini, B. (1990). Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Technical report, Penn Engineering.

Schuurman, I. and Windhouwer, M. (2011). Explicit semantics for enriched documents. what do ISOcat, RELcat and SCHEMAcat have to offer. In *2nd Supporting Digital Humanities conference (SDH 2011), Copenhagen*.

Thompson, H. and McKelvie, D. (1997). Hyperlink semantics for standoff markup read-only documents. In *Proceedings of SGML Europe-97*.

Windhouwer, M. (2012). RELcat: a relation registry for isocat data categories. In *Proceedings of LREC 2012*, pages 3661–3664.

Semantic Annotation Issues in Parallel Meaning Banking

Johan Bos
University of Groningen
johan.bos@rug.nl

Abstract

If we try to align meaning representations of translated sentences, we are faced with the following problem: even though concepts and relations ought to be independent from specific natural languages, the non-logical symbols present meaning representations in usually resemble language-specific words. In faithful translations, such symbols can be easily aligned. In informative translations (where more information is provided by the target translation), symbols can be aligned by a symbol denoting an inclusion relation. In loose translations, we need a third combinator to combine symbols with similar but not identical meanings. We show how this can be done with several concrete, non-trival English-German translation pairs. The resulting formalism is a first step towards constructing parallel meaning banks.

Keywords: Semantic Representations, Meaning Banking, Parallel Corpora

1. Introduction

The ingredients of meaning representations can roughly be divided into two categories: the logical symbols, and the non-logical symbols. To the first category belong the quantifiers, the variables, and the boolean operators (negation, conjunction). The members of the second category, the non-logical symbols, are based on the language that is undergoing semantic analysis. For example, a meaning representation for a simple sentence like “John doesn’t smoke” would contain the logical symbols $\neg$ and $\wedge$, several variables, and the non-logical symbols JOHN (representing the entity referring to John) and SMOKE (representing the event of smoking). But now suppose I have a good translation of this English sentence into, say, German or Dutch. Arguably, the meaning representation for this translation should not differ a great deal. But what would it look like precisely?

One possible solution is to take a (neutral) auxiliary language for defining the vocabulary of non-logical symbols. But soon one will discover that this option isn’t feasible. In natural (non-literal) translations, the source is sometimes more general, sometimes more specific than the target translation. This information will be lost when one relies on a single language. Moreover, phrasal translations will be hard to capture by a single language of symbols.[1] The alternative, and one that will be explored in this paper, is to combine the non-logical symbols of the source and target of a translated sentence into a single meaning representation.

In order to investigate this possibility, we follow a strongly data-driven method. We take non-trivial translation examples from an existing corpus (see Figure 1) and produce the meaning representations for each language. Then we will compare the respective meaning representations, and examine how we could align the two representations. Here we will just consider pairs of English-German translations — the choice for these two close languages makes sense for a pilot study of this kind.

We employ Discourse Representation Theory, DRT (Kamp and Reyle, 1993), as the formal theory of meaning, mainly because it is well-known among semanticists and has covers many linguistic phenomena, but we would like to emphasize that any meaning representation with variables and n-place relations could have been adopted to integrate the ideas put forward in this paper. We will introduce new machinery for representing parallel meanings. We will bring three new operators into play for combining non-logical symbols dealing with faithful translations, informative translations, and loose translations. To make this more readable, we just assume that the non-logical symbols represent the right sense of the concepts expressed by the surface strings. We also assume that each non-logical symbol carries the information of its source language (here: English or German), but don’t explicitly show it in the meaning representations for reasons of clarity.

2. Faithful Translations: $\equiv$

Faithful translations are among the easiest to align, because they are often based on word-by-word translations. Consider the examples and corresponding meaning representations given below in Example 1. Here, and in the examples that follow, we show the meaning representation for an English expression and one its German translation, and a parallel meaning representation comprising both source and target language. The mono-lingual meaning representations also show the mappings of discourse referents to surface strings (where dotted variables indicate substitutions that need to take place) for the reader’s convenience.

EXAMPLE 1

x p	x p
x ↦ “the chance to ṗ”	x ↦ “die Gelegenheit zu ṗ”

CHANCE(x)
TO(x,p)
	GELEGENHEIT(x)
	ZU(x,p)

↘ ↙

x p
CHANCE≡GELEGENHEIT(x) TO≡ZU(x,p)

[1] Although there are initiatives, notably the Abstract Meaning Representation project (Banarescu et al., 2013), pursuing closely related goals.

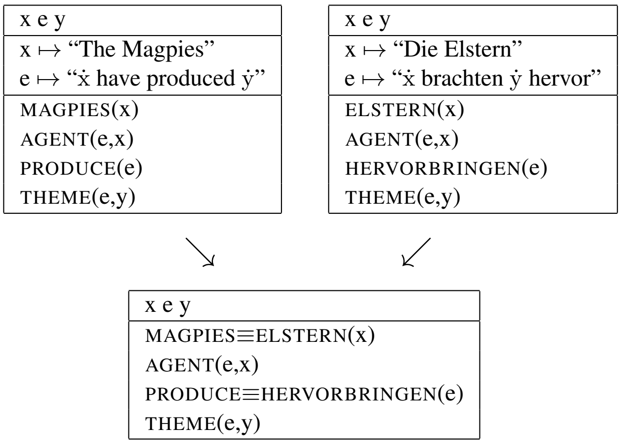
English (en)	**German** (de)
Pubs also provide good value for money, the chance to taste a pint of beer and have a chat with the locals.	Pubs bieten auch ein gutes Preis-Leistungsverhältnis, die Gelegenheit ein Glas Bier zu trinken und mit den Einheimischen zu plaudern.
The "Magpies", Newcastle United Football Club, have produced some of Britain's finest players.	Die "Elstern", wie der Newcastle United Football Club auch genannt wird, brachten einige der besten Fußballspieler Großbritanniens hervor.
Due to the possibility of animals and birds bringing disease to the UK, bringing them with you on holiday is not recommended.	Da Haustiere und Vögel Krankheiten nach Großbritannien einschleppen können, wird davon abgeraten, sie mit in die Ferien zu nehmen.

Figure 1: Examples considered in this study. Source: The English-German Translation Corpus, `http://ell.phil.tu-chemnitz.de/`.

This example illustrates a faithful, literal translation, and as a pleasant consequence there is a simple one-to-one mapping between the non-logical symbols of the source and target language. To arrive at a parallel meaning representation, we combine the non-logical symbols (with the same arity) originating from different languages by simply concatenating them with the help of a new operator: ≡. For instance, the German-originating two-place relation ZU and the English-originating two-place relation TO are combined to yield a new compound non-logical symbol TO≡ZU.

Now consider Example 2, illustrating some basic neo-Davidsonion event structure.[2] It makes sense to assume that the thematic roles are universal and therefore language independent. Therefore it is not necessary to align the conditions for the roles in the parallel meaning representation: they are shared. However, it could be the case that there are languages that explicitly express a role (for instance, by a preposition), in which case the non-logical symbol denoting that role could be based on it.[3]

EXAMPLE 2



x e y
x ↦ "The Magpies"
e ↦ "ẋ have produced ẏ"
MAGPIES(x)
AGENT(e,x)
PRODUCE(e)
THEME(e,y)

x e y
x ↦ "Die Elstern"
e ↦ "ẋ brachten ẏ hervor"
ELSTERN(x)
AGENT(e,x)
HERVORBRINGEN(e)
THEME(e,y)

x e y
MAGPIES≡ELSTERN(x)
AGENT(e,x)
PRODUCE≡HERVORBRINGEN(e)
THEME(e,y)

We will give meaning to this new operator by extending a translation function from the meaning representation to first-order logic, $[.]^{fol}$, on the same lines as earlier work in Discourse Representation Theory (Bos, 2004; Kamp and Reyle, 1993). We can define ≡ as follows:

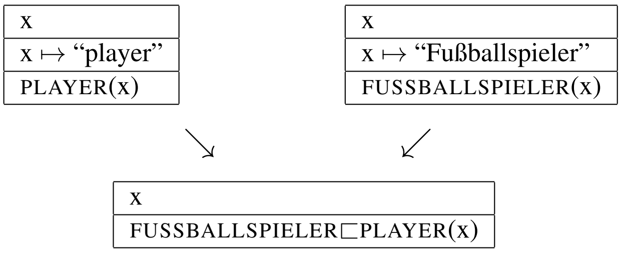
$$[S_i{\equiv}S_j(\mathrm{x}_1,\ldots,\mathrm{x}_n)]^{fol} = S_i(\mathrm{x}_1,\ldots,\mathrm{x}_n) \wedge \forall \mathrm{u}_1,\ldots,\mathrm{u}_n(S_i(\mathrm{u}_1,\ldots,\mathrm{u}_n) \leftrightarrow S_j(\mathrm{u}_1,\ldots,\mathrm{u}_n))$$

This simply says that all these symbols are synonyms, and applied to n of variables, result in logically equivalent meanings. One could compare this to a WordNet (Fellbaum, 1998) synset: given the compound symbol A≡B, then A and B belong to the same cross-lingual synset.

3. Informative Translations: $\sqsubset$

Translations, however, are rarely as literal and faithful as the previous examples suggest. Consider for instance Example 3, where the English noun "players" is translated into German with the more specific "Fußballspieler". Even though it is clear from the context in the English sentence that we talk about players that practice the game of football, it isn't stated explicitly. It would therefore be wrong to align the meanings of these words with the ≡ operator. What we propose to do instead is introducing a new operator, $\sqsubset$, that combines two symbols and specifies that the first is more specific (carries more information) than the second.

EXAMPLE 3



x
x ↦ "player"
PLAYER(x)

x
x ↦ "Fußballspieler"
FUSSBALLSPIELER(x)

x
FUSSBALLSPIELER⊏PLAYER(x)

As can be seen in the parallel meaning representation in Example 3, we specified that FUSSBALLSPIELER is more informative than PLAYER. This seems to be a common phenomenon in translation. What's left to do is giving a formal definition for $\sqsubset$, and we define it in first-order logic as:

$$[S_i{\sqsubset}S_j(\mathrm{x}_1,\ldots,\mathrm{x}_n)]^{fol} = S_i(\mathrm{x}_1,\ldots,\mathrm{x}_n) \wedge \forall \mathrm{u}_1,\ldots,\mathrm{u}_n(S_i(\mathrm{u}_1,\ldots,\mathrm{u}_n) \rightarrow S_j(\mathrm{u}_1,\ldots,\mathrm{u}_n))$$

[2] For simplicity we assume that proper names introduce one-place relations.

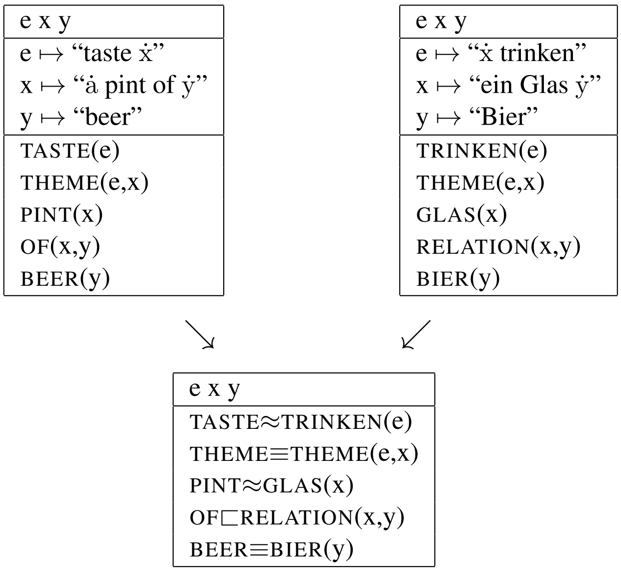
[3] An example that comes to mind is the passive construction in English, where the *agent* role is marked by the preposition "by". A further example is the semantic role of *recipient* expressed by the preposition *to*, in constructions like "Mary gives the book to John". See also Example 6.

For instance, given the compound symbol A$\sqsubset$B applied to x, then A(x) holds, and if A(x) holds then also B(x) holds. In the parlance of WordNet (Fellbaum, 1998) practitioners, A would be a hyponym of B.

4. Loose Translations: $\approx$

An old proverb says that a translation cannot be both faithful and beautiful. Loose translations often just sound better. A case in point is "taste a pint of beer" and its German rendering "ein Glas Bier trinken": a pint (a unit of measurement) isn't the same as a glass (a container), and tasting isn't the same as drinking, although in WordNet (Fellbaum, 1998) they are both co-troponyms of *consume*. To align such loose translations we propose a new operator for symbol alignment: $\approx$, illustrated by Example 4.

EXAMPLE 4



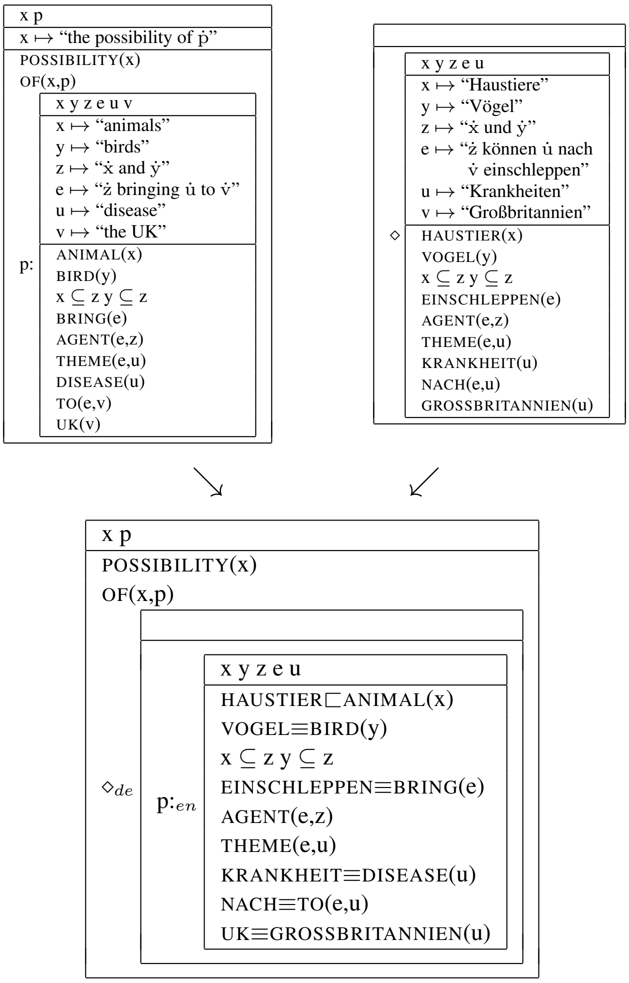
The $\approx$ combiner is used to align non-logical symbols that have approximately the same meaning, and therefore cannot be described by $\equiv$ or $\sqsubset$. It is defined as follows:

$$[S_i{\approx}S_j(\mathrm{x}_1,\ldots,\mathrm{x}_n)]^{fol} \;=\; S_i(\mathrm{x}_1,\ldots,\mathrm{x}_n) \;\wedge\; \forall \mathrm{u}_1,\ldots,\mathrm{u}_n(\neg S_i(\mathrm{u}_1,\ldots,\mathrm{u}_n) \rightarrow S_j(\mathrm{u}_1,\ldots,\mathrm{u}_n))$$

5. Aligning Embedded Contexts

So far we have looked at what we believe are the basic ways to align meaning representations for parallel texts. But there are further issues in meaning alignment, and as a matter of fact the machinery proposed so far isn't able to account for some problems that we encounter when we consider modals and negation. Consider the English sentence "The possibility of animals and birds bringing disease to the UK" and its German translation "Haustiere und Vögel können Krankheiten nach Großbritannien einschleppen." Both sentences contain a modal expression, expressed by a noun in English, and by a modal verb in German. Analogously to Example 1, we could analyze the English modal by introducing a hybrid modal operator (Bos, 2004). Now suppose that the German modal verb is semantically interpreted by the modal possibility operator $\diamond$. This would give the meaning representation as shown in Example 5.

EXAMPLE 5



There is some discrepancy between the monolingual semantic analyses: the hybrid modal operator (the colon :) that connects a propositional discourse referent with an embedded context in the English case, and the modal operator $\diamond$ in the German case. We could say that in such a case we would need to revise the semantics analysis either on the English or on the German side, to arrive at the same logical operator. An alternative solution, shown here in Example 5, is to decorate logical operators with a *language mode*. This way, we can combine several operators triggered by different languages into one and the same parallel meaning representation. A similar semantic mismatch arises with translating "not recommended" with the German verb "abraten". On the one side we face an explicit negation, and on the other side an implicit negation. Further empirical study is required to shed more light on this issue and evaluate the various possibilities for semantic alignment.

6. Discussion

In this paper we proposed a new formalism to align meaning representations of translated texts. We illustrated the formalism with several non-trivial examples for English–German translations. Certainly, there are many things that we did not consider: light verbs, tense, aspect, discourse relations, pronouns, anaphoric phenomena. Hence, a sensible question to ask is how representative the examples considered in this pilot study are and how and whether this method scales up to other phenomena and languages more distant from English than German.

The only answer we can give to this question is that one just needs to try and investigate, using the empirical method explored here. It is probably fair to point though that the examples that we discussed were not selected because they were easy to model. In fact we tried deliberately to find challenging examples with syntactic mismatches (such as the implicit vs. explicit negation). It seems that for closely related languages such as English and German the approach put forward in this paper is promising. For more distant languages, it could be that the same message is conveyed with very different syntactic structures, as the English–Korean pair[4] (“I have a headache” and its translation “nan-nun meri-ka aphuta”) in Example 6.

EXAMPLE 6

x y e
x ↦ “I” y ↦ “head” e ↦ ẋ have a ẏache”

HAVE-ACHE(e)
RECIPIENT(e,x)
THEME(e,y)
HEAD(y)

x y e
x ↦ “nan” y ↦ “meri” e ↦ “ẋ-nun ẏ-ka aphuta”

APHUTA(e)
NUN(e,x)
KA(e,y)
MERI(y)

↘ ↙

x y e
HAVE-ACHE≈APHUTA(e) RECIPIENT≡NUN(e,x) THEME≡KA(e,y) HEAD≡MERI(y)

This is an interesting example because to ensure a smooth alignment between the English and Korean sentence, it forces us to produce a non-literal semantic analysis of the English sentence. It also shows that thematic roles, at least under the analysis put forward here, are more commonly overtly expressed in languages other than English. But then, even within a single language, paraphrases with different syntactic structure should receive similar meaning representations: consider for instance “my head hurts” and “I have a headache”. In this particular case, a proper analysis of light verbs would strengthen semantic alignment.

Finally, we would like to remark that the assumptions that we have made for semantic representations are humble: meaning is described with the help of variables, n-place relations, a stock of non-logical symbols, and a couple of logical operators (the usual suspects, i.e. negation, disjunction, modalities). This is standard practice carried out by formal semanticists studying Germanic languages, and we don’t see any reason why it wouldn’t extend to more distant languages. It is an exercise that could lead not only to interesting language resources for machine translation applications, but also to get a better general understanding of cross-lingual semantic analysis.

[4] This example was kindly suggested to me by one of the anonymous reviewers of this paper.

7. References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August.

Bos, J. (2004). Computational Semantics in Discourse: Underspecification, Resolution, and Inference. *Journal of Logic, Language and Information*, 13(2):139–157.

Fellbaum, C., editor. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT.* Kluwer, Dordrecht.

A Proof-Based Annotation Platform of Textual Entailment

Assaf Toledo[1], Stavroula Alexandropoulou[1], Sophie Chesney[2],
Robert Grimm[1], Pepijn Kokke[1], Benno Kruit[3], Kyriaki Neophytou[1],
Antony Nguyen[1], Yoad Winter[1]

[1] - Utrecht University [2] - University College London [3] - University of Amsterdam
{a.toledo,s.alexandropoulou,y.winter}@uu.nl
sophie.chesney.10@ucl.ac.uk, {pepijn.kokke,bennokr}@gmail.com
{r.m.grimm,k.neophytou,a.h.nguyen}@students.uu.nl

Abstract

We introduce a new platform for annotating inferential phenomena in entailment data, buttressed by a formal semantic model and a proof-system that provide immediate verification of the coherency and completeness of the marked annotations. By integrating a web-based user interface, a formal lexicon, a lambda-calculus engine and an off-the-shelf theorem prover, the platform allows human annotators to mark linguistic phenomena in entailment data (pairs made up of a premise and a hypothesis) and to receive immediate feedback whether their annotations are substantiated: for positive entailment pairs, the system searches for a formal logical proof that the hypothesis follows from the premise; for negative pairs, the system verifies that a counter-model can be constructed. This novel approach facilitates the creation of textual entailment corpora with annotations that are sufficiently coherent and complete for recognizing the entailment relation or lack thereof. A corpus of several hundred annotated entailments is currently being compiled based on the platform and will be available for the research community in the foreseeable future.

Keywords: Annotation Platform, Semantic Annotation, Proof System, Formal Model, Textual Entailment, RTE

1. Introduction

The Recognizing Textual Entailment (RTE) corpora (Dagan et al., 2006; Bar Haim et al., 2006; Giampiccolo et al., 2008, a.o) present the challenge of automatically determining whether an entailment relation obtains between a naturally occurring text *T* and a manually composed hypothesis *H*.[1] These corpora, which are currently the only available resources of textual entailments, mark entailment candidates as positive/negative.[2] For example:

Example 1

- T: For their discovery of ulcer-causing bacteria, Australian doctors Robin Warren and Barry Marshall have received the 2005 Nobel Prize in Physiology or Medicine.
- H: Robin Warren was awarded a Nobel Prize.[3]
- Entailment: Positive

However, the linguistic phenomena that underlie entailment in each particular case and their contribution to inferential processes are not indicated in the corpora. In the absence of a gold standard that identifies linguistic phenomena triggering inferences, the inferential processes employed by entailment systems to recognize entailment are not directly accessible and, as a result, cannot be evaluated or improved straightforwardly.

We address this problem through the SemAnTE (Semantic Annotation of Textual Entailment) platform introduced in this paper. The platform allows human annotators to elucidate some of the central inferential processes underlying entailments in the RTE corpus. In 80.65% of the positive pairs in RTE 1–4, annotators found the recognition of entailment to rely on inferences stemming, *inter alia*, from the semantics of appositive, restrictive or intersective modification (Toledo et al., 2013). We decided to focus on the above three phenomena for two reasons. First, they are prevalent in the RTE datasets and, second, their various syntactic expressions can be modeled semantically using a limited set of logical concepts, such as equivalence, inclusion and conjunction.

The annotation platform allows the annotators to mark the above three modification patterns when they are involved in the recognition of entailment by binding the words and constructions in sentences to a lexicon of abstract semantic denotations. The proposed semantic modeling offers an important advantage: it licenses the system to search for formal proofs that substantiate manual annotations and to describe how the modeled phenomena interact and contribute to the recognition process. This is achieved by employing a lambda-calculus engine and a theorem prover.

The platform is currently employed for the preparation of a new corpus of several hundred annotated entailments comprising both positive and negative pairs. In the future, we plan to extend the semantic model to cover other, more complex phenomena.

[1] A short software demonstration paper describing the SemAnTE annotation platform is included in the EACL 2014 proceedings.

[2] Pairs of sentences in RTE 1-3 are categorized in two classes: *yes-* or *no-entailment*; pairs in RTE 4-5 are categorized in three classes: *entailment*, *contradiction* and *unknown*. We label the judgments *yes-entailment* from RTE 1-3 and *entailment* from RTE 4-5 as *positive*, and the other judgments as *negative*.

[3] Pair 222 from the development set of RTE 2.

2. Semantic Model

We model entailment in natural language based on order theory, on a working assumption that entailment describes a *preorder* relation on the set of all possible sentences. Thus, any sentence trivially entails itself (reflexivity); and given two entailments $T_1 \Rightarrow H_1$ and $T_2 \Rightarrow H_2$ where H_1 and T_2 are identical sentences, we assume $T_1 \Rightarrow H_2$ (transitivity). We use a standard model-theoretical extensional semantics, whereby each model M assigns sentences a truth-value in the set $\{0, 1\}$ – the domain of *truth-values* on which we assume the simple *partial order* $\leq$. We adapt Tarski's (1944) theory of truth to entailment relations and consider a theory of entailment adequate if the intuitive entailment preorder on sentences can be described as the pairs of sentences T and H whose truth-values $[\![T]\!]^M$ and $[\![H]\!]^M$ satisfy $[\![T]\!]^M \leq [\![H]\!]^M$ for all models M.

The function of annotations is to link between textual representations in natural language and model-theoretic representations. To this end, the words and structural configurations in T and H are marked with lexical labels that encode semantic meanings for the linguistic phenomena being modeled. These lexical labels are defined formally in a lexicon, as illustrated in Table 1 for major lexical categories over types: e for *entities*, t for *truth-values*, and the functional compounds of e and t.

Category	Type	Example	Denotation
Proper Name	e	Dan	**dan**
Indef. Article	$(et)(et)$	a	A
Def. Article	$(et)e$	the	ι
Copula	$(et)(et)$	is	IS
Noun	et	bacteria	**bacteria**
Intrans. verb	et	sit	**sit**
Trans. verb	eet	receive	**receive**
Pred. Conj.	$(et)((et)(et))$	and	AND
Res. Adj. (Mod)	$(et)(et)$	short	$R_m(\mathbf{short})$
Res. Adj. (Pred)	et	short	$P_r(\mathbf{short})$
Res. Adj. (Mod)	$(et)(et)$	thin	$R_m(\mathbf{thin})$
Res. Adj. (Pred)	et	thin	$P_r(\mathbf{thin})$
Int. Adj. (Mod))	$(et)(et)$	Dutch	$I_m(\mathbf{dutch})$
Int. Adj. (Pred))	et	Dutch	**dutch**
Exist. Quant.	$(et)(et)t$	some	SOME

Table 1: Lexicon Illustration

Denotations that are assumed to be arbitrary are given in boldface. For example, the intransitive verb *sit* is assigned the type et, which describes functions from entities to truth-values, and its denotation **sit** is an arbitrary function of this type. The denotations of several other lexical items are restricted by the given model M. As illustrated in Figure 1, the coordinator and is assigned the type $(et)((et)(et))$, and its denotation is a function that takes a function A of type et and returns a function that takes a function B, also of type et, and returns a function that takes an entity x of type e and returns 1 if and only if x satisfies both A and B.

Attaching lexical labels to words and syntactic constructions enables annotators to mark the linguistic phenomena manifested in the data. Moreover, by virtue of its formal foundation, this approach allows annotators to verify that the entailment relation (or lack thereof) that obtains between the textual forms of T and H is also present between their respective semantic forms. This latter step ensures that the annotations provide sufficient information for recognizing the entailment relation in a given pair based on the semantic abstraction. For example, consider the simple entailment *Dan is short and thin* ⇒ *Dan is short* and assume annotations of *Dan* as a proper name, *short* and *thin* as restrictive modifiers in predicate position, and *and* as predicate conjunction. The formal model can be used to verify these annotations by constructing a proof as follows:

For each model M, $[\![\textit{Dan [is [short [and thin]]]}]\!]^M$

$$\begin{array}{lll}
= & (\textsc{is}((\textsc{and}(P_r(\mathbf{thin})))(P_r(\mathbf{short}))))(\mathbf{dan}) & \text{analysis} \\
= & (((\lambda A_{et}.\lambda B_{et}.\lambda x_e.A(x) \wedge B(x)) & \text{def. of IS} \\
 & (P_r(\mathbf{thin})))(P_r(\mathbf{short})))(\mathbf{dan}) & \text{and AND} \\
= & P_r(\mathbf{thin})(\mathbf{dan}) \wedge P_r(\mathbf{short})(\mathbf{dan}) & \text{func. app.} \\
\leq & P_r(\mathbf{short})(\mathbf{dan}) & \text{def. of } \wedge \\
= & (\textsc{is}(P_r(\mathbf{short})))(\mathbf{dan}) & \text{def. of IS} \\
= & [\![\textit{Dan is short}]\!]^M & \text{analysis}
\end{array}$$

$$\begin{array}{l}
\textsc{a} = \textsc{is} = \lambda A_{et}.A \\
\iota = \lambda A_{et}.\begin{cases} a & A = (\lambda x_e.x = a) \\ \text{undefined} & \text{otherwise} \end{cases} \\
\textsc{who}_A = \lambda A_{et}.\lambda x_e.\iota(\lambda y.y = x \wedge A(x)) \\
R_m = \lambda M_{(et)(et)}.\lambda A_{et}.\lambda x_e.M(A)(x) \wedge A(x) \\
P_r = \lambda M_{(et)(et)}.\lambda x_e.M(\lambda y_e.1)(x) \\
\textsc{some} = \lambda A_{et}.\lambda B_{et}.\exists x.A(x) \wedge B(x) \\
\textsc{and} = \lambda A_{et}.\lambda B_{et}.\lambda x_e.A(x) \wedge B(x)
\end{array}$$

Figure 1: Functions in the Lexicon

3. Platform Architecture

The platform's architecture is based on a client-server model, as illustrated in Figure 2.

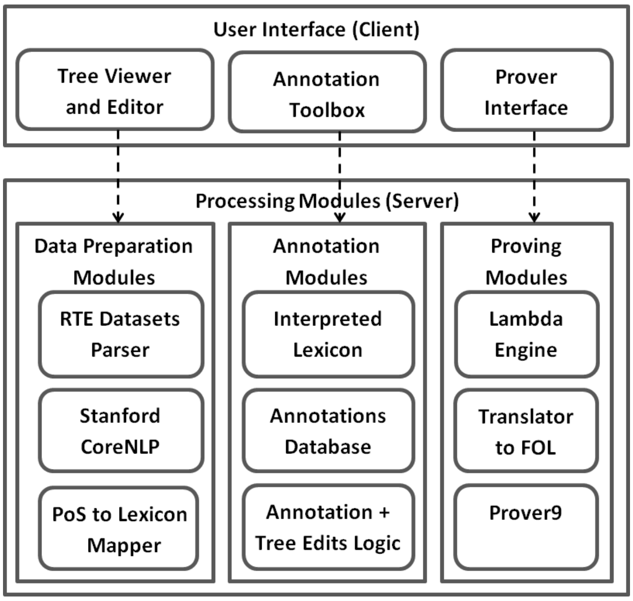


Figure 2: Platform Architecture

The user interface (UI) is implemented as a web-based client using Google Web Toolkit (Olson, 2007) and allows multiple annotators to access the RTE data, to annotate them, and to substantiate their annotations. These operations are done by invoking corresponding remote procedure calls at the server side. We describe the system components as we go over the work-flow of annotating Example 1.

Data Preparation: We extract T-H pairs from the RTE datasets XML files and use the Stanford CoreNLP (Klein and Manning, 2003; Toutanova et al., 2003; de Marneffe et al., 2006) to parse each pair and to annotate it with part-of-speech tags.[4] Subsequently, we apply a naive heuristic to map the PoS tags to the lexicon.[5] This process is performed as part of the platform's installation and when annotators need to simplify the original RTE data in order to avoid syntactic/semantic phenomena that the semantic engine does not support. For example, the fronted *for*-phrase *For their discovery...* is moved after the object of the verb *receive* as fronted adjuncts are not supported. Additionally, the phenomenon of distributivity manifested in the inference *Robin Warren and Barry Marshall have received...* → *Robin Warren has received...*, which is required for recognizing the entailment in this example. We do not model this inference and the construction must therefore be simplified. These simplifications yield T_{simple} and H_{simple} as follows:

- T_{simple}: The Australian doctor Robin Warren has received the great Nobel Prize in Physiology-Medicine for the discovery of the ulcer-causing bacteria.
- H_{simple}: Robin Warren was awarded a Nobel Prize.

Annotation: The annotation is done by marking the tree-leaves with entries from the lexicon. For example, *receives* is annotated as a transitive verb, *ulcer-causing* is annotated as a restrictive modifier (*MR*) of the noun *bacteria*, and *Australian* is annotated as an intersective modifier of the noun *doctors*. In addition, annotators add leaves that mark semantic relations. For instance, a leaf that indicates the apposition between *The Australian doctor* and *Robin Warren* is added and annotated as WHO_A. Furthermore, the annotators fix parsing mistakes as in *the great Nobel Prize in Physiology–Medicine* which was parsed as: [the [great [Nobel Prize]]] [in Physiology–Medicine] and fixed to: [the [great [[Nobel Prize] [in Physiology–Medicine]]]]. The server stores a list of all annotation actions. Figure 3 shows the tree-view, lexicon, prover and annotation history panels in the UI.

Defining Lexical Relations: Our modeling of modification phenomena does not address inferences that rely on lexical knowledge, as in: "Robin Warren has received a prize" → "Robin Warren was awarded a prize". Such lexical relations between the text and hypothesis are marked by the annotators and translated into logical formulas by the proof-system.

Proving: Once all leaves are annotated and the tree structures of T_{simple} and H_{simple} are manipulated, the annotators use the prover interface to request a search for a proof indicating that their annotations are substantiated. First, the system uses lambda calculus reductions to create logical forms that represent the meanings of T_{simple} and H_{simple} in higher-order logic. At this stage, type errors may be reported due to erroneous parse-trees or annotations. In this case an annotator will fix the errors and re-run the proving step. Second, once all type errors are resolved, the higher-order representations are lowered to first order and Prover9 (McCune, 2010) is executed to search for a proof between the logical expressions of T_{simple} and H_{simple}.[6] The proofs are recorded in order to be included in the corpus release. Figure 4 shows the result of translating T_{simple} and H_{simple} to an input to Prover9.

4. Corpus Preparation

We have so far completed annotating 40 positive entailments based on data from RTE 1-4. The annotators are thoroughly familiar with the data and have extensive experience in recognizing entailments stemming from appositive, restrictive and intersective modification. While compiling a corpus of several hundred entailment pairs, we are also working to extend our model to recognize inferences produced by a wider range of linguistic phenomena. The objective is to minimize the need for simplifying the input utterances so as to make them compatible to the model.

```
formulas(assumptions).
% Pragmatics:
all x0 (((nobel_prize(x0) & in_nobel_prize(Physiology_
Medicine, x0)) & great_nobel_prize_in(Physiology_Medicine,
x0)) ↔ x0=c219).
all x0 ((doctor(x0) & australian_doctor(x0)) ↔ x0=c221).
all x0 ((x0=c221 & x0=Robin_Warren) ↔ x0=c220).
all x0 ((bacteria(x0) & ulcer_causing_bacteria(x0)) ↔
x0=c223).
all x0 ((discovery(x0) & of_discovery(c223, x0)) ↔
x0=c222).

% Semantics:
(received(c219, c220) & for_received(c219, c222, c220)).
all x0 (all x1 (received(x0, x1) → awarded(x0, x1))).

end_of_list.

formulas(goals).
exists x0 (nobel_prize(x0) & awarded(x0, Robin_Warren)).
end_of_list.
```

Figure 4: Input for Theorem Prover

5. Conclusions

This paper proposes a novel concept for an annotation platform buttressing a proof-system designed to substantiate a semantic annotation scheme for inferences stemming from modification phenomena. This method guarantees that the manual annotations constitute a complete description of a given entailment relation and facilitates the creation of a

[4] Stanford CoreNLP version 1.3.4

[5] This heuristic is naive in the sense of not disambiguating verbs, adjectives and other types of terms according to their semantic features. It is meant to provide a starting point for the manual annotation process.

[6] Prover9 version 2009-11A

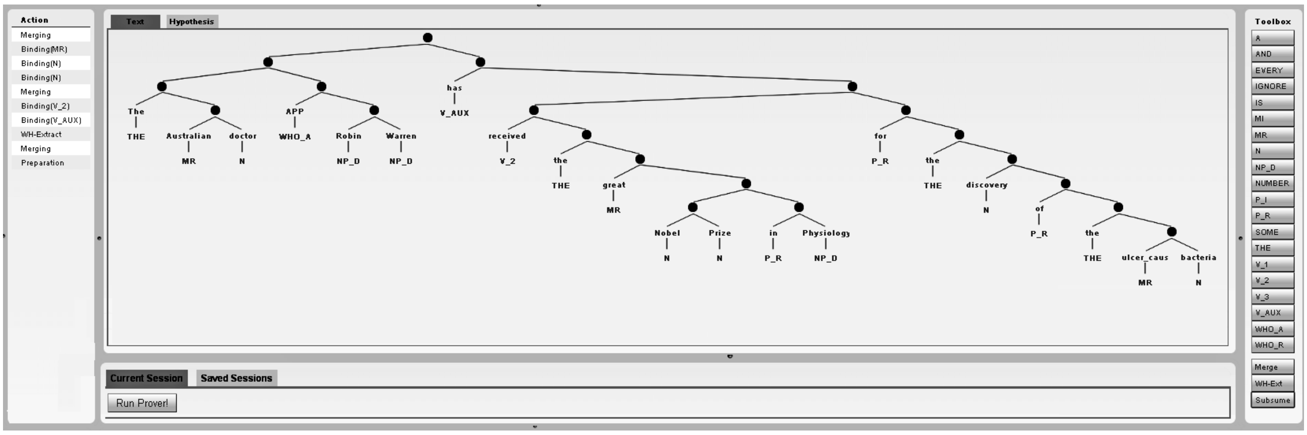


Figure 3: User Interface Panels: Annotation History, Tree-View, Prover Interface and Lexicon Toolbox

gold-standard of such phenomena. A new corpus is currently being developed and will be publicly available for the research community in the foreseeable future.

Acknowledgments

The work of Stavroula Alexandropoulou, Robert Grimm, Sophie Chesney, Pepijn Kokke, Benno Kruit, Kyriaki Neophytou, Antony Nguyen, Assaf Toledo and Yoad Winter was supported by a VICI grant number 277-80-002 by the Netherlands Organisation for Scientific Research (NWO).

6. References

Bar Haim, Roy, Dagan, Ido, Dolan, Bill, Ferro, Lisa, Giampiccolo, Danilo, Magnini, Bernardo, and Szpektor, Idan. (2006). The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Dagan, Ido, Glickman, Oren, and Magnini, Bernardo. (2006). The pascal recognising textual entailment challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190.

de Marneffe, Marie-Catherine, MacCartney, Bill, and Manning, Christopher D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology*. The Stanford Natural Language Processing Group.

Giampiccolo, Danilo, Dang, Hoa Trang, Magnini, Bernardo, Dagan, Ido, and Cabrio, Elena. (2008). The fourth pascal recognising textual entailment challenge. In *TAC 2008 Proceedings*.

Klein, Dan and Manning, Christopher D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. ACL.

McCune, William. (2010). Prover9 and Mace4. `http://www.cs.unm.edu/~mccune/prover9/`.

Olson, Steven Douglas. (2007). *Ajax on Java*. O'Reilly Media.

Tarski, Alfred. (1944). The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research*, 4(3):341–376.

Toledo, Assaf, Alexandropoulou, Stavroula, Katrenko, Sophia, Klockmann, Heidi, Kokke, Pepijn, and Winter, Yoad. (2013). Semantic Annotation of Textual Entailment. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 240–251, Potsdam, Germany, March. Association for Computational Linguistics.

Toutanova, Kristina, Klein, Dan, Manning, Christopher D., and Singer, Yoram. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. ACL.

Semantic Annotation of the Danish CLARIN Reference Corpus

Bolette S. Pedersen[1], Sanni Nimb[2], Sussi Olsen[1], Anders Søgaard[1], Nicolai Sørensen[2]
[1]University of Copenhagen & [2]The Society for Danish Language and Literature
E-mail: bspedersen@hum.ku.dk, sn@dsl.dk, saolsen@hum.ku.dk, soegaard@hum.ku.dk, nhs@dsl.dk

Abstract

The newly initiated project "Semantic Processing across Domains" is granted by the Danish Research Council for Culture and Communication and runs for the period 2013-2016. It focuses on Danish as a low-resourced language and aims at increasing the level of technological resources available for the Danish HLT community. A primary project goal is to provide semantically annotated text corpora of Danish following agreed standards and to let these serve as training data for advanced machine learning algorithms which will address data scarcity and domain adaptation as central problem areas. The Danish CLARIN Reference Corpus - supplemented by a selection of additional text types from social media and the web - are being sense and role annotated. We experiment with an adaptation of PropBank roles to Danish as well as with a scalable sense inventory of Danish. This inventory spans from supersense annotations (semantic classes) to wordnet-derived sense annotations which rely on a distinction between ontological types and main and subsenses. The annotation tool WebAnno, which is being developed as part of the German CLARIN project, is applied for the annotation task.

Keywords: Semantic annotation, word sense disambiguation, semantic role labelling, DK-CLARIN

1. Semantic annotation and Danish HLT resources

The newly initiated project "Semantic Processing across Domains" is granted by the Danish Research Council for Culture and Communication and runs for the period 2013-2016. It is a collaborate project between the University of Copenhagen and The Society for Danish Language and Literature (DSL), which is an independent institution editing and publishing Danish texts and dictionaries on a scholarly basis. The project focuses on Danish as a low-resourced language and aims at increasing the level of technological resources available for the Danish HLT community, which according to the META-NET White Paper Series (See Pedersen et al. 2012) falls in the category 'fragmentary' compared to other European languages. To this end, a primary project goal is to provide semantically annotated text corpora of Danish following agreed standards and to let these serve as training data for machine learning algorithms which address data scarcity and domain adaptation as central problem areas (Søgaard 2013).

2. CLARIN resources used and developed in the project

The DK-CLARIN project (2008-2011) aimed at initiating a Danish research infrastructure for the Humanities by integrating written, spoken, and visual records in a common technological infrastructure. This work is currently continued within the Danish DigHumLab[1] and European CLARIN [2] projects, and several resources deriving from this work serve as background resources for the present annotation project. This regards in particular the CLARIN Reference Corpus (Asmussen 2012) which is the main target for our annotation even if the material is augmented with a selection of additional text types from social media. The CLARIN Reference Corpus contains approx. 45 million words covering newspapers, magazines, oral (but transcribed) congress debates, web pages, blogs etc. DK-CLARIN also financed the finalization of the Danish wordnet, DanNet (Pedersen et al. 2009), which is applied as a lexical resource for word sense annotation in the project and which – together with a medium-sized dictionary of Danish, The Danish Dictionary (DDO, developed by DSL), forms the basis for experiments with a scalable sense inventory. Both the corpus and the wordnet are available via the META-SHARE and DK-CLARIN platforms.

3. Three semantic annotation tasks

The project is concerned with semantic annotation at word and sentence level. Currently, three specific annotation tasks are embarked:

- A lexical sample task with a selected subset of nouns and verbs to be sense annotated on the basis of DanNet and DDO[3]
- An all-words task with coarse-grained sense annotations based on so-called supersenses (or 'semantic classes')
- Annotation of semantic roles relying on a transfer of PropBank roles (Palmer et al. 2005) to Danish.

In the following the three tasks are described in more detail.

3.1 Word sense annotation based on DanNet

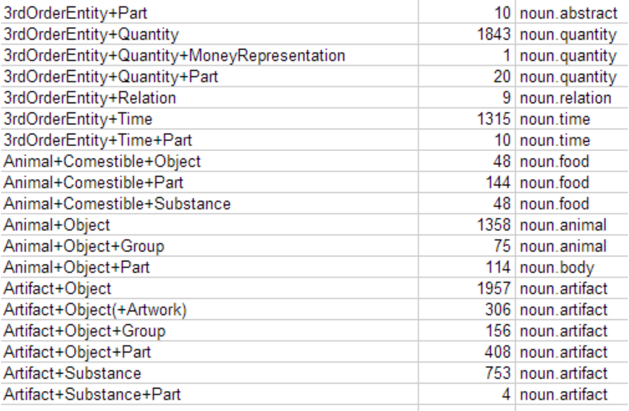
One aim of the project is to experiment with a scalable

[1] See http://dighumlab.com/
[2] See http://www.clarin.eu/

[3] For this task, the CLARIN Reference Corpus will be augmented with manually selected examples from additional corpora and from the web.

sense inventory for word sense disambiguation (wsd). In contrast to most other wordnets, DanNet is compiled from the corpus-based definitions and sense distinctions provided in DDO with clear distinctions between main and sub-senses. This distinction opens for the possibility of automatic generation of sense inventories of varying granularity and for an examination of the inter-coder agreement that is achieved with these individual inventories (a comparable approach is employed in the MASC project at a larger scale, cf. Passonneau et al. 2012 and de Melo et al. 2012). We hypothesize that the main senses in DanNet in combination with its ontological types (such as Person, Semiotic Artifact, Building, Time, Measurement[4]) will provide the basis for a practically more adequate and theoretically well-founded sense inventory for word sense disambiguation than what is seen in several comparable wordnet resources where rather finegrained and unstructured sense enumerations are applied (for discussions, see Ide & Wilks 2007, Kilgarriff 2007, Brown et al. 2010, Vossen et al. 2011, de Melo et al. 2012 among many others). An initial approach is therefore to automatically collapse sub-senses of a word with their main sense, unless a sub-sense has another ontological type or topic than the main sense – a case which is typically seen with metaphorical or very specialized senses.

For comparison to an even coarser inventory, and in order to provide also sense annotations that are directly comparable and interoperable across semantic corpus resources for other languages, all DanNet senses are mapped to the so-called 'supersense'-inventory (corresponding roughly to semantic classes) derived from the wordnet lexico-grapichal classes (cf. Ciaramita & Johnson 2003 and http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html)[5] using a transfer scheme (Figure 1).



3rdOrderEntity+Part	10	noun.abstract
3rdOrderEntity+Quantity	1843	noun.quantity
3rdOrderEntity+Quantity+MoneyRepresentation	1	noun.quantity
3rdOrderEntity+Quantity+Part	20	noun.quantity
3rdOrderEntity+Relation	9	noun.relation
3rdOrderEntity+Time	1315	noun.time
3rdOrderEntity+Time+Part	10	noun.time
Animal+Comestible+Object	48	noun.food
Animal+Comestible+Part	144	noun.food
Animal+Comestible+Substance	48	noun.food
Animal+Object	1358	noun.animal
Animal+Object+Group	75	noun.animal
Animal+Object+Part	114	noun.body
Artifact+Object	1957	noun.artifact
Artifact+Object(+Artwork)	306	noun.artifact
Artifact+Object+Group	156	noun.artifact
Artifact+Object+Part	408	noun.artifact
Artifact+Substance	753	noun.artifact
Artifact+Substance+Part	4	noun.artifact

Figure 1. Extract of transfer scheme from EuroWordNet ontological types used in DanNet to supersenses (numbers in middle column indicate number of synsets)

To exemplify the distinction between supersenses and the finer sense inventory that we generate from DanNet and DDO, consider the lexemes *pande* (pan, forehead) and *kort* (map, card, playing card). The two unrelated meanings of *pande* will be maintained in both approaches (noun.artifact and noun.body in supersense terms), whereas the different meanings of *kort* will be collapsed into one with the supersense approach (corresponding to the supersense noun.artifact), but maintained in the more fine-grained approach. However, both approaches will generalize over the dictionary distinctions between postcards, admissions cards and id cards since these are all considered to be subsenses in DDO with the same ontological type in DanNet.

Previous to the manual annotation, all corpus data are pre-annotated based on DanNet (for annotation tool, see Section 4). In cases of more than one sense or supersense, the annotator will choose between the pre-annotated ones. In all situations, the annotator can overrule a pre-annotated sense and assign an alternative sense. Unknown words (which are mostly compounds) are obviously not pre-annotated; in these cases the annotators will pick the most appropriate sense from a pick list. Figure 2 shows an annotation task with pre-annotated data.

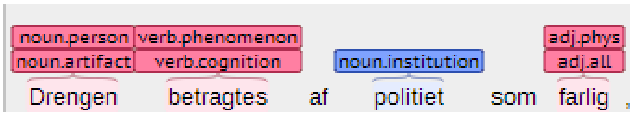


Figure 2. Pre-annotated corpus extract for the annotator to refine (all-words task) (lit: '*The boy is considered by the police as dangerous'*)

Three annotators will work on each corpus extract and a gold standard annotation will be decided upon by the third annotator, who has the role of curator.

3.2 Annotation of semantic roles

This part of the annotation project plans to relate to recent ISO standards for semantic role annotation (Bunt & Palmer 2013), and will include a transfer of PropBank roles to Danish, relying in addition on existing descriptive works on Danish verbs. These include The Danish Thesaurus (DT) which is recently being published by DSL (cf. Nimb et al. 2013) and which group verbs in a FrameNet-like fashion. The thesaurus consists of scenarios described by the same words as the ones evoking a semantic frame. Information about the function of these groups in terms of arguments in a frame description is implicitly given in the metadata of the semantic subsections, i.e. in the type of group and from assigned relations of the type involved_agent and involved_patient, making DT a rich background resource for the constructing a lexical resource for semantic role

[4] The ontological types in DanNet are adapted from the EuroWordNet Ontology (Vossen et al. 1999).

[5] For the all-words task, only the supersenses are annotated.

annotation. Further, a FrameNet-like resource has been developed by Bick (2011). This resource uses the semantic verb classification of DanNet and includes a set of all in all 38 semantic roles.

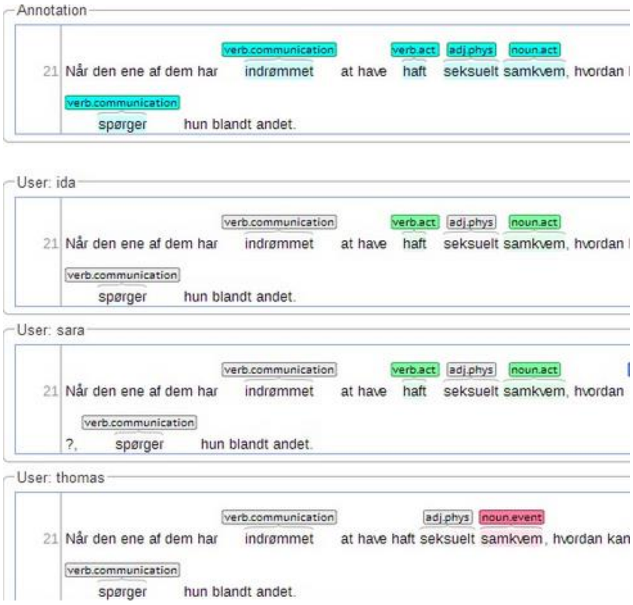
Following roughly a transfer approach tested by the Dutch Language Corpus Initiative (D-Coi) (Monachesi & Chapman 2006), we foresee the following steps:

1. Localize the verb sense
2. If the verb sense is not yet part of the Danish frames file, translate it to English
3. Check the verb's frames file in PropBank
4. Localize the arguments and modifiers of the verb; compare and adjust according to DT and Danish FrameNet
5. Extent the Danish frame file with synonyms from DT

In contrast to PropBank which employs phrase structure trees for the syntactic structure, we plan to assign semantic roles onto a dependency parsed version of the corpus.

4. Applied annotation tool: WebAnno

For the first annotation task, which was initiated in January 2014 and concerns the lexical all-words task, we apply the browser-based tool WebAnno. This tool is being developed as part of the German CLARIN project (CLARIN-D), cf. Yimam et al. 2013.



21 Når den ene af dem har indrømmet at have haft seksuelt samkvem, hvordan
spørger hun blandt andet.

21 Når den ene af dem har indrømmet at have haft seksuelt samkvem, hvordan
spørger hun blandt andet.

21 Når den ene af dem har indrømmet at have haft seksuelt samkvem, hvordan
?, spørger hun blandt andet.

21 Når den ene af dem har indrømmet at have haft seksuelt samkvem, hvordan kan
spørger hun blandt andet.

Figure 3. The curator function in WebAnno where the gold standard is developed based on previous manual annotations (‘*When one of the two has admitted to have had a sexual relationship, how can..*’)

The tool supports the annotation of a variety of linguistic levels and it is interoperable with a variety of data formats. Further it supports project management of the annotation tasks and allows for dynamic quality judgments by integrating measures of inter-annotator agreement and curator facilities. Figure 3 shows how annotation disagreements are detected and the gold standard developed.

Another central feature for the tasks foreseen in the project is the previously mentioned possibility of including pre-annotations as seen in Figure 2.

To our knowledge, we are the first team to apply WebAnno for *semantic* annotation. Since new features and some optimization is still foreseen in the experimental phase, the task is currently being hosted by Darmstadt University (where WebAnno is being developed), but the hosting role will be taken over by the University of Copenhagen during spring 2014.

5. Concluding remarks and future work

The original DK-CLARIN resources included both text and lexical-semantic resources, but hardly any links between the two. With this project we move towards an integration of text and lexical resources by annotating a part of the CLARIN Reference Corpus with word senses and semantic roles following agreed standards. Further, the annotations will serve as training data to machine learning algorithms that will enable us to automatically annotate larger amounts of data with a certain margin of correctness. The main focus in this part of the project will be to investigate methods for making joint learning of SRL-WSD less sensitive to the amounts of annotated data available and to domain differences (cf. Søgaard 2013).

DSL foresees to use the manually and automatically annotated corpora as an extra on-line citation resource for their dictionaries, illustrating the specific dictionary senses and realizations via direct links from the dictionary sense to examples in a corpus. At present DSL already offers advanced searching for part-of-speech, morphological features and some syntactical ones. Adding the possibility of enhancing the search pattern with word senses and semantic roles is the next logical step, and such an extension will allow for example to observe semantic meaning changes over time in a more systematic way.

References

Asmussen, Jørg (2012). CLARIN-Referencekorpus. Talk at Sprogteknologisk Workshop, University of Copenhagen October 31, 2012. http://cst.ku.dk/Workshop311012/sprogtekno2012.pdf

Bick, Eckhard (2011). A FrameNet for Danish. In: Proceedings of NODALIDA 2011, May 11-13, Riga, Latvia. *NEALT Proceedings Series, Vol 11*, pp.34-41.

Brown, S.W., T. Rood, & M. Palmer. (2010). Number or Nuance: Which factors restrict reliable word sense annotation? *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*. Valetta, Malta.

Bunt, H. and M. Palmer (2013) Conceptual and representational choices in defining an ISO standard for semantic semantic role annotation. In: *Proceedings Ninth Joint ISO - ACL Workshop on Interoperable Semantic Annotation (ISA-9)*, Potsdam, March 2013.

Ciaramita, Massimiliano, Mark Johnson (2003). Supersense Tagging of Unknown Nouns in WordNet *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

Ida, Nancy & Yorick Wilks (2007) Making Sense About Sense. In: E. Agirre & P. Edmonds: *Word Sense Disambiguation - Algorithms and Applications*. Springer.

Kilgarriff, Adam (2007). Word senses. In: E. Agirre & P. Edmonds: *Word Sense Disambiguation - Algorithms and Applications*. Springer.

Melo, Gerhard de, Collin F. Baker, Nancy Ide, Rebecca J. Passonneau, Christiane Fellbaum (2012) Empirical Comparisons of MASC Word Sense Annotations. *Proceedings from LREC 2012*, Istanbul, Turkey.

Monachesi, P. and J. Trapman. (2006). Merging FrameNet and PropBank in a corpus of written Dutch. In Proceedings of the workshop Merging and layering linguistic information. *Workshop held in conjunction with LREC 2006, Genoa - Italy, 23* May 2006. pp. 32-39.

Nimb, Sanni, Bolette S. Pedersen, Anna Braasch, Nicolai H. Sørensen and Thomas Troelsgård (2013) Enriching a wordnet from a thesaurus. *Workshop Proceedings on Lexical Semantic Resources for NLP from the 19th Nordic Conference on Computational Linguistics (NODALIDA*). *Linköping Electronic Conference Proceedings; Volume 85.* Oslo, Norway.

Palmer, M. D. Gildea, P. Kingsbury. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles, *Computational Linguistics Journal, 31:1*

Passonneau, Rebecca, Collin Bakery, Christiane Fellbaumz, Nancy Ide (2012). The MASC Word Sense Sentence Corpus. *Proceedings from LREC 2012*, Istanbul, Turkey.

Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen**,** L. Trap-Jensen, H. Lorentzen (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation, Computational Linguistics Series. Volume 43, Issue 3, Page 269-299.*

Pedersen, B.S, J. Wedekind, S. Kirchmeier-Andersen, S. Nimb, J.E. Rasmussen, L.B. Larsen, S. Bøhm-Andersen, H.Erdman Thomsen, P. J. Henrichsen,J. O. Kjærum, P. Revsbech, S.Hoffensetz-Andresen, B. Maegaard (2012). The Danish Language in the Digital Age - Det danske sprog i den digitale tidsalder. *META-NET White Paper Series, Springer Verlag.*

Søgaard, Anders. (2013). *Semi-supervised learning and domain adaptation in natural language processing*. Morgan & Claypool.

Yimam, S.M., Gurevych, I., Eckart de Castilho, R., and Biemann C. (2013): WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of ACL-2013, demo session*, Sofia, Bulgaria.

Vossen, Piek (ed). (1999). *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. The Netherlands: Kluwer Academic Publishers.

Vossen, Piek, Attila Görög, Fons Laan, Maarten van Gompel, Rubén Izquierdo, Antal van den Bosch (2011) DutchSemCor: Building a semantically annotated corpus for Dutch. *Proceedings of Electronic Lexicography in the 21st century: New Applications for new users (eLEX2011),* Slovenia.

Semantic Annotation of Anaphoric Links in Language

Kiyong Lee

Korea University
Seoul 137-767, Korea
ikiyong@gmail.com

Abstract

This paper attempts to integrate several existing coreference annotation schemes into an extended annotation scheme $\mathcal{AS}_{ana}$. The proposed $\mathcal{AS}_{ana}$ allows some other types of the anaphor-antecedent relation, called 'anaphoric link', than the canonical type of coreference that implies the referential identity between an anaphor and its antecedent. The structure of $\mathcal{AS}_{ana}$ itself is very simple, consisting of a single entity type for mentions and a single anaphoric relation, each of which is characterized by a small set of attribute-value specifications. Constrained by these specifications, $\mathcal{AS}_{ana}$ supports a two-step annotation procedure: For a given text T, (1) identification of a set of mentions M in the text T that refer to something in the universe of discourse referents as its *markables*, (2a) identification of a set of pairs of the mentions in M that are anaphorically related, and (2b) specification of the type of such a relation.

anaphor, anaphoric link, annotation, annotation scheme (AS), antecedent, coreference, discourse referent, markable, mention

1. Introduction

There appeared a little gossip item in *The Telegraph*.

(1) By Lucy Kinder 5:45PM GMT 24 Feb 2014
{{A former head teacher}$_{mt1}$ of {a primary school}$_{mt2}$ }$_{mt3}$ spanked {his$_{mt4}$ wife$_{mt5}$}$_{mt6}$ as punishment$_{mt7}$ for having {an affair}$_{mt8}$ after drawing up {pie charts}$_{mt9}$ to detail how much pain$_{mt10}$ she$_{mt11}$ had caused him$_{mt12}$.
{Graham Jones}$_{mt13}$, 44, told {his$_{mt14}$ wife$_{mt15}$}$_{m16}$ that she$_{mt17}$ would receive {eight blows}$_{mt18}$ - one$_{mt19}$ for {each day}$_{mt20}$ he$_{mt21}$ had known about {her$_{mt22}$ relationship$_{mt23}$}$_{mt24}$.
Jones$_{mt25}$ bent {his$_{mt26}$ wife$_{mt27}$}$_{mt28}$ over {a freezer }$_{mt29}$ in {their$_{mt30}$ cellar$_{mt31}$}$_{mt32}$ and spanked her$_{mt33}$ with {a plimsoll}$_{mt34}$ before she$_{mt35}$ escaped to {the kitchen}$_{mt36}$.
There$_{mt37}$ he$_{mt38}$ told her$_{mt39}$ she$_{mt40}$ had {four more hits}$_{mt41}$ to come and used {a spatula}$_{mt42}$ to deliver them$_{mt43}$.
Crying and screaming, {Mrs Jones}$_{mt44}$ was left with bruising and reddening.

This short news item contains at least 44 noun phrases (NPs) of various forms that mention or refer to something in the universe of discourse, a non-empty set of discourse referents (see Kamp (1981) and Kamp & Ryele (1993)). Each of these phrases, which are technically called 'mentions', is uniquely identified with an integer and its prefix *mt* that stands for *mention* in the text.[1]

Most of the mentions are referentially grounded, namely referring expressions, and some of them are also identified as having coreferentially related to others. To understand the whole story told by this news article, one should recognize these coreferential relations among the so-called mentions. The noun phrase *A former head teacher*$_{mt1}$ or the larger phrase *A former head teacher of a primary school*$_{mt3}$ and the name *Graham Jones*$_{mt13}$, for instance, refer to the same person and so do the name *Mrs Jones*$_{mt44}$ and the noun phrase *his wife*$_{mt6}$.

Each of such pairs that refer to the same entity in the discourse referents consists of two mentions, one called 'anaphor' and the other its 'antecedent'. They are also said to corefer or be coreferential.

The antecedent of an anaphor may be split into more than one. The pronoun *their*$_{mt25}$ in the news article, for instance, has its antecedent split into two mentions, *Jones*$_{mt21}$ and *his wife*$_{mt23}$.[2] There are half a dozen occurrences of pronouns such as *he, his, she*, and *her* that refer to either *Graham Jones*$_{mt13}$ or *his wife*$_{mt6}$ and there is also a locative pronoun *There*$_{mt37}$ that refers to *the kitchen*$_{mt36}$. The two expressions, *an affair*$_{mt8}$ and *her relationship*$_{mt24}$, may also be understood to corefer, for they can be interpreted as referring to the same event or state of affairs, provided events are included as first-class citizens among discourse referents.

Reference or coreference resolution is a big issue in computational linguistics, especially in the area of information extraction (IE). There have been several important publications that deal with that issue: to cite some, we have: Hirschman & Chinchor (1997), Chen & Hacioglu (2006), Haghighi & Klein (2009), Haghighi & Klein (2010), Rahman & Ng (2011), Stoyanov et al. (2010), Stoyanov & Eisner (2012), Ratinov & Roth (2012), and van Deemter & Kibble (2000).

While referring to these works and others to be cited, this paper aims at constructing a semantic annotation scheme (AS) for coreference and other anaphoric link phenomena in a language (English) that may be proposed as an ISO standard for language resources management. First, this paper reviews the four existing ASs: (1) the TEI-based ASs, Bruneseaux and Romary (1997) and TEI P5 (2014), (2) Hirschman & Chinchor (1997)'s MUC-7 Coreference Task Definition (CTD), (3) Müller & Strube (2006)'s MMAX2-based multi-level AS, and (4) Pustejovsky et al. (2013)'s ISO-Space AS.

[1] Not every mention refers to an (individual) entity (e.g., *no man*).

[2] Pustejovsky et al. (2013) calls this a case of split antecedent, whereas Rullmann (2003) views this as an instance of multiple antecedent.

Second, this paper proposes to integrate all these ASs into a two-level AS: Given an input text, it specifies (1) ways of identifying the whole set of mentions as possible markables and then (2) ways of (a) selecting possible anaphors from the set of possible markables, (b) pairing each of the selected anaphors with their antecedent(s), and (c) specifying the type of each anaphor-antecedent relation, called 'anaphoric' link. The second level is complex: it consists of three steps (a), (b), and (c). Each of these steps is constrained and triggered by a small set of attribute-value specifications for each anaphor-antecedent pair of the mentions and each type of the anaphoric link.
The rest of the paper develops as follows: Section 2 Review of Existing Annotation Schemes, Section 3 Specification of the Proposed Integrated Annotation Scheme, and Section 4 Illustrations, Section 5 Semantic Interpretations, and Section 6 Concluding Remarks.

2. Review of Existing Annotation Schemes

2.1. Preliminaries

There are several types of the anaphoric link. The best known type is coreference, an equivalence (symmetric, transitive, and reflexive) relation, that holds between two terms, called 'mentions', if and only if their denotations are identical.

(2) Two terms t_1 and t_2 corefer iff $[|t_1]^M$ = $[|t_2|]^M$, where $[|t_i|]^M$ is the denotation or referent of a term t_i with respect to a model M.

Coreference and binding are two different, but related linguistic phenomena, often discussed together. Consider:

(3) a. *John* loves *his* mother. [coreference]
b. *Everyone* loves *his* mother. [binding]

In (a), the pronoun *his* may be understood as coreferring with the name *John* as its antecedent. In (b), on the other hand, the pronoun *his* does not corefer with the quantified noun phrase *everyone*, but is treated in formal semantics as a variable bound by the universal quantifier. This paper attempts to accommodate both the type of coreference, as defined in (2), and some other types of anaphoric phenomena into the proposed $\mathcal{AS}_{ana}$.

2.2. Coindexing

In linguistics, coreference and binding are both annotated in the same manner by coindexing, as shown below.

(4) a. John$_i$ loves his$_i$ mother. [coreference]
b. Everyone$_i$ loves his$_i$ mother. [binding]

Split antecedents can also be represented by coindexing with a set index such as $\{x, y\}$. Here are some examples, taken from Rullmann (2003), (5a,b,c):

(5) a. Mary$_1$ told John$_2$ that they$_{\{1,2\}}$ should invest in the stock market.
b. Every woman$_1$ told [her$_1$ husband]$_2$ that they$_{\{1,2\}}$ should invest in the stock market.
c. Every man$_1$ told [each of his$_1$ girlfriends]$_2$ that they$_{\{1,2\}}$ were going to get married.

Coindexing is not, however, expressively powerful enough to mark up details of anaphoric relations. The treatment of reciprocal pronouns is one of such cases.

(6) a. *They$_i$* love *each other$_i$*.
b. If *everyone$_i$* were to love *one another$_i$*, then *they$_i$* wouldn't want me to make a sacrifice, ...

More than mere coindexing is called for an adequate interpretation of the anaphoric phenomena shown here.

2.3. The TEI-based Annotation Schemes

There are two almost identical ASs for coreference: Bruneseaux and Romary (1997) and TEI P5 (2014). They are both based on XML and also on *the* TEI *Guidelines*. They differ from each other mainly because Bruneseaux and Romary (1997) followed a much earlier version of *the* TEI *Guidelines*, while TEI P5 (2014) is the most recent version, updated January 2014.

2.3.1. Bruneseaux and Romary (1997)

The AS, proposed by Bruneseaux and Romary (1997), consists of two (XML) elements: `<rs>` for referring strings and `<link>` for coreference. For each of the two elements, we can specify their associated attributes and possible values:[3]: (1) Attributes for the Element `<rs>`: `attributes = type, key; type = "object"; key = ID`; and (2) Attributes for the Element `<link>`: `attributes = type, targets; type = "coref"; targets = IDRef IDRef.`

Here is an example:[4]

(7)
```
<s>Kill <rs type="object" key="01">
an active, plump chicken</rs>.  Prepare
<rs type="object" key="02">it</rs>
for the oven, cut <rs type="object"
key="03"> it</rs> into <rs
type="object" key="04">four
pieces</rs> and roast <rs type="object"
key="05">it</rs> with thyme for I
hour.</s>
<link type="coref" targets="02 01"/>
<link type="coref" targets="03 02"/>
<link type="coref" targets="04 03"/>
<link type="coref" targets="05 04"/>
```

The chain of coreference relations may not preserve the original identity. Here the live chicken$_{01}$ was killed and cut into four pieces$_{04}$ and became roasted chicken$_{05}$.[5]

[3] As a specification language, we adopt ISO 14977 (1996) Extended BNF (Backus-Naur Form) that can be converted into its document type definition (DTD).

[4] Bruneseaux and Romary (1997) annotated a French version of the original English example (16) that occurs in Brown and Yule (1983), page 202.

[5] The original annotation of the French version given in Bruneseaux and Romary (1997) was modified here to suit the English example.

2.3.2. TEI P5 (2014)

TEI P5 (2014) (16.4.1) discusses correspondence between textual segments:

(8)
```
<title xml:id="SHIRLEY">Shirley</title>,
which made its Friday night debut
only a month ago, was not listed on
<name xml:id="NBC">NBC</name>'s new
schedule, although <seg corresp="#NBC"
xml:id="NETWORK">the network</seg>
says <seg corresp="#SHIRLEY"
xml:id="SHOW">the show</seg> still
is being considered.
```

In this textual fragment, the name *Shirley* is annotated as the title of a show being broadcast over NBC, a television network. The text contains no pronominal forms, but the two nominal forms, *the show* and *the network*, are understood as corresponding to the two names *Shirley* and NBC, respectively. The annotation of the two segments, namely those nominal forms, right above introduces the attribute `@corresp` to indicate such a coreferential relation for each of the two.
As shown below, the use of the elements `<linkGrp>` and `<link>` makes correspondence relations more explicit:

(9)
```
<linkGrp type="anaphoric_link"
targFunc="antecedent anaphor">
<link target="#Shirley #show"/>
<link target="#NBC #network"/>
</linkGrp>
```

The above annotation represents two instances of the anaphoric link involving anaphors and their antecedents. The attribute `@target` has two arguments, as specified by the attribute `@targFunc`: the first argument is `antecedent` and the second `anaphor`. The element `<linkGrp>` allows several instances of the anaphoric link to be grouped together, while simplifying the specification of the element `<link>` with a single attribute `@target`.

2.4. The MUC-7 Coreference Task Definition

Hirschman & Chinchor (1997)'s CDT lists four purposes of constructing an AS in the order of their importance. The first two are: (1) to support the MUC (Message Understanding Conference) information extraction tasks and (2) to be able to achieve good (ca. 90%) inter-annotator agreement. Creation of a corpus for research on coreference and discourse phenomena is the last goal.

2.4.1. Markables

Hirschman & Chinchor (1997)'s CDT restricts the set of its markables to nouns, that is, names, noun phrases, or pronouns[6]. Noun phrases include dates (*January 23*), currency expressions (*$1.2 billion*), percentages (*17 %*), and temperatures (*70 degrees*) that contain numerical values. Possessive, demonstrative, and reflexive pronouns are markables. So are the first person pronouns. The interrogative pronouns (*who, which engine*) are not markables.
In Hirschman & Chinchor (1997), verbs and other verbal forms such as gerunds (*Slowing the economy*) are not markables.[7] Implicit pronouns, that is, null anaphora (*Bill*$_i$ *called John and* e_i *spoke with him for an hour.*) and presumptive or intrusive pronouns (*the movie*$_i$ *which I saw* t_i)[8] as well as relative pronouns (complementizers) are not treated as markables.

2.4.2. Extents

The extent of a markable is a maximal string, while its head is marked with an attribute MIN (minimal string). The maximal noun phrases thus include their modifiers, appositional phrases, non-restrictive relative clauses, and prepositional phrases (*Fred Frosty, the ice cream king of Tyson's Corner*, MIN="`Fred Frosty`").

2.4.3. Coreference Links

Coreference in Hirschman & Chinchor (1997) is not restricted to referential identity. Here is the general principle for annotation coreference that they proposed:

(10) Two markables are coreferential if they both refer to sets, and the sets are identical, or they both refer to types, and the types are identical.

This principles thus allows the possible coreferentiality between bound anaphora and quantified NPs that are their antecedents.
Here are examples for the bound anaphoric relation that are treated in Hirschman & Chinchor (1997):[9]

(11) a. {Most computational linguists}$_i$ prefer their$_i$ own parsers.

b. {Every TV network}$_i$ reported its$_i$ profits yesterday. They$_i$ plan to release full quarterly statements tomorrow.

2.4.4. An SGML Serialization

Hirschman & Chinchor (1997) represents its coreference annotation in SGML. It introduces only one element <COREF> for the annotation of markables and also of their coreferential link type IDENT with the following specification of attribute-values:

(12) List of Attributes and Possible Values for <COREF>
ID = INTEGER;
MIN = CDATA; {* Head of the whole extent *}
REF = IDRef; {* Antecedent *}
TYPE = IDENT;
STATUS = OPT; {* if the reader is uncertain about the identity relation.*}

[6] All of the examples given in subsection 2.4 are copied from Hirschman & Chinchor (1997).

[7] The phrases *program trading, excessive trading, slowing of the economy* are noun-like, so they are treated as markables.

[8] See Cooper (1979), Evans (1977), Evans (1980), and Wechsler (2006) for detailed discussions of interpreting various uses of pronouns and Sells (1984) for presumptive and intrusive pronouns.

[9] Each extent is marked with a pair of stars (*) in Hirschman & Chinchor (1997), but these stars are replaced with curly brackets in this paper.

2.4.5. Illustrations

Here are some illustrations, copied from Hirschman & Chinchor (1997):

```
(13) <COREF ID="100">Lawson Mardon Group
     Ltd.</COREF> said <COREF ID="101"
     TYPE="IDENT" REF="100">it</COREF> ....
```

```
(14) Our <COREF ID="102" MIN="Board of
     Education">Board of Education</COREF>
     budget is just too high, the Mayor said
     <COREF ID="102" STATUS="OPT" TYPE="IDENT"
     REF="102">Livingstone Street</COREF> has
     lost control.
```

Here, the reader is uncertain about the IDENT relation of the Board of Education and Livingstone Street, although they are locally identical. That is why `STATUS="OPT"` is introduced into the annotation.

```
(15) <COREF ID="1" MIN="boys and girls">The
     sleepy boys and girls</COREF>
     enjoy <COREF ID="2" REF="1"
     TYPE="IDENT">their</COREF> breakfast.
```

Conjoined noun phrases are treated as one extent.

```
(16) <COREF ID="5">Fred</COREF> resigned
     as <COREF ID="6" MIN="president"
     REF="5">president of IBM</COREF>;
     next month, <COREF ID="7">the
     president</COREF> will be <COREF ID="8"
     REF="7">Mary</COREF>
```

The chain of coreference links `<ID="5", ID="6", ID="7">` is cut off by not specifying the coreference type `TYPE="IDENT"` in `<COREF ID="6"...REF="7" >` and `<COREF ID="8" REF="7">`.

Bound anaphors are treated here.

```
(17) <COREF ID="1" MIN="man">every man
     who knows<COREF ID="2" REF="1"
     TYPE="IDENT">his<COREF>own mind</COREF>
```

The entire string *every man who knows his own mind* and the pronoun *his* are annotated as coreferring with `IDENT`.

There are many other illustrations, for instance, for the annotation of apposition, predicate nominals and time-dependent identity, types and tokens, functions and values, and metonymy.

2.5. The MMAX2 Multi-level Annotation Scheme

Schäfer et al. (2012) uses Müller & Strube (2006)'s GUI-based MMAX2 annotation tool for coreference resolution to build a fully coreference-annotated large corpus of 266 scholarly papers from the ACL anthology. Here we briefly introduce the MMAX2 coreference AS.

2.5.1. Markables

Only proper names, noun phrases, and pronouns are markables, called possible entity 'mentions'. There are 8 mention types: (1) def-np (definite NPs), (2) pper (personal pronouns), (3) ne (proper names including citations), (4) ppos (possessive pronouns/determiners), (5) indef-np (indefinite NPs), (6) conj-np (coordinations), (7) pds (demonstrative pronouns), and (8) preflexive (reflexive pronouns).

Unlike Hirschman & Chinchor (1997), this AS treats relative pronouns (*who, which, whose, that, ...*) as markables and excludes bound anaphora ($\{$*Every teacher*$\}_i$ *likes his*$_i$ *job.*) as well as predicative nominals ($\{$*A mason*$\}_i$ *is* $\{$*a workman*$\}_i$.).

2.5.2. Anaphoric Links

Following van Deemter & Kibble (2000) and their two other related works, Kibble & van Deemter (1999) and Kibble & van Deemter (2000), this AS differentiates coreference from other types of the anaphoric link. It suggests that the annotation of coreference proper be separated from other tasks such as annotation of bound anaphors and of the relation between a subject and a predicative NP. It calls for a division of labor that achieves better inter-annotator agreement.

2.6. The Brandeis ISO-Space Annotation Guidelines

The ISO-Space Working Group at Brandeis University produced a manual of annotation guidelines for spatial information (Pustejovsky et al., 2013) and proposed it as an annex of ISO-Space (2013), an ISO international standard for spatial annotation. First, the 2013 version of this annex annotates spatial entities (*I am sitting in the* **car**$_{se}$.)[10] as referring expressions. Second, it introduces an element `<metaLink>` to annotate the three different types of coreference between these spatial entities: (1) coreference, (2) subcoreference, and (3) split coreference. Third, it specifies a set of attributes and their possible values for the element `<metaLink>`, as shown below:

(18) Attributes and Possible Values for `<metaLink>`

```
     attributes = id, [objectID1],
     [objectID2], relType, [comment];
     {* the attributes in square brackets
     are implied.*}
     id = "meta"INTEGER;
     objectID1 = IDRef;
     objectID2 = IDRef;
     relType = "coref", "subCoref",
     "splitCoref";
     comment = CDATA;
```

Instead of using XML as its representation language, Pustejovsky et al. (2013) adopts the predicate-logic-like forms to represent the three different types of coreference. Here are examples:

(19) a. $\{$Two cars$\}_{se1}$ are on the street. One$_{se2}$ of them$_{se3}$ turns left.

[10] A spatial entity is introduced into ISO-Space (2013) as an element of a type of entity that participates in location-involving motions or non-motion events. Its identifier is marked with a prefix `se`.

b.
```
spatialEntity(id=se1, extent=two
cars, countable=yes)
spatialEntity(id=se2, extent=one,
countable=yes)
spatialEntity(id=se3, extent=them,
countable=yes)
metaLink(id=meta1, objetID1=se1,
objectID2=se3, relType=coref)
metaLink(id=meta2, objetID1=se1,
objectID2=se2, relType=subCoref)
```

The above annotation can be interpreted as stating that the referents of two spatial entities *cars*$_{se1}$ and *them*$_{se3}$ are identical, while *one*$_{se2}$ partially corefers with the spatial entity type expressions *cars*$_{se1}$.
Here is another example:

(20) a. {John$_{se6}$ and Mary$_{se7}$}$_{se8}$ met at the store. They$_{se9}$ went shopping.

b.
```
metaLink(id=meta4 objectID1=se8
objecID2=se9 relType=coref)
metaLink(id=meta3 objectID1=se7
objectID2=se8 relType=subCoref)
metaLink(id=meta4 objectID1=se9
objectID2=se6 relType=splitCoref)
metaLink(id=meta5 objectID1=se9
objectID2=se7 relType=splitCoref)
```

Here, the names *John*$_{se6}$ and *Mary*$_{se7}$ are each treated as a referring expression. At the same time, the whole phrase {*John*$_{se6}$ *and Mary*$_{se7}$}$_{se8}$ as a group is also treated as a referring expression. And then the antecedent of the plural pronoun *They*$_{se9}$ is split into two: *John*$_{se6}$ and *Mary*$_{se7}$.

3. Specification of the Proposed Integrated Annotation Scheme

3.1. Overview

Given an input text, the task of coreference or other anaphoric link annotation is three-fold: (1) identification of a set of mentions in the text that refer to something in the domain of discourse referents as its *markables*, (2a) identification of a set of anaphor-antecedent pairs of the mentions that are anaphorically related and (2b) specification of the type of such a relation. To trigger and constrain these annotation steps, the entity type of mentions and the anaphoric relation are assigned a set of required or implied attribute-value specifications. (2a) and (2b) constitute two sub-processes unified into one, for they depend on each other. The input text can be of any size. It can range from a short sentence to a very large corpus.

3.2. Identifying Markables and Extents

The set of possible markables consists of terms or mentions, which comprise both referring and non-referring expressions in a text. As attested quantitatively by various reference resolution experiments such as Chen & Hacioglu (2006), Haghighi & Klein (2010), Stoyanov et al. (2010), and Raghunathan et al. (2010), these mentions are mostly noun phrases of the following four forms: (1) proper names, (2) definite or indefinite nominals with *plurality* and other agreement specifications, (3) (generalized) universal or existential quantifiers, and (4) definite or indefinite pronouns with *gender* and *number* specifications, which are subclassified into: personal pronouns, reflexives, reciprocals, and demonstratives.[11]
The list of these features just specifies what morphosyntactic features are required or implied for the identification of mentions. The annotation of these features could be done at earlier stages of annotating raw data such as tokenization and morphosyntactic annotation. The process, whether manual or automatic, of marking up these mentions as markables should be straightforward at this basic level.

3.3. Anaphoric Links

The main task of annotating coreference and other types of the anaphoric link is to recognize antecedent-anaphor pairs among the set of markables and also to identify the type of their anaphoric link.

3.3.1. Anaphor-antecedent Pairs

Anaphors are part of the set of mentions, being mostly pronouns and other pronominal forms (see Keenan (1993a)). They are thus easily identified.

(21) a. Bob loves Jane, but *she* doesn't love *him*.

b. Bob was tired, and *so* was I.

Some definite noun phrases can be anaphors, too. Here are some examples:

(22) a. {The project leader}$_i$ is refusing to help. {The jerk}$_i$ thinks only of himself.

b. {Hilary Clinton}$_i$, {Bill's wife}$_i$.

Among the list of pronouns, we may also include the use of *it* referring to propositions, facts, actions, etc., or the use of *so* that may involve so-called sloppy identities, as shown below:

(23) a. John said {he has been to heaven}$_i$, but I don't believe {it}$_i$.

b. John {loves his wife}$_i$ and { does}$_{i?}$ Bob.

Examples such as these are often discussed in linguistic literature, but have been seldom treated in computational work.
The so-called expletive *it* and *there*, the complementizer *that*, and the impersonal use of the pronoun *it*, as shown below, are excluded from the list of possible anaphors as well as from the list of possible markables.

(24) *It*$_{exp}$'s impossible to go out now, for *it*$_{imp}$'s raining cats and dogs. *It*$_{exp}$ is also reported *that*$_{comp}$ *there*$_{exp}$ is a storm approaching from the south.

The identification of anaphors as well as mentions can also be triggered by the morphosyntactic features of markables.

[11] Interrogatives are excluded.

3.3.2. Types of the Anaphoric Link

Unlike anaphors, antecedents can be of any class of a word, phrase, or clause. It should, however, be a subset of markables as specified by the first step of annotation. If verbal forms are excluded from the set of markables for some practical reasons, then they would not be in the set of possible anaphors or antecedents.

The extent of antecedents is not restricted to a single word or phrase, but may extend to larger phrases such as conjoined phrases:

(25) {The boys$_i$ and the girls$_j$}$_k$ met at a party and they$_k$ danced all night.

Antecedents may not be contiguous, either, but split into two or more phrases, as in:

(26) I$_i$ met {a farmer}$_j$ and {his$_j$ dog}$_k$ and we$_{\{i,j,k\}}$ all walked together.

There are at least two uses of pronouns:

(27) a. anaphoric: *John$_i$ loves his$_i$ wife.*

b. indexical or deictic: *Look at him$_1$. He$_1$ is naked.* Context: the speaker pointing to a person over there.

In the anaphoric use, the pronoun his_i finds its antecedent $John_i$ in the given text. In the indexical use, the antecedent of the pronoun him_1 is not found in the text, but provided contextually.

Pronouns can be antecedents as well as anaphors in the chain of an anaphoric link.

(28) John$_{i_1}$ loves {his$_{i_2}$ wife}$_j$ and she$_j$ also loves him$_{i_3}$.

Anaphoric links may be forward or backward. The term that corefers with a pronoun normally precedes it, thus being called ‘antecedent’. This so-called antecedent may also come after its related anaphor, as in:

(29) When she_i returned home, Sue_i was surprised to find her dog gone.

In such a case, the pronoun is often called ‘cataphor’.

Sometimes it is difficult to decide which is an anaphor and which is its antecedent, as especially in appositive cases (*Seoul$_i$*, {*the capital of South Korea*}$_j$, where i and j corefer.). In such cases, we simply have to state that they correspond to or corefer with each other.

The antecedent-anaphor relation is normally a one-to-one relation, but there are cases in which the antecedent of an anaphor is split into many. Besides this case of split coreference, Pustejovsky et al. (2013) lists `subCoreference` as another type of coreference:

(30) I have {two cars}$_i$, but one$_j$ of them$_i$ broke down.

Here *one$_j$* is a member of the set of *two cars$_i$*.

3.4. Formal Description

Bunt (2010) provides a formal description of the annotation structure, consisting of two levels of syntax: one is an abstract level of an annotation, called ‘abstract syntax’, and another, a concrete level of representing annotations, called ‘concrete syntax’. Every abstract syntax for semantic annotations must be supported by an explicit (formal) semantics. An XML-serialization of an abstract syntax is an instance of a concrete syntax. The semantics of a concrete syntax is defined as the semantics of the abstract syntax for which it defines a concrete representation. (Different representations of the same abstract syntax thus have the same semantics.)

3.4.1. Abstract Syntax

The abstract syntax of an annotation scheme consists of two parts: (1) a conceptual inventory, that specifies the basic concepts from which annotation structures are built up; (2) a specification of the possible ways of combining elements of the conceptual inventory into annotation structures. An annotation structure is a set consisting of two kinds of elements: *entity structures* and *link structures*. Entity structures provide linguistic information about a region of primary data; link structures provide information about the semantic relation between regions of primary data. In the case of annotating coreference and other anaphoric link types, entity structures correspond to the entities that are related by anaphoric links, and link structures to the linkings of anaphoric expressions to their antecedents.

An entity structure is a pair $\langle m, a\rangle$ where m is a markable that identifies a region of primary data, and a is the specification of the semantic information that the annotation provides about that region of primary data. In the annotation scheme $\mathcal{AS}_{ana}$ for coreference and other anaphoric link types the a component of an entity structure is an n-tuple, $3 \leq n \leq 6$ consisting maximally of a semantic type t, a definiteness d, a morphosyntactic form f, a natural gender g, a plurality p, and a collectiveness c (more about these elements below). The fact that the length n of these n-tuples may vary, reflects the optionality of some of the elements.

A link structure is a triplet $\langle \epsilon_1, \epsilon_2, r\rangle$ consisting of two entity structures (for anaphor and antecedent) and a relation corresponding to the type of anaphoric link between them.

For the abstract syntax of the annotation scheme $\mathcal{AS}_{ana}$ the conceptual inventory is a 9-tuple $\langle M, T, D, F, G, P, C, Q, R\rangle$, where (1) M is a non-empty set of markables, (2) T is a set of semantic types; (3) D is a set of definiteness values; (4) F is a set of morphosyntactic forms; (5) G is a set of natural genders; (6) P is a set of singular/plural values; (7) C is a set of ‘collectivity values’; (8) Q is a set of generalized quantifiers and (9) R is a set of binary relations over the set of entity structures, corresponding to the various types of anaphoric links. The annotation structures are defined by an assignment @ that specifies the semantic components of entity structures. For each markable m in M, @(m) generates an n-tuple, $3 \leq n \leq 7$, of elements from $T \times D \times F \times F \times P \times C \times Q$. [12]

[12] The specification of morphosyntactic forms, and several of

3.4.2. Concrete Syntax

Here is an XML-based concrete syntax $\mathcal{AS}_{anX}$, corresponding to the abstract syntax of the proposed annotation scheme $\mathcal{AS}_{ana}$. First, it introduces two elements `<entity>` and `<anaLink>` that correspond to entity structures and link structures, respectively, as defined in $\mathcal{AS}_{ana}$. Both of these XML elements have an `@identifier` attribute in order to allow references from within the representation of a certain link structure to the representations of specific entity structures or other link structures. Moreover, `<entity>` structures have a `@target` attribute for representing the markables that they associate linguistic information with. Second, the assignment @ for M can be transduced into $\mathcal{AS}_{anX}$ as below:

(31) Attributes and Values for the element `<entity>`

```
attributes = identifier, target,
type, [def], form, [naturalGender],
[plurality], [collectivity], [quant],
[comment];[13]
identifier = "entINTEGER";[14]
target = IDRef, CDATA;[15]
type = "person", "organization", CDATA;
def= "yes", "no";
form = "name", "nom", "pro";[16]
naturalGender = "male", "female";[17]
plurality = "yes";
collectiveness = "yes";
quant = CDATA;
comment = CDATA;
```

By allowing the value `quant` as a possible value for the attribute `form`, $\mathcal{AS}_{anX}$ treats the anaphoric link between a quantifier and a bound anaphor (e.g., pronoun). It does not treat reciprocals (e.g., *each other, one another*).

(32) Attributes and Values for the element `<anaLink>`

```
attributes = identifier, anaphor,
antecedent, type, [comment];
identifier = "anaINTEGER";
anaphor = IDRef;
{* By an XML convention, IDRef's are
prefixed with a star.  *}
antecedent = IDRef*;
{* This allows multiple antecedents.
The indexical use of a pronoun may
not have an antecedent in the element
<entity>.  *}
type = "ident", "partIdent",
"setIdent", "qBound";
comment = CDATA;
```

The attribute `@type` introduces values other than `ident` for referential identity. These values allow the types of the anaphoric link other than the type of coreference proper. The use of each of the values of the attribute `@type` is illustrated below:

(33) a. `ident`: referential individual-level identity;
John$_1$ *loves Jane*$_2$*, but she*$_2$ *dislikes him*$_1$*.*

b. `partIdent`: referential partial identity;
John owns $\{$*two cars*$\}_i$*. One*$_{i1}$ *of them*$_i$ *broke down.*

c. `setIdent`: set or group-level identity between an anaphor and its identity;
$\{$*Every farmer*$\}_1$ *owns a donkey. They*$_1$ *beat it.*
$\{$*The whole army*$\}_2$ *surrendered themselves*$_2$*.*

d. `qBound`: case of bound anaphors;
Every$_x$ *farmer loves his*$_x$ *wife.*

The set-level anaphoric identity assumes that the denotation of an anaphor is a set and also that that set is also the denotation of its antecedent so that they are identical as sets. For example, the denotation $[|every\ farmer|]^M$ of *every farmer* with respect to a model M is understood to be a set $\{X|\ [|farmer|]^M \subseteq X\}$ of supersets of the set of farmers.

We introduce `<isoAna>` as the root element for XML documents in the concrete XML annotation scheme $\mathcal{AS}_{anX}$ for coreference and other types of anaphoric link.

4. Illustrations

Here is a segment of the news item given earlier. The proposed $\mathcal{AS}_{anX}$ can annotate it, as shown below:

(34) a. $\{$Graham Jones$\}_{ent13}$, 44, told $\{$his$_{ent14}$ wife$_{ent15}$ $\}_{ent16}$ that she$_{ent17}$ would receive $\{$eight blows$\}_{ent18}$ - one$_{ent19}$ for $\{$each day$\}_{ent20}$ . . . .

b. Step Two: Identification of Possible Anaphors

```
<isoAna xml:id="ana1">
<entity xml:id="ent13"
target="Graham Jones"
form="name"type="person" />
<entity xml:id="ent14" target="his"
form="pro "type="person" >
<entity xml:id="ent16" target="his
wife" form="nom" type="person"
def="yes" naturalGender="female"/>
<entity xml:id="ent17" target="she"
form="pro" type="person"
naturalGender="female"/>
<entity xml:id="ent18"
target="eight blows" form="quant"
plurality="yes"/>
<entity xml:id="ent19" target="one"
form="pro" def="no"/>
<anaLink xml:id="ana01"
anaphor="#ent14" antecedent="#ent13"
type="ident"/>
<anaLink xml:id="ana03"
anaphor="#ent17" antecedent="#ent16"
type="ident"/>
```

the other elements may refer to other levels of annotation.

[13] Attributes in square brackets are optional or implied.

[14] The identifier is tagged `xml:id` for XML documents, otherwise `id`. Examples are: `"ent3", "ent20"`.

[15] The attribute `@target` has an extent ID in a tokenized source text or the extent itself as its value. This value can be a (possibly null or non-contiguous) sequence of tokens or their IDs.

[16] Verbal forms including sentential or adjectival forms are excluded.

[17] Optional attributes have a value 'unspecified' as default.

```
<anaLink xml:id="ana05"
anaphor="#ent19" antecedent="#ent18"
type="partIdent"/>
</isoAna>
```

The indefinite pronoun *one*$_{ent19}$ is treated as partially identical with the quantified nominal $\{$*eight blows*$\}_{ent18}$, thus being annotated `type ="partIdent"` in `<anaLink xml:id= "ana05">`.

Here are well-known donkey sentences:

(35) a. Every farmer who owns a donkey beats it.[18]
b. If Pedro owns a donkey, he beats it.

Example (b) can be annotated as below:

(36) a. If Pedro$_{ent1}$ owns $\{$a donkey$\}_{ent2}$, he$_{ent3}$ beats it$_{ent4}$.

b.
```
<isoAna xml:id="ana2">
<entity xml:id="ent1" target="Pedro"
form="name" type="person"/>
<entity xml:id="ent2" target="a
donkey" form="nom" type="animal"
def="no"/>
<entity xml:id="ent3" target="he"
form="pro" type="person"/>
<entity xml:id="ent4" target="it"
form="pro" type="person" def="yes"/>
<anaLink xml:id="ana01"
anaphor="#ent3" antecedent="#ent1"
type="ident"/>
<anaLink xml:id="ana02"
anaphor="#ent4" antecedent="#ent2"
type="ident"/>
</isoAna>
```

Here both of the anaphor-antecedent pairs, <*he*, *Pedro*> and <*it*, *a donkey*> are treated as coreferring. Note that indefinite descriptions are treated as referential terms, not existential quantifiers (see Kamp (1981).)

5. Semantic Interpretations

As stated earlier, every semantic annotation must be accompanied by an explicitly defined semantics. The use of the lambda calculus in the line of Montague (1974) or that of the discourse representation structures (DRSs), proposed by Kamp (1981) and Kamp & Ryele (1993), can, for instance, be linked to the abstract syntax to provide such a semantics for semantic annotations. Attempts have been made by Katz (2007), Pratt-Hartman (2007), Bunt (2007), Bunt & Overbeeke (2008a), Bunt & Overbeeke (2008b), Lee (2008) to develop an annotation-based semantics with the use of lambda calculus or by Bunt (2010) with the use of DRSs . The use of lambda abstraction has run into the problem of complexity especially in dealing with multiple quantification and embedded adjunct structures. This should be the case with the treatment of various anaphoric phenomena. There are at least two interesting works to overcome this complexity problem: One is an earlier work by Muskens (1996) which proposed a way of combining Montague semantics with DRSs and another is the most recent work by Bunt (2014) which directly addresses to the treatment of anaphoric phenomena by combining underspecified representation (USR) that arises because of the presence of context-dependent expressions such as pronouns with representation of annotation information (AIR). In constructing these representation structure, Bunt (2014) shows how useful and necessary it is to combine the introduction of discourse referents in DRSs with markables in the annotation into USR and AIR, especially when there are multiple occurrences of identical anaphoric expressions, that is, pronouns, in a text.

We leave detailed discussion of ways of interpreting anaphoric links as a work item for the future. Here are some remarks on the interpretation of various anaphoric expressions. First, names and definite descriptions are referential terms, both referring to some unique entities in the domain of discourse referents. Indefinite descriptions are also treated as referential terms, as mentioned earlier.

Second, as proposed and discussed in formal model-theoretic semantics (see Montague (1974), Barwise and Cooper (1981), Link (1987), and Keenan & Westerst°ahl (2010)), proper names, definite descriptions, indefinite singular (*a dog*) or bare plural (*donkeys*) noun phrases as well as quantified noun phrases (*three students, every man*) are also interpreted as referring to sets of sets or properties, in the world. In our treatment, universally quantified expressions are differentiated from other types of generalized, but existentially quantified expressions.

Third, pronouns, on the other hand, do not refer directly to any entities in the world, but only through being coreferential with some other terms in the text (anaphoric use) or by referring to some entities that are provided contextually in a discourse situation (indexical use) (see Keenan (2007)). Nevertheless, pronouns are also marked up as referring expressions or mentions in coreference annotation (see Cooper (1979), Evans (1977), Evans (1980).)

6. Concluding Remarks

The purpose of this paper has been to integrate several existing ASs for anaphoric links into a unified $\mathcal{AS}_{ana}$ that may be accepted as an ISO standard for language resources management. One big issue in designing $\mathcal{AS}_{ana}$ is a choice between theoretical granularity and practical sustainability. If an AS is theoretically fine-grained, then the range of its applications may be wider, provided that various conditions of its sustainability are guaranteed such as the ease of its use with a high score of inter-annotator agreement and the cost-effectiveness of developing language resources through its use, as mentioned in Hirschman & Chinchor (1997).

The aspect of granularity here mainly concerns (morphosyntactic) forms of anaphors and types of anaphoric link. The pronoun *his* in Examples (3a,b), for instance, is treated as a typical anaphor with a reasonable claim that every pronominal form is an anaphor. The name *John* and the quantifier *Everyone* are easily recognized as their respective antecedents. A question now is whether their anaphoric links are of the same type or not. The proposed $\mathcal{AS}_{ana}$ treats them both as instances of the anaphoric link, but of

[18] Originally, from Geach (1962).

different types, without elevating the notion of coreference from the level of referential identity to that of set-identity or type-identity as in Hirschman & Chinchor (1997)'s MUC-7 CDT. While preserving the classical definition of coreference as referential identity, $\mathcal{AS}_{ana}$ can easily modify its scheme in the line of van Deemter & Kibble (2000) with a division of labor.

7. Acknowledgments

Partially supported by the ID R&D program of MSIP/KEIT No. 10044457 "Development of Autonomous Intelligent Collaboration Framework for Knowledge Bases and Smart Devices". I owe thanks to Harry Bunt for his great suggestions and also to Roland H. Hausser, Hwan-Mook Lee, and Do-sam Hwang for reading the pre-final version of this paper with encouraging comments.

8. References

Muskens, Reinhard. 1996. Combining Montague semantics and discourse representation. *Linguistics and Philosophy* 19: 143-186.

Barwise, Jon, and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4: 159219.

Brown, Gillian, and George Yule. 1983. *Discourse Analysis*. Cambridge University Press, Cambridge.

Bruneseaux, Florence, and Laurent Romary. 1997. Codage des références et coréférences dan les DHM. *ACHALLC'97*, 169-173.

Bunt, Harry. 2007. The semantics of semantic annotation. In *Proceedings of the 21st Pacific Asia Conference on Language, Information, and Computation (PACLIC-21), pp. 13-29. Korean Society for Language and Information, Seoul.*

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

Bunt, Harry. 2011. Introducing abstract syntax + semantics in semantic annotation, and its consequences for the annotation of time and events. pp. 157-205. In Eunryoung Lee and Aesun Yoon (eds.), Recent Trends in Language and Knowledge Processing. *Hankookmunhwasa, Seoul.*

Bunt, Harry. 2014. Annotations that effectively contribute to semantic interpretation. In Harry Bunt, Johan Bos and Stephen Pulman (eds.), Computing Meaning*, vol. 4, 49-69. Springer, Berlin.*

Bunt, Harry, and C. Overbeeke. 2008a. An extensible, compositional semantics of temporal annotation. In Proceedings of LAW-II: The Second Linguistic Annotation Workshop, *LREC-2008. Marrakech.*

Bunt, Harry, and C. Overbeeke. 2008b. Towards formal interpretation of sematnic annotation. In Proceedings of the 6th Edition of LREC (Language Resources and Evaluation Conference) 2008. Marrakech.

Burnard, Lou and Syd Bauman (eds.). 2014. TEI P5: *Guidelines for Electronic Text Encoding and Interchange*, Version2.6.0. Last updated on 20th January 2014, revision 12802. Text Encoding Initiative Consortium, Charlollotesville, VA.

Chen, Ying, and Kadri Hacioglu. 2006. Exploration of coreference resolution: The ACE entity detection and recognition task. In Petr Sojka, Ivan Kope[č]ek, and Karel Pala (eds.), *Proceedings of the 9th International Conference, TSD 2006*, LNAI 4188, pp.301-308.

Cooper, Robin. 1979. The interpretation of pronouns. In Frank Heny and Helmut S. Schnelle (eds.), *Syntax and Semantics 10: Selections from the 3rd Groningen Round Table*, 61-92. Academic Press, New York.

Evans, Gareth. 1977. Pronouns, quantifiers, and relative clauses (I). *The Canadian Journal of Philosophy*, 7.3, 467-536.

Evans, Gareth. 1980. Pronouns. *Linguistic Inquiry*, 6, 353375.

Geach, Peter. 1962. *Reference and Generality*. Cornell University Press, Ithaca, N.Y.

Haghighi, Aria, and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic Features. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 11521161. Singapore, 6-7 August 2009. ACL and AFNLP 2009.

Haghighi, Aria, and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proceedings of HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 385-393.

Hirschman, Lynette, and Nancy Chinchor. 1997. MUC(Message Understanding Conference)-7 coreference task definition, version 3.0. Updated 13 July 1997.

ISO/IEC JTC 1 Information Technology. 1996. *ISO/IEC 14977:1996(E), Information technology - Syntactic metalanguage - Extended BNF.*

ISO/TC 37/SC 4/WG 2. 2013. *ISO CD 24617-7 Language resource management - Semantic annotation framework - Part 7: Spatial information (ISO-Space).* The International Organization for Standardization, Geneva.

Kamp, Hans. 1981. A theory of truth and semantic representation. In *Formal Methods in the Study of Language*, part 1, 277-322. MC Tract 135. Stichting Mathematisch Centrum, Amsterdam.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.

Karttunen, Lauri 1969. Pronouns and variables. In Robert Binnick, Alice Davidson, Georgia Green, and Jerry Morgan (eds.), *Proceedings from the fifth regional meeting of the Chicago Linguistic Society*, pp. 108-116. University of Chicago Department of Linguistics.

Katz, Graham. 2007. Towards a denotational semantics for TimeML. In F. Schilder, Graham Katz, and James Pustejovsky (eds.), *Annotation, Extraction, and Reasoning about Time and Events*, 88-106. Springer, Dordrecht.

Keenan, Edward. 1993a. Identifying anaphors. In J Guenter, B. Kaiser, and C. Zoll (eds), *Proceedings of*

BLS 19, 503-516. Berkeley Linguistics Society, UC Berkeley.

Keenan, Edward. 1993b. Anaphor-antecedent asymmetry. In Utpal Lahiri and Zachary Wyner (eds.), *Proc. of Semantics and Linguistic Theory*, III: 117-144. Dept. of Modern Languages and Linguistics, Cornell University.

Keenan, Edward. 2007. On the denotations of anaphors. *Research on Language and Computation* 5.1:5-17. Formerly C54.

Keenan, Edward, and Dag Westerst°ahl. 2010. Generalized quantifiers in linguistics and logic, In Johan van Benthem and Alice ter Meulen (eds.), *Handbook of Logic and Linguistics*, 2nd rev. ed., 859-910. Springer, Berlin.

Kibble, Rodger, and Kees van Deemter. 1999. What is coreference, and what should coreference annotation be? *Proceedings of ACL workshop on Coreference and its applications*. University of Maryland, June 1999.

Kibble, Rodger, & Kees van Deemter. 2000. Coreference annotation: Whither? *Proceedings of LREC-2000*, pages 90-96. Athens, Greece.

Lee, Kiyong. 2008. Formal semantics for interpreting temporal annotation. In Piet van Sterkenburg (ed.), *Unity and Diversity of Languages: Special Lectures for the 18th International Conference of Linguists*, pp. 97-108. Benjamins, Ambsterdam.

Link, Godehard. 1987. Generalized quantifiers and plurals. In P. Gärdenfors (ed.), *Generalized Quantifiers: Linguistic and Logical Approaches*, p. 151-180. Reidel, Dordrecht.

Montague, Richard. 1974. The proper treatment of quantification in ordinary English. In Richmond Thomason (ed.), *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven.

Müller, Christoph, and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Murkherjee (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197-214. Peter Lang, Frankfurt a.M..

Pratt-Hartman, Ian. From TimeML to Interval Temporal Logic. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)*, pp.166-180. Tilburg, Netherlands.

Pustejovsky, James, Jessica Moszkowicz, and Zachary Yocum 2013. *ISO-Space Annotation Guidelines*, Version 1.7.0, (April 2013). The ISO-Space Working Group, Brandeis University.

Raghunathan, K., H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. 2010. A multi-pass sieve for coreference resolution. In EMNLP.

Rahman, Altaf, and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 814-824.

Ratinov, Lev-Arie, and Dan Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 12341244. Jeju Island, Korea, 1214 July 2012, Association for Computational Linguistics.

Rullmann, Hotze. 2003. Bound-variable pronouns and the semantics of number. In Brian Agbayani, Paivi Koskinen, and Vida Samiian (eds.), *Proceedings of the Western Conference on Linguistics: WECOL 2002*. Department of Linguistics, California State University, Fresno, pp. 243-254.

Schäfer, Ulrich, Christian Spurk, and Jörg Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the ACL Anthology. *Proceedings of COLING 2012: Posters*, pages 10591070.

Sell, Peter. 1984. *Syntax and semantics of resumptive pronouns*. Doctoral thesis, University of Massachusetts, Amherst.

Stoyanov, Veselin, Claire Cardie, Nathan Gilber, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with Reconcile. *Proceedings of the ACL 2010 Short Papers*, pages 156-161.

Stoyanov, Veselin, and Jason Eisner. 2012. Easy-first coreference resolution. *Proceedings of COLING 2012: Technical Papers*, pages 2519-2534, COLING 2012, Mumbai, December 2012.

Soon, Wee Meng, Daniel Chung Yong Lim, and Hwee Tou Ng. 2001. A machine learning approach to coreference resolution of noun phrases.

van Deemter, Kees, and Rodger Kibble. 2000. On Coreferring: Coreference annotation in MUC and related schemes. *Computational Linguistics* 26.4, pp. 615-623.

Wechsler, Stephen. 2006. Why are the lazy so agreeable? In Hans-Martin Gärtner, Sigrid Beck, Regine Eckardt, Renate Musan & Barbara Stiebels (eds.), *Between 40 and 60 Puzzles for Krifka*. Centre for General Linguistics, Typology and Universals Research (ZAS), Berlin.

Towards Extending the ISOcat Data Category Registry with Zulu Morphosyntax

Laurette Pretorius, Sonja Bosch
University of South Africa
PO Box 392, UNISA, Pretoria, South Africa 0003
pretol@unisa.ac.za, boschse@unisa.ac.za

Abstract

The importance of the semantic annotation of morphological data for agglutinating languages is the departure point of this paper. It discusses the principled extension of the ISOcat data category registry (DCR) to include Zulu morphosyntactic data categories. The focus is on the Zulu noun. Where existing data categories are found appropriate they are used and where new additions are required the published guidelines are followed. The expectation is that these extensions will also be useful for languages that are related to Zulu and share its morphosyntactic structure. The inclusion of the other Zulu word categories forms part of future work.

Keywords: ISOcat, data categories, Zulu, morphosyntax, semantic interoperability

1. Introduction

Our point of departure is the increasing emphasis on the semantic interoperability of linguistic data, and more specifically, morphological data. In languages with agglutinative morphologies much information, both syntactic and semantic, is encoded in the morphology of words. The accurate annotation of this information is often key to the reliability of language processing technologies, tools and applications for these languages and for their interoperability with other languages and, for example, in the Semantic Web.

Interoperability can be defined as a measure of the degree to which diverse systems or language resources are able to work or be used together to achieve a common goal (Ide and Pustejovsky, 2010). Interoperability is typically defined in terms of syntactic and semantic interoperability.

“Whereas syntactic interoperability provides for the exchange of clearly defined classes of data, semantic interoperability enables the automatic recognition of the individual data exchanged.” Also in linguistics, “syntax refers to the grammar and formal rules for defining sets of data, while semantics define the meaning and the use of these data. In other words, on the semantic layer data becomes information” (Kubicek et al., 2011). This is also true for morphological data.

In the project, reported on in this paper, we follow a principled approach to the creation of data categories to facilitate the syntactic and semantic interoperability of morphological information of the Bantu language family. For this purpose we use the ISOcat Data Category Registry[1] (DCR), based on the ISO 12620 standard. Indeed, ISOcat may be considered a key resource and a *de facto* standard in its own right. In terms of language we focus on Zulu, an agglutinative Bantu language spoken in Southern Africa.

As a first step towards future syntactic and semantic annotation of Zulu morphological data, using ISOcat, it is necessary to ensure that all the relevant linguistic concepts, referred to as data categories, occurring in Zulu morphology are available in ISOcat. This paper therefore reports on a first attempt to extend the ISOcat DCR for Zulu by restricting our attention to the Zulu noun and its morphological structure.

The structure of the paper is as follows: Section 2 briefly introduces ISOcat and shows how it facilitates syntactic and semantic annotation and interoperability of linguistic data. Since our focus is on the Zulu noun, section 3 provides a short exposition of Zulu noun morphology. Section 4 represents the core contribution. It discusses the proposed Zulu morphosyntax extension to the ISOcat DCR. Section 5 concludes the paper by sharing acquired insights, discussing further extensions to the ISOcat Zulu morphosyntax and identifying future work, also for other related languages.

2. ISOcat DCR and interoperability

The main aim of the ISOcat DCR is to define widely accepted linguistic concepts in a stable and persistent way. Each concept is assigned a so-called persistent identifier (PID) in the form of a cool Uniform Resource Identifier (URI). It provides a “framework for defining data categories compliant with the ISO/IEC 11179 family of standards. According to this model, each data category is assigned a unique administrative identifier, together with information on the status or decision-making process associated with the data category. In addition, data category specifications in the DCR contain linguistic descriptions, such as data category definitions, statements of associated value domains, and examples. Data category specifications can be associated with a variety of data element names and with language-specific versions of definitions, names, value domains and other attributes.” (ISOcat, nd). Although data categories are stored as a flat list, there is the option of creating customized registry instances for specific subdisciplines of interest. This is achieved through the so-called Data Category Selections (DCSs).

Interoperability is achieved when users annotate their data with references to registered concepts, thereby allowing others to interpret their data. Interoperability is further enhanced by ensuring that there is the minimum of duplication in the registry. The ISOcat DCR makes it possible to

[1] ISOcat data category registry, `http://www.isocat.org/`

work across projects, disciplines and languages by providing a mechanism to make the semantics of different tag sets explicit through referencing of registered ISOcat concepts.

3. Morphology of the Zulu noun

The morphological structure of the Zulu noun is characterised by a nominal classification system that categorises nouns into a number of noun classes, as determined by prefixal morphemes also known as noun prefixes. These noun prefixes have, for ease of analysis, been divided into classes with numbers by scholars who have worked within the field of the Bantu language family. Table 1 shows examples of Meinhof's (Meinhof, 1932, 48) numbering system of some of the noun class prefixes:

Prefix	Class	Word form	English
umu-	1	*umuntu*	"person"
aba-	2	*abantu*	"persons"
u-	1a	*unozinti*	"goalkeeper"
o-	2a	*onozinti*	"goalkeepers"
umu-	3	*umuzi*	"homestead"
imi-	4	*imizi*	"homesteads"
i(li)-	5	*idolo*	"knee"
ama-	6	*amadolo*	"knees"
u(lu)-	11	*ukhezo*	"wooden spoon"
izin-	10	*izinkezo*	"wooden spoons"
ubu-	14	*ubusuku*	"night"

Table 1: Meinhof's (1932:48) numbering system of noun class prefixes

Noun prefixes usually indicate number, with the uneven class numbers designating singular and the corresponding even class numbers designating plural. However, this is not always the case, since some nouns in so-called plural classes do not have a singular form; plurals of class 11 nouns are found in class 10, while a class such as 14 is not associated with number at all. The noun prefix typically constitutes two parts, namely a preprefix (the initial vowel) and a basic prefix, but in some classes such as 1a and its plural class 2a a basic prefix does not feature. In other instances such as classes 11 and 14 the basic prefixes are often discarded, with the result that only the preprefix appears in the surface form.

Other morphemes that may be suffixed to the noun in Zulu are the diminutive, augmentative and feminine, or combinations thereof:

- Diminutive nouns are usually formed by the suffixation of a diminutive suffix *-ana* to a noun, e.g.
  isikhathi ("time") > *isikhashana* ("little/short time").

- Augmentative nouns are usually formed by the suffixation of an augmentative suffix *-kazi* to a noun, e.g.
  itshe ("stone") > *itshekazi* ("huge boulder").

- Feminine nouns are sometimes formed by the suffixation of a feminine suffix *-kazi* (alternatively *-azi*) to a noun[2] , e.g.
  imvu ("sheep") > *imvukazi* ("ewe"); *inkomo* ("head of cattle") > *inkomazi* ("cow").

- A combination of diminutive and feminine: A feminine noun formed by the suffixation of a diminutive suffix *-kazi* or *-azi* may be followed by a diminutive suffix as well, e.g.
  izimvu ("sheep") > *izimvukazana* ("small ewes"),
  imbuzi ("goat") > *isibuzazana* ("young she-goat - not yet having given birth").

Other examples of the use of these suffixes are as follows:

- Diminutives of adjective stems, e.g.
  -khulu ("big/large") > *-khulwana* or *-khudlwana* ("somewhat large").

- Augmentative suffix with adjective stem, e.g.
  -khulu ("big/large") > *-khulukazi* (expresses additional greatness).

- Feminine suffix with adjective stem added to the adjective stem *-de* to bring about harmony with feminine nouns, e.g.
  inkomazi endekazi ("a tall cow")
  amantombazane amadekazi ("tall girls").

We will use *izimvukazana* ("small ewes") as example throughout. Its morphological analysis is as follows:

izin-	+ *-vu*	+ *-kazi*	+ *-ana*
noun stem.10	("ewe")	fem suffix	dim suffix

The concepts that are explicit in the morphological analysis of *izimvukazana* are shown in Table 2.

Morpheme	Syntactic concept	Semantic concept
izin-	Prefix of class 10 has syntactic role in sentence (nominal classification)	Class 10 indicates plural
-vu	Stem	"sheep"
-kazi	Suffix	Feminine
-ana	Suffix	Diminutive

Table 2: Morphological concepts in *izimvukazana*

Other relevant concepts are the augmentative suffix, class gender, which subsumes class 10 and all the other classes, and affix which subsumes prefix and suffix. At the lexical level relevant concepts for *izimvukazana* are, for example, diminutive noun and noun, but this forms part of future work.

4. Extending the DCR

Now that we have established the concepts that we need for the syntactic and semantic annotation of the morphological information in *izimvukazana* and, in general the Zulu noun, we proceed to select or create data categories for these concepts in order to extend the ISOcat DCR.

[2] Usually a common gender noun, while a different noun represents the masculine form, e.g. *inqama* ("ram") and *inkunzi* ("bull"), cf. (Taljaard and Bosch, 1998, 144).

4.1. The procedure

Our extension is based on the principle that if an appropriate data category already exists, we use it, if not, we add a new data category to the DCR. If there are multiple options available we consider those (see section 4.2) and select an appropriate DC. If not, we extend the registry, as described in (Anon., 2010). For now the DCs that are added to the DCR via the ISOcat web interface are marked as private categories. Once they have been tested, moderated and evaluated they will be marked as public. Since the broader aim is to add Zulu data categories in such a way that these categories are also useful for other related languages that share morphosyntactic structure with Zulu, we attempt to provide the most general definitions without losing essential information. Table 3 shows the DCs for the concepts in Table 2, for the augmentative suffix and the mentioned subsuming concepts.

DC	Existing/New	Key
Prefix	Existing	3417
Stem	Existing	3485
Suffix	Existing	3501
Class 10	New	6171
Feminine	Existing	1880
Diminutive	Existing	3046
Augmentative	New	6542
Class gender	New	6016
Affix	Existing	3072

Table 3: DCs for Zulu noun morphology, both existing and new

4.2. Discussion of the selection of existing DCs

In cases where more than one DC is appropriate, we opted for the DC of the GOLD ontology (Farrar and Langendoen, 2003) by way of consistency and coherence.

- **Prefix**: The selected DC is 3417: "An affix which is added to the front of a root or stem". This definition goes hand in hand with that of the affix (3072), and does not necessarily serve to change the meaning of a word as indicated in 293 and 1365, but may also serve to "change a word according to the grammatical context." (Kosch, 2006, 8).

- **Stem**: The selected DC is 3485: "Stem is the class of morphological units that are analysable into a root and possibly one or more derivational units. Stems can occur alone and are the basis for adding inflectional units." This definition provides more information on the nature of the stem vs. the root than for instance 1389.

- **Suffix**: The selected DC is 3501: "An affix, consisting of a letter, syllable, or syllables that follow a stem or word modifying its meaning. Suffixes may be inflectional or derivational." This definition is closer to the general function of suffixes in the Bantu languages than for instance 294 and 1395.

- **Affix**: The selected DC is 3072: "An affix is a morpheme with an abstract meaning which can only be used when added to a root morpheme. These are classified in four different ways, depending on their position with reference to the root: suffix, prefix, circumfix and infix." The significant part of this definition for the case of the Bantu languages is that the term referred to as an affix is a morpheme that cannot occur independently (cf. (Kosch, 2006, 8)). Some of the other definitions such as those in 291 and 1234 do not make explicit reference to this dependency.

- **Feminine**: The selected DC is 1880: "Of, relating to, or constituting the gender that ordinarily includes most words or grammatical forms referring to females." This definition caters for the feminine suffix in Zulu which may occur with nouns as well as with adjectives compared to 3197 which caters more specifically for languages with grammatical gender.

- **Diminutive**: The selected DC is 3046: "Form expressing smallness". This definition covers the diminutive suffix in Zulu which is not only suffixed to nouns but also to adjectives, and is therefore more appropriate than for instance 2225 where only a diminutive noun is mentioned in the definition.

4.3. The addition of new DCs

In the ISOcat online template for defining new data categories a justification, profile and status have to be provided for each new addition. For all the new DCs, given below, the justification is that they are "Bantu language identifiers", their profile is "morphosyntax", and their status is "private". The essential concept specific information, required in the template, is as follows:

English name: class 10
Key: 6171
PID: `http://www.isocat.org/datcat/DC-6171`
Identifier: class_10
Definition: Class designation used in the Bantu languages generally for miscellaneous nouns, including many animal names, in the plural.
Source: (Doke, 1967).

English name: augmentative
Key: 6542
PID: `http://www.isocat.org/datcat/DC-6542`
Identifier: augmentative
Definition: Form expressing augmentation.
Source: (Doke, 1967).

It was deemed necessary to create a new DC for the augmentative in particular for the Bantu languages since an existing DC such as 3094 "A special form of a noun that signals that the object being referred to is large relative to the usual size of such an object" would not make provision for augmentatives in adjective stems.

English name: class gender
Key: 6016
PID: `http://www.isocat.org/datcat/DC-6016`
Identifier: classGender
Definition: The morphosyntactic classification of nouns in the Bantu languages that generate grammatical agreement by means of class prefixes, also termed gender number prefixes.
Source: (Kosch, 2006).
Conceptual domain: class gender is a complex closed DC, which assumes as values the simple DCs class_1, class_1a, class_2, ..., class_15 and class_16-18.

The complete list of DCs relating to class number is shown in Table 4, which follows the references in section 6.

4.4. Defining a Data Category Selection for Zulu

As a final step we define a Data Category Selection (DCS) specifically for Bantu to ensure optimal interoperability between the Bantu languages. By selecting existing DCs, where possible, wider interoperability with other languages will also be achieved.

5. Conclusion and future work

In this paper we described our first efforts in extending the ISOcat DCR for the future annotation, both syntactic and semantic, of Zulu noun morphology. This forms part of a larger, longer term project towards extending the DCR with all the morphological concepts for Zulu and even other Bantu languages.

There seems to be a good mix of existing and new data categories, attesting to, on the one hand, the commonality between languages and the potential for interoperability and, on the other hand, the inherent difference between language families, which is also to be expected.

6. References

Anon. (2010). DCR Style Guidelines. `http://www.isocat.org/manual/DCRGuidelines.pdf`.

Doke, C. M. (1967). *The Southern Bantu Languages*. Dawsons of Pall Mall, London.

Farrar, S. and Langendoen, D. T. (2003). A linguistic ontology for the Semantic Web. *GLOT International*, 7(3):97–100. `http://linguistics-ontology.org/paper/12`.

Ide, N. and Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China. `http://www.cs.vassar.edu/~ide/papers/ICGL10.pdf`.

ISOcat. (n.d.). ISOcat Data Category Registry. `http://www.isocat.org/`.

Kosch, I. M. (2006). *Topics in Morphology in the African Language Context*. University of South Africa, Pretoria. ISBN 9781868883691.

Kubicek, H., Cimander, R., and Scholl, H. J. (2011). Layers of interoperability. In Kubicek, H., Cimander, R., and Scholl, H. J., editors, *Organizational Interoperability in E-Government: Lessons from 77 European Good-Practice Cases*. Springer, Berlin. ISBN: 9783642225024 (Online).

Meinhof, C. (1932). *Introduction to the phonology of the Bantu languages*. Dietrich Reimer/Ernst Vohsen, Berlin.

Taljaard, P. C. and Bosch, S. E. (1998). *Handbook of isiZulu*. J.L. van Schaik, Pretoria. ISBN 0627018742.

English name	Key	Identifier	Definition
class 1	6017	class_1	Class designation used in the Bantu languages generally for nouns denoting human beings in the singular.
class 1a	6540	class_1a	Class designation used in the Bantu languages generally for proper names, kinship terms, names of personified animals and objects, and nouns of foreign origin in the singular.
class 2	6163	class_2	Class designation used in the Bantu languages generally for nouns denoting human beings in the plural.
class 2a	6541	class_2a	Class designation used in the Bantu languages generally for proper names, kinship terms, names of personified animals and objects, and nouns of foreign origin in the plural.
class 3	6164	class_3	Class designation used in the Bantu languages generally for impersonal nouns, including names of plants, trees, some body parts, names of spirits, diseases, rivers and abstract nouns in the singular.
class 4	6165	class_4	Class designation used in the Bantu languages generally for impersonal nouns, including names of plants, trees, some body parts, names of spirits, diseases, rivers and abstract nouns in the plural.
class 5	6166	class_5	Class designation used in the Bantu languages generally for miscellaneous nouns, including majority of names of fruits in the singular.
class 6	6167	class_6	Class designation used in the Bantu languages generally for miscellaneous nouns, including majority of names of fruits in the plural; nouns denoting things occurring in pairs, fluids, abstract nouns.
class 7	6168	class_7	Class designation used in the Bantu languages generally for miscellaneous nouns, including names of languages, customs and habits, nature and the physical world, material objects and instruments in the singular.
class 8	6169	class_8	Class designation used in the Bantu languages generally for miscellaneous nouns, including names of languages, customs and habits, nature and the physical world, material objects and instruments in the plural.
class 9	6170	class_9	Class designation used in the Bantu languages generally for miscellaneous nouns, including many animal names, in the singular.
class 10	6171	class_10	Class designation used in the Bantu languages generally for miscellaneous nouns, including many animal names, in the plural.
class 11	6172	class_11	Class designation used in the Bantu languages generally for miscellaneous nouns and long objects, in the singular. Class 11 nouns take their plurals mainly in three classes, viz. 6, 10 and 14. The choice of the plural class is language dependent.
class 14	6173	class_14	Class designation used in the Bantu languages generally for abstract nouns and non-abstract nouns mostly collective, usually in the singular. Some nouns in this class take plural forms and these are mainly found in class 6.
class 15	6174	class_15	Class designation used in the Bantu languages for verbal infinitives. By the very nature of their meaning, infinitive forms do not show a distinction between singular and plural.
class 16-18	6175	class_16-18	Class designation used in the Bantu languages for the so-called locative noun classes. Very few nouns occur in these classes. No longer productive noun classes. Class 16 and 17 prefixes indicate locative adverbials.

Table 4: Complete data categories for class number added to the ISOcat Data Category Registry (Doke, 1967)

Understanding questions and finding answers: semantic relation annotation to compute the Expected Answer Type

Volha Petukhova

Spoken Language Systems, Saarland University, Germany
v.petukhova@lsv.uni-saarland.de

Abstract

The paper presents an annotation scheme for semantic relations developed and used for question classification and answer extraction in an interactive dialogue based quiz game. The information that forms the content of this game is concerned with biographical facts of famous people's lives and is often available as unstructured texts on internet, e.g. Wikipedia collection. Questions asked as well as extracted answers, are annotated with dialogue act information (using the ISO 24617-2 scheme) and semantic relations, for which an extensive annotation scheme is developed combining elements from TAC KBP slot filling and TREC QA tasks. Dialogue act information, semantic relations and identified focus words (or word sequences) are used to compute the Expected Answer Type (EAT). Our semantic relation annotation scheme is defined and validated according to ISO criteria for design of a semantic annotation scheme. The obtained results show that the developed tagset fits the data well, and that the proposed approach is promising for other query classification and information extraction applications where structured data, for example, in the form of ontologies or databases, is not available.

Keywords: semantic annotation, annotation scheme design, semantic relations

1. Introduction

According to the ISO Linguistic Annotation Framework (ISO, 2009), the term 'annotation' refers to linguistic information that is added to segments of language data and/or nonverbal communicative behaviour. Semantic annotations have been proven to be useful for various purposes. Annotated data is used for a systematic analysis of a variety of language phenomena and recurring structural patterns. Corpus data annotated with semantic information are also used to train machine learning algorithms for the automatic recognition and prediction of semantic concepts. Finally, semantically annotated data is used to build computer-based services and applications. One of the first steps in obtaining such annotations is the design of a semantic annotation scheme that fits the data well. The International Organization for Standards (ISO) has set up a series of projects for defining standards for the annotation of various types of semantic information, together forming the so-called Semantic Annotation Framework (SemAF). Different parts of SemAF are concerned with (1) time and events; (2) dialogue acts; (3) semantic roles; (4) spatial information; and (5) discourse relations. They define general theoretically and empirically well-founded domain- and language-independent concepts. This presents a good starting point for designing domain-specific schemes, if desired.

In this paper we discuss the design of a domain-specific annotation scheme for semantic relations used for a domain-specific Question Answering (QA) application. In a domain-specific QA, questions are expected about a certain topic; if a question outside that topic is asked, it will not be answered by the system.

The system described here is an interactive guessing game in which players ask questions about attributes of an unknown person in order to guess his/her identity. The player may ask ten questions of various types, and direct questions about the person's name or alias are not allowed. Moreover, the system is a Question Answering Dialogue System (QADS), where answers are not just pieces of extracted text or information chunks, but full-fledged natural language dialogue utterances. The system has all components that any traditional dialogue system has: Automatic Speech Recognition (ASR) and Speech Generation (e.g. TTS) modules, and the Dialogue Engine. The Dialogue Engine, in turn, consists of four components: the interpretation module, the dialogue manager, the answer extraction module and the utterance generation module. The dialogue manager (DM) takes care of overall communication between the user and the system. It gets as input a dialogue act representation from the interpretation module (IM), which it is usually about a question which is uttered by the human player. Questions are classified according to their communicative function (e.g. Propositional, Check, Set and Choice Questions) and semantic content. Semantic content is determined by Expected Answer Type (EAT), e.g. LOCATION as semantic relation, and the focus word, e.g. *study*. To extract the requested information, a taxonomy is designed comprising 59 semantic relations to cover the most important facts in human life, e.g. birth, marriage, career, etc. The extracted information is mapped to the EAT, and both the most relevant answer and a strategy for continuing the dialogue are computed. The DM then passes the system response along for generation, where the DM input is transformed into a dialogue utterance (possibly a multimodal and multifunctional one).

The paper is structured as follows. Section 2 gives an overview of previous approaches to designing semantic relation tagsets for QA applications. Section 3 discusses design criteria for the new semantic relation annotation scheme. Section 4 defines the semantics of the relations and groups them into a hierarchical taxonomy. Section 5 describes the collection of dialogue data and annotations, with indicated reliability of the defined annotation scheme in terms of inter-annotator agreement. In Section 6 classification results using semantic relations in questions and for answer extraction are presented. Section 6 concludes the reported study and outlines future research.

2. Related work

A major breakthrough in QA has been made by (Moldovan et al., 2000) when designing an end-to-end open-domain QA system. This system achieved the best result in the TREC-8 competition[1] with an accuracy of 77.7%. Their system contains the three components: question processing, paragraph indexing and answer processing. First, the question type, question focus, question keyword and expected answer type are specified. There are 9 question classes (e.g. *'what'*, *'who'*, *'how'*) and 20 sub-classes (e.g. *'basic what'*, *'what-who'*, *'what-when'*). Additionally, expected answer type is determined, e.g. *person*, *money*, *organization*, *location*. Finally, a focus word or a sequence of words is identified in the question, which disambiguates it by indicating what the question is looking for (see Moldovan et al., 2000 for an overview of defined classes for 200 of the most frequent TREC-8 questions).

Li and Roth (2002) proposed another question classification scheme, also based on determining the expected answer type. This scheme is a layered hierarchical one having two levels. The first level represents coarse classes like *Date*, *Location*, *Person*, *Time*, *Definition*, *Manner*, *Number*, *Price*, *Title*, *Distance*, *Money*, *Organization*, *Reason* and *Undefined*. The second level has 50 fine-grained classes like *Description*, *Group*, *Individual* and *Title* for the upper-level class of *Human*.

The most recent work comes from the TAC KBP slot filling task (Joe, 2013) aiming to find filler(-s) for each identified empty slot, e.g. for a person (e.g. date_of_birth, age, etc.) and/or for an organization (e.g. member_of, founded_by, etc). Pattern matching, trained classifiers and Freebase[2] are used (Min et al., 2012) and (Roth et al., 2012) to find the best filler. The best system performance achieved in terms of F-score is 37.28% (see Surdeanu, 2013 and Roth et al., 2013).

We see that semantic relations are commonly used to compute an expected answer type. Our task, domain and data differ from the above mentioned approaches in that (1) our domain is closed, (2) the content is mainly unstructured internet articles, and (3) the answers are not just extracted chunks or slot fillers, but rather full dialogue utterances. These aspects cannot be captured by existing annotation approaches. Therefore, we propose a new semantic relation annotation scheme and when developing it we rely on criteria formulated for semantic annotation ISO standards design (see e.g. ISO 24617-2). These criteria support well-founded decisions when designing the conceptual content and structure of the annotation scheme. We discuss the criteria in the next Section.

3. Annotation scheme design criteria

The design of a scheme for annotating primary language data with semantic information is subject to certain methodological requirements, some of which have been made explicit in various studies (Bunt and Romary, 2002; Ide et al., 2003; Bunt and Romary, 2004), and some of which have so far remained implicit. For example, Bunt and Romary (2002) introduce the principle of *semantic adequacy*, which is the requirement that semantic annotations should have a semantics. This is because a semantic annotation is meant to capture something of the meaning of the annotated stretch of source text, but if the annotation does not have a well-defined semantics, then there is no reason why the annotation should capture meaning any better than the source text itself.

A semantic annotation scheme is intended to be applied to language resources, in particular to collections of empirical data. It should therefore contain concepts for dealing with those phenomena which are found in empirical data, allowing good coverage of the phenomena of interest.

Finally, an annotation scheme should be practically useful, i.e. be effectively usable by human annotators and by automatic annotation systems; it should not be restricted in applicability to source texts in a particular language or group of languages; and it should incorporate common concepts of existing annotation schemes where possible.

From these considerations, the following general criteria can be distilled:

- *compatibility*: incorporate common concepts of existing annotation schemes, thus supporting the mapping from existing schemes to the new one, and ensuring the interoperability of the defined scheme.
- *theoretical validity*: every concept defined has a well-defined semantics.
- *empirical validity*: concepts defined in the scheme correspond to phenomena that are observed in corpora.
- *completeness*: concepts defined in the scheme provide a good coverage of the semantic phenomena of interest.
- *distinctiveness*: each concept defined in the scheme is semantically clearly distinct from the other concepts defined.
- and *effective usability*: concepts defined in the scheme are learnable for both humans and machines with acceptable precision.

We will show in this paper that each of these criteria is fulfilled, supporting well-founded decisions when designing the conceptual content and structure of the proposed annotation scheme.

4. Semantic relations

In order to find the answer to a certain question, semantic role information can be used. A semantic role is a relational notion (between an event and its participant) and describes the way a participant plays in an event or state (first defined as such in (Jackendoff, 1972) and (Jackendoff, 1990)), as described mostly by a verb, typically providing answers to questions such as "who" did "what" to "whom," and "when," "where," "why," and "how." Several semantic role annotation schemes have been developed in the past, e.g. FrameNet (ICSI, 2005), PropBank (Palmer et al., 2002), VerbNet (Kipper, 2002) and Lirics (Petukhova and Bunt, 2008).

[1] http://trec.nist.gov/pubs/trec8

[2] http://www.freebase.com/

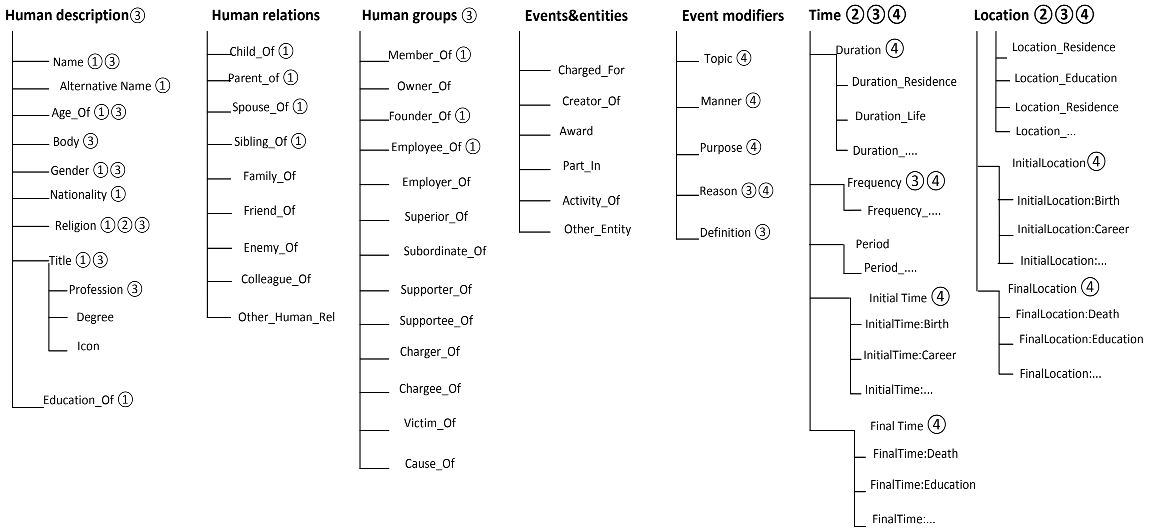


Figure 1: Semantic relations taxonomy. (① means that the relation is also defined in TAC KBP slot filling task; ② in TREC-08 QA task; ③ in TREC 2002 QA task, i.e. annotation scheme proposed by (Li and Roth, 2002); and ④ in LIRICS semantic role set)

Communicative function	%
Propositional Questions	22.4
Set Questions	38.8
Choice Questions	10.4
Check Questions	23.9
Unspecified Question Type	4.5

Table 1: Distribution of information-seeking communicative functions in the annotated data.

Along with semantic roles, relations between participants are also relevant for our domain, e.g. the relation between Agent and Co-Agent (or Partner) involved in a 'work' event may be a COLLEAGUE_OF relation.

To decide on the set of relations to investigate, we analysed available and collected new dialogue data. As a starting point, we analysed recordings of the famous US game 'What's my line?' that are freely available on Youtube (www.youtube.org). However, the latter differs from our scenario: during the TV-show participants may ask only propositional questions with expected 'yes' or 'no' answers;, our game allows any question type from the user. Therefore, we collected data in pilot dialogue experiments, where one participant was acting as a person whose name should be guessed and the other as a game player. 18 dialogues were collected of total duration of 55 minutes comprising 360 system's and user's speaking turns. To evaluate the relation set and to train classifiers, we performed large scale gaming experiments in a Wizard of Oz setting (see Section 4).

Pilot experiments showed that all players tend to ask similar questions about gender, place and time of birth or death, profession, achievements, etc. To capture this information we defined 59 semantic relations. We proposed a multi-layered taxonomy: a high level, coarse annotation comprising 7 classes and a low-level, fine-grained annotation, comprising 52 classes. This includes the HUMAN DESCRIPTION class defined for basic facts about an individual like age, title, nationality, religion, etc.; HUMAN RELATIONS for parent-child and other family relations; HUMAN GROUPS for relations between colleagues, friends, enemies, etc.; EVENTS&NON-HUMAN ENTITIES class for awards, achievements, products of human activities, etc.; EVENT MODIFIERS for specifying manner, purpose, reasons, etc.; the TIME class to capture temporal information like duration, frequency, period, etc.; and the LOCATION class to capture spatial event markers for places where events occur. Some of the second-level classes are broken down into even more specific classes. For example, TITLE has three classes such as PROFESSION for official name(s) of the employment and occupation/job positions; DEGREE for unofficial and official names of obtained degrees and degrees within an organization, e.g. 'highest paid athlete', 'doctor in physics', 'senior leader', etc.; and ICON for unofficial or metaphorical titles that do not refer to an employment or membership position, e.g. 'public figure', 'hero', 'sex symbol', etc. Figure 1 shows the defined hierarchical taxonomy with an indication of what concepts can be found in existing schemes for annotating semantic relations and semantic roles. It should be noted here that the majority of the concepts defined here are domain-specific, i.e. tailored to our quiz game application. The approach could however be adapted for designing comparable annotation schemes for other domains; this has for example been done for the food domain (see Wiegand and Klakow, 2013).

From a semantic point of view, each relation has two arguments and is one of the following types:

- RELATION(Z,?X), where Z is the person in question and X the entity slot to be filled, e.g. CHILD_OF(einstein,?X);
- RELATION(E_1, ?E_2) where E_1 is the event in question and E_2 is the event slot to be filled, e.g. REA-

RELATION	%	RELATION	%	RELATION	%	RELATION	%
ACTIVITY_OF	10.21	LOC_BIRTH	2.34	AGE_OF	3	LOC_DEATH	1.69
AWARD	4.4	LOC_RESIDENCE	1.69	BODY	1.5	MANNER	1.12
CHARGED_FOR	4.21	MEMBER_OF	2.43	CHILD_OF	1.5	NAME	1.87
COLLEAGUE_OF	1.03	NATIONALITY	1.22	CREATOR_OF	6.09	OWNER_OF	1.97
DESCRIPTION	4.12	PARENT_OF	1.31	DURATION	1.31	REASON	1.22
EDUCATION_OF	3.65	RELIGION	2.53	EMPLOYEE_OF	1.59	SIBLING_OF	0.94
ENEMY_OF	1.12	SPOUSE_OF	1.4	FAMILY_OF	1.59	SUPPORTED_BY	0.94
FOUNDER_OF	1.87	TIME	7.96	FRIEND_OF	1.03	TIME_BIRTH	2.06
GENDER	1.69	TIME_DEATH	1.59	LOCATION	4.68	TITLE	11.14

Table 2: Question types in terms of defined semantic relations and their distribution in data (relative frequency in %).

SON(death,?E$_2$); and

- RELATION(E,?X) where E is the event in question and X the entity slot to be filled, e.g. DURATION(study,?X).

The slots to be filled are categorized primarily based on the type of entities which we seek to extract information about. However, slots are also categorized by the *content* and *quantity* of their fillers.

Slots are labelled as *name*, *value*, or *string* based on the content of their fillers. *Name* slots are required to be filled by the name of a person, organization, or geo-political entity (GPE). *Value* slots are required to be filled by either a numerical value or a date. The numbers and dates in these fillers can be spelled out (December 7, 1941) or written as numbers (42; 12/7/1941). *String* slots are basically a "catch all", meaning that their fillers cannot be neatly classified as names or values.

Slots can be *single-value* or *list-value* based on the number of fillers they can take. While single-value slots can have only a single filler, e.g. date of birth, list-value slots can take multiple fillers as they are likely to have more than one correct answer, e.g. employers.

5. Data collection and annotations

In order to validate the proposed annotation scheme empirically, two types of data are required: (1) dialogue data containing player's questions that are more realistic than youtube games and larger than our pilots; and (2) descriptions containing answers to player's questions about the guessed person. This data is also required to build an end-to-end QADS.

To collect question data we explored different possibilities. There is some question data publicly available, e.g. approximately 5500 questions are provided by the University Illinois[3] annotated according to the scheme defined in (Li and Roth, 2002). However, not all of this data can be used for our scenario. We filtered out about 400 questions for our purposes. Since this dataset is obviously too small, we generated questions automatically using the tool provided by (Heilman and Smith, 2009) from the selected Wikipedia articles and filtered them out manually. Out of the generated 3000 questions relevant ones were selected: grammatically broken questions were fixed and repetitions deleted. Additionally, synonyms from WordNet[4] were used to generate different variations of questions for the same class. Questions collected in pilot experiments were added to this set as well. The final question set consists of 1069 questions. These questions are annotated with (1) communicative function type according to ISO 24617-2; (2) with semantic relations as defined in Section 3; and (3) with question focus word or word sequence. Table 1 provides an overview of the types of information-seeking communicative functions in the collected data and those relative frequencies.

Table 2 illustrates the distribution of question types based on the EAT's semantic relation.

A focus word or word sequence describes the main event in a question, usually specified by a verb or eventive noun. The focus word (sequence) is extracted from the question to compute the EAT and formulate the query. For example,

(1) Question: When was his first album released?
Assigned semantic relation: TIME
Focus word sequence: first album released
EAT: TIME_release(first_album)
Query:
TIME_release(first_album) :: (E, ?X) :: QUALITY(VALUE) :: QUANTITY(SINGLE)

The question set is currently enriched with questions from large scale Wizard of Oz experiments. The data collection procedure was similar to that of pilots. A Wizard (English native speaker) simulated the system's behaviour and the other participant played the game. 21 unique subjects, undergraduates of age between 19 and 25, who are expected to be related to our ultimate target audience, participated in these experiments. 338 dialogues were collected of a total duration of 16 hours comprising about 6.000 speaking turns. An example from this dialogue collection can be found in the Appendix.

Answers were retrieved from 100 selected Wikipedia articles in English containing 1616 sentences (16 words/sentence on average), 30.590 tokens (5.817 unique tokens). Descriptions are annotated using complex labels consisting of an IOB-prefix (**I**nside, **O**utside, and **B**eginning), since we aim to learn the exact answer boundaries, and semantic relation tag, the same as used for classifying questions. We mainly focus on labeling nouns and noun phrases. For example:

[3] http://cogcomp.cs.illinois.edu/page/resources/data

[4] urlhttp://wordnet.princeton.edu/

RELATION	%	RELATION	%	RELATION	%	RELATION	%	RELATION	%
ACCOMPLISHMENT	4.0	DURATION	1.8	LOC_DEATH	0.8	PART_IN	3.6	TIME	14.6
AGE_OF	2.1	EDUCATION_OF	4.2	LOC_RESIDENCE	3.2	RELIGION	0.7	TIME_BIRTH	2.8
AWARD	2.5	EMPLOYEE_OF	2.2	MEMBER_OF	1.8	SIBLING_OF	2.3	TIME_DEATH	1.0
CHILD_OF	3.6	FOUNDER_OF	1.2	NATIONALITY	3.1	SPOUSE_OF	1.9	TITLE	14.2
COLLEAGUE_OF	1.7	LOC	5.6	OWNER_OF	1.1	SUBORDINATE_OF	1.3		
CREATOR_OF	8.5	LOC_BIRTH	5.0	PARENT_OF	3.7	SUPPORTEE_OF	1.1		

Table 3: Answer types in terms of defined semantic relations and their distribution in data (relative frequency in %)

(2) *Gates graduated from* ***Lakeside School*** *in 1973.*

The word *Lakeside* in (2) is labeled as the beginning of an EDUCATION_OF relation (B-EDUCATION_OF), and *school* is marked as inside of the label (I-EDUCATION_OF). Table 3 illustrates the distribution of answer types based on the identified semantic relation.

Since the boundaries between semantic classes are not always clear, we allowed multiple class labels to be assigned to one entity. For example:

(3) *Living in Johannesburg, he became involved in anti-colonial politics, joining the ANC and becoming a founding member of its* ***Youth League****.*

Here, *Youth League* is founded by a person (FOUNDER_OF relation), but the person is also a member of the *Youth League*. There are also some overlapping segments detected as in example (4):

(4) *He served as* ***the commander-in-chief of the Continental Army*** *during the American Revolutionary War.*

The entity *commander-in-chief of the Continental Army* in (4) is marked as TITLE, while *the Continental Army* is recognized as MEMBER_OF. Both of these relations are correct, since if a person leads an army he/she is also a member of it.

To assess the reliability of the defined tagset, the inter-annotator agreement was measured in terms of the standard Kappa statistic (Cohen, 1960). For this, 10 randomly selected descriptions and all 1069 questions were annotated by two trained annotators. The obtained *kappa* scores were interpreted as annotators having reached good agreement (averaged for all labels, kappa = .76).

6. Semantic relation classification and learnability

To investigate the learnability of the relations we defined in a data-oriented way and to evaluate the semantic relation set, we performed a number of classification experiments. Moreover, we partition the training sets in such a way that we can assess relation learnability by plotting learning curves for each relation given an increasing amount of training data.

Classifiers used were statistical ones, namely, Conditional Random Fields (CRF) (Lafferty et al., 2001) and Support Vector Machines (SVM) (Joachims et al., 2009).[5].

The selected feature set includes **word & lemma tokens**; **n-grams** and **skip n-grams** for both tokens and their lemmas; **POS** tags from the Stanford POS tagger (Toutanova et al., 2003); **NER** tags from three different NER tools: Stanford NER (Finkel et al., 2005), Illinois NER (Ratinov and Roth, 2009), and Saarland NER (Chrupala and Klakow, 2010); **chunking** using OpenNLP [7] to determine NP boundaries; **key word** to determine the best sentence candidate for a particular relation, e.g. *marry, married, marriage, husband, wife, widow, spouse* for the SPOUSE_OF relation.

System	Precision	Recall	F-score	Accuracy
Baseline	76.87	73.79	73.72	97.38
System 1	80.18	77.71	78.05	97.89

Table 4: Question classification results

To assess the system performance standard evaluation metrics are used, precision (P), recall (R) and F-score (F1). In particular, precision is important, since it is worse for the system to provide a wrong answer than not to provide any answer at all, e.g. to say it cannot answer a question.[8] It should be noted that for answer extraction sequential classifiers were trained and their predictions were considered as correct iff both the IOB-prefix and the relation tag fully correspond to those in the referenced annotation.

6.1. Question classification

In the 10-fold cross-validation classification experiments, classifiers were trained and evaluated in two different settings: (1) *Baseline*, where classification is based solely on the bag-of-words features; (2) and *System 1*: best system performance after trying different sets of features and selection mechanisms, namely, on bag-of-words plus bigrams generated from bag-of-lemmas. Table 4 presents the classification results.

It may be observed that System 1 clearly outperforms the baseline. The results are also better than those of the state-of-art systems on this task. To compare, the system reported in (Dell and Wee Sun, 2003) using SVM reached 80.2% accuracy (using bag-of-words) and 79.2% (using bag-of-ngrams) for the 50 question classes defined in (Li and Roth, 2002) and on their data. The reported in (Huang et al., 2008) the accuracies of SVM and Maximum Entropy (ME) classifiers were 89.2% and 89.0% respectively on the data and taxonomy of (Li and Roth, 2002). The best performance in terms of accuracy reported by Li and Roth (2006)

[5] We used two CRF implementations from CRF++[6]

[7] http://opennlp.apache.org/

[8] Each WoZ experiment participant filled in a questionnaire, where among other things they indicated that 'not-providing' an answer was entertaining; giving wrong information, by contrast, was experienced as annoying.

	Baseline			System 1			System 2		
	P	R	F1	P	R	F1	P	R	F1
CRF ++	0.56	0.34	0.42	0.68	0.52	0.59	0.82	0.55	0.66
SVM-HMM	0.59	0.28	0.38	0.53	0.51	0.52	0.72	0.47	0.57
Pattern*	-	-	-	-	-	-	0.74	0.62	0.67

Table 6: Overall system performance. *) applied only to 12 most frequently occurring relations

of the tagset was 89.3% using the SNoW learning architecture for a hierarchical classifier .

The performance of the classifiers (System 1 setting) on each relation in isolation has also been assessed. Table 5 presents the obtained results.

Our classifiers achieved reasonably high accuracy in detecting all relations. In terms of F-score, three relations were rather problematic, namely OWNER_OF, DESCRIPTION and SUPPORTEE_OF. For the latter, the number of training instances was rather low as we will show in our learnability experiments (see Section 5.3). For the first one, we have concluded that this relation requires a more clear definition to make better distinctions with other classes, e.g. it is often confused with CREATOR_OF and FOUNDER_OF. Similarly, the DESCRIPTION relation has a rather vague definition and tends to be applied for many unclassifiable instances. We introduce two relations instead: DEFINITION and TOPIC (see Figure 1).

6.2. Answer extraction

In the 5-fold cross-validation classification experiments, classifiers were trained and evaluated in three different settings: (1) *Baseline* obtained when training classifiers on word token features only; (2) *System 1* where classification is based on automatically derived features such as n-grams for tokens and lemmas (trigrams), POS, NER tags and chunking; joint classification on all relations; (3) and *System 2*: pattern matching and classification on the same features as System 1 applied for each relation separately.

Both CRF++ and SVM-HMM classifiers in System 1 and 2 settings show gains over the baseline systems. To appreciate how good statistical classifiers generally are on relation recognition for answer extraction, consider the performance of distant supervision SVM[9] with precision of 53.3, recall of 21.8 and F-score of 30.9 (Roth et al., 2013) on the TAC KBP relations. However, we emphasize that our task, relation set, application and data are different from those of TAC KBP.

As can be observed from Table 6, the CRF++ classifier achieves the best results in terms of precision and F-score. Although the running time was not measured, the classification runs faster than the SVM-HMM. System 2 outperforms System 1 (6-11% increase in F-score). When training on each relation in isolation, feature weights can be adjusted more efficiently, while not affecting other classifiers' performances.

More detailed results from CRF++ on each semantic relation classification can be seen in Table 7.

Relation	P	R	F1	Relation	P	R	F1
ACCOMPLISHMENT	0.73	0.44	0.55	NATIONALITY	0.92	0.73	0.81
AGE_OF	0.95	0.76	0.84	OWNER_OF	0.76	0.40	0.48
AWARD	0.80	0.62	0.70	PARENT_OF	0.79	0.54	0.63
CHILD_OF	0.74	0.58	0.65	PART_IN	0.25	0.05	0.08
COLLEAGUE_OF	0.78	0.32	0.43	RELIGION	0.60	0.16	0.24
CREATOR_OF	0.64	0.17	0.26	SIBLING_OF	0.92	0.69	0.78
DURATION	0.97	0.64	0.76	SPOUSE_OF	0.76	0.42	0.52
EDUCATION_OF	0.84	0.65	0.72	SUBORDINATE_OF	0.81	0.19	0.31
EMPLOYEE_OF	0.77	0.19	0.28	SUPPORTEE_OF	1.00	0.40	0.54
FOUNDER_OF	0.65	0.26	0.36	MEMBER_OF	0.65	0.14	0.21
LOC	0.77	0.33	0.45	TIME	0.90	0.83	0.86
LOC_BIRTH	0.94	0.84	0.89	TIME_BIRTH	0.92	0.89	0.90
LOC_DEATH	0.90	0.55	0.67	TIME_DEATH	0.94	0.79	0.86
LOC_RESIDENCE	0.86	0.55	0.66	TITLE	0.84	0.66	0.74

Table 7: CRF++ performance on System 2.

6.3. Learnability

The outcome from the learnability experiments is presented in Figure 2. From these graphs, we can clearly observe that larger training data positively correlates with higher F-score. The SUPPORTEE_OF is the most sensitive relation to the amount of training data, followed by LOC_DEATH and SUBORDINATE_OF.

7. Discussion and conclusions

We propose an annotation scheme for question classification and answer extraction from unstructured textual data based on determining semantic relations between entities. Semantic relation information together with the focus words (or word sequences) is used to compute the Expected Answer Type. Our results show that the relations that we have defined help the system to understand user's questions and to capture the information, which needs to be extracted from the data. The proposed scheme fits the data and is reliable, as evidenced by good inter-annotator agreement. Semantic relations can be learned successfully in a data-oriented way. We found the ISO semantic annotation scheme design criteria very useful. Following them supported our decisions when defining concepts and the structure of the scheme. The proposed approach is promising for other query classification and information extraction tasks for domain-specific applications.

There is a lot of room for further research and development, and the annotation scheme is far from perfect. For instance, observed inter-annotator agreement and classification results indicate that some relations need to be re-defined. We will test how generic the proposed approach is by testing it on the TAC and TREC datasets. Moreover, since some relations, in particular of RELATION(E$_1$, ?E$_2$) and RELATION(E,?X) types, are very close to semantic roles, there is a need to analyse semantic role sets (e.g. ISO semantic roles (Bunt and Palmer, 2013)) and study the possible overlaps.

From the QADS development point of view, we need to evaluate the system in real settings. For this, the ASR is currently retrained, i.e. generic language and acoustic models are adapted to our game scenario. For now, all classification experiments were run on data transcribed by a human. It is a semi-automatic process, when the ASR output has been corrected. The real system, however, needs to operate on ASR output lattices (list of hypotheses for each token with

[9] Distant supervision method is used when no or little labeled data is available, see (Mintz et al., 2009).

Relation	P	R	F1	Accuracy (in %)	Relatio	P	R	F1	Accuracy (in %)
ACTIVITY_OF	0.61	0.72	0.67	92.56	AGE_OF	1.00	0.93	0.96	99.78
AWARD	0.83	0.85	0.84	98.59	BODY	0.54	0.59	0.57	98.64
CHARGED_FOR	0.96	0.87	0.91	99.27	CHILD_OF	0.85	0.76	0.81	99.45
COLLEAGUE_OF	0.63	0.65	0.64	99.25	CREATOR_OF	0.73	0.69	0.71	96.58
DESCRIPTION	0.32	0.42	0.36	93.86	DURATION	0.93	0.99	0.96	99.90
EDUCATION_OF	0.91	0.79	0.85	98.97	EMPLOYEE_OF	0.91	0.75	0.83	99.49
ENEMY_OF	0.81	0.56	0.66	99.35	FAMILY_OF	0.45	0.88	0.59	98.07
FOUNDER_OF	0.85	0.66	0.74	99.14	FRIEND_OF	1.00	0.72	0.84	99.71
GENDER	1.00	0.97	0.99	99.95	LOCATION	0.78	0.91	0.84	98.38
LOC_BIRTH	0.99	0.92	0.95	99.79	LOC_DEATH	0.80	0.89	0.84	99.44
LOC_RESIDENCE	0.93	0.71	0.81	99.42	MANNER	1.00	0.92	0.96	99.91
MEMBER_OF	0.92	0.67	0.77	99.04	NAME	0.95	0.91	0.93	99.73
NATIONALITY	0.97	0.48	0.64	99.34	OWNER_OF	0.42	0.22	0.29	97.86
PARENT_OF	0.74	0.91	0.82	99.46	REASON	1.00	0.61	0.76	99.52
RELIGION	0.99	0.74	0.85	99.34	SIBLING_OF	0.98	0.80	0.88	99.80
SPOUSE_OF	0.78	0.59	0.67	99.19	SUPPORTEE_OF	0.69	0.20	0.31	99.17
TIME	0.94	0.95	0.95	99.16	TIME_BIRTH	0.95	0.85	0.90	99.61
TIME_DEATH	1.00	0.71	0.83	99.53	TITLE	0.73	0.89	0.80	95.01

Table 5: Question classification results for each relation in isolation.(*presented in alphabetic order)

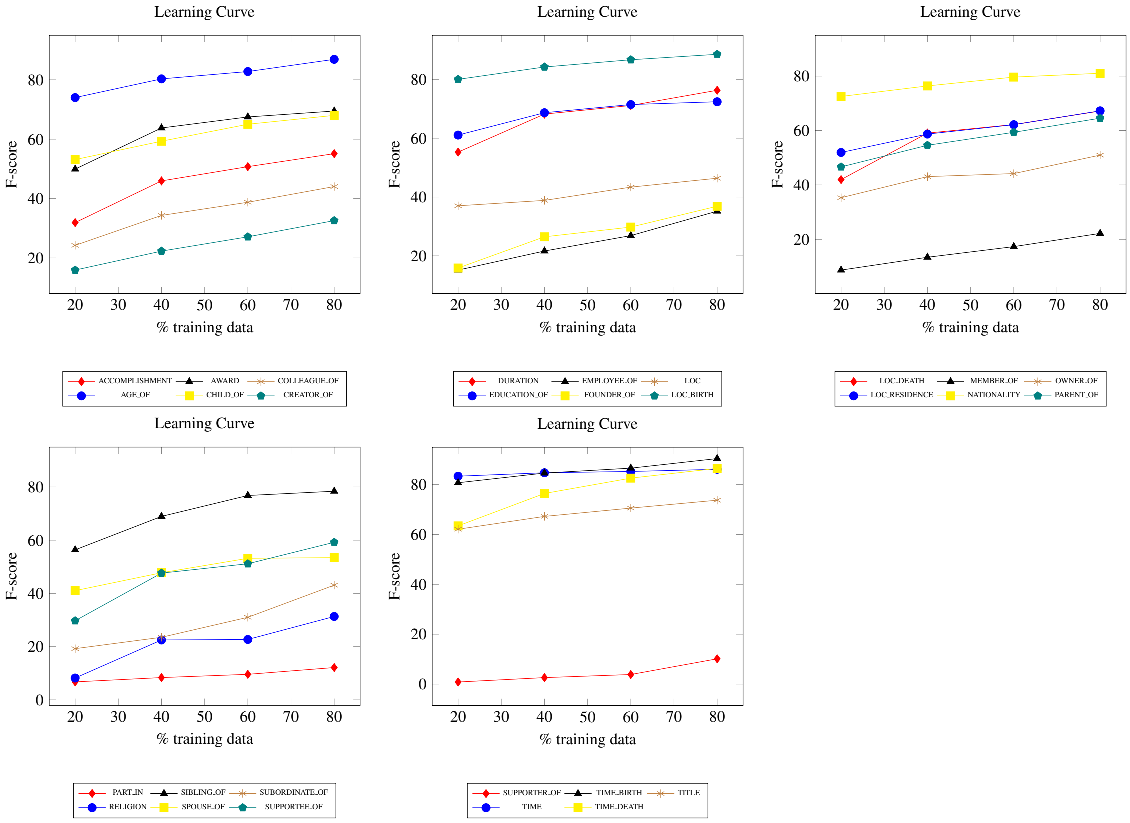


Figure 2: Learning curves for the defined relations

the recognizer's confidence scores). Therefore, in the nearest future we will test the question classifiers performance on the actual ASR output.

8. Acknowledgments

The research reported in this paper was carried out within the DBOX Eureka project under number E! 7152. The author is particularly thankful to her Master students Desmond Darma Putra and Humayun Faiz who performed machine-learning classification experiments.

9. References

H. Bunt and M. Palmer. 2013. Conceptual and representational choices in defining an iso standard for semantic role annotation. In *Proceedings Ninth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, Potsdam.

H. Bunt and L. Romary. 2002. Towards multimodal semantic representation. In *Proceedings of LREC 2002 Workshop on International Standards of Terminology and Language Resources Management*, pages 54–60, Las Palmas, Spain.

H. Bunt and L. Romary. 2004. Standardization in multimodal content representation: Some methodological issues. In *Proceedings of LREC 2004*, pages 2219–2222, Lisbon, Portugal.

G. Chrupala and D. Klakow. 2010. A named entity labeler for german: Exploiting wikipedia and distributional clusters. In *Proceedings of LREC'10*, Valletta, Malta. European Language Resources Association (ELRA).

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46.

Z. Dell and L. Wee Sun. 2003. Question classification using support vector machines. In *Proceedings of SIGIR*, pages 26–32, Toronto, Canada.

J. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL '05*, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Heilman and N. Smith. 2009. Question generation via overgenerating transformations and ranking. Language Technologies Institute, Carnegie Mellon University Technical Report CMU-LTI-09-013.

Z. Huang, M. Thint, and Z. Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of EMNLP*, pages 927–936.

ICSI. 2005. Framenet. Available at `http://framenet.icsi.berkeley.edu`.

N. Ide, L. Romary, and E. de la Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of HLT-NAACL Workshop on The Software Engineering and Architecture of Language Technology*, Edmunton.

ISO. 2009. *ISO 24612:2009 Language resource management: Linguistic annotation framework (LAF)*. ISO, Geneva.

ISO. 2012. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2*. ISO Central Secretariat, Geneva.

R.S. Jackendoff. 1972. *Semantic interpretation in generative grammar*. MIT Press.

R.S. Jackendoff. 1990. *Semantic structures*. MIT Press.

T. Joachims, T. Finley, and C.-N. Yu. 2009. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59.

E. Joe. 2013. Tac kbp 2013 slot descriptions.

K. Kipper. 2002. Verbnet: A class-based verb lexicon. Available at `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

X. Li and D. Roth. 2002. Learning question classifiers. In *Proceedings of the COLING '02*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

X. Li and D. Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, pages 229–249.

B. Min, X. Li, R. Grishman, and S. Ang. 2012. New york university 2012 system for kbp slot filling. In *Proceedings of the 5th Text Analysis Conference (TAC 2012)*.

M. Mintz, R. Bills, S.and Snow, and Jurafsky D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJCNLP)*, page 10031011.

D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum, and V. Rus. 2000. The structure and performance of an open-domain question answering system. In *Proceedings of ACL '00*, pages 563–570, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Palmer, D. Gildea, and P. Kingsbury. 2002. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

V. Petukhova and H. Bunt. 2008. Lirics semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the sixth international conference on language resources and evaluation (LREC 2008)*. Paris: ELRA.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL '09*, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Roth, G. Chrupala, M. Wiegand, M. Singh, and D. Klakow. 2012. Saarland university spoken language systems at the slot filling task of tac kbp 2012. In *Proceedings of the 5th Text Analysis Conference (TAC 2012)*, Gaithersburg, Maryland, USA.

B. Roth, T. Barth, M. Wiegand, M. Singh, and D. Klakow. 2013. Effective slot filling based on shallow distant supervision methods. In *TAC KBP 2013 Workshop*, Gaithersburg, Maryland USA. National Institute of Standards and Technology.

M. Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *TAC KBP 2013 Workshop*, Gaithersburg, Maryland USA. National Institute of Standards and Technology.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL '03*, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Wiegand and D. Klakow. 2013. Towards the detection of reliable food-health relationships. In *Proceedings of the NAACL-Workshop on Language Analysis in Social Media (NAACL-LASM)*.

Appendix: dialogue example

S: Hello
P: Hello
S: Good afternoon almost evening
S: What is your name
P: My name is James
S: Hello James it's nice to meet you
P: Nice to meet you
S: How are you doing today?
P: Good, thank you
S: Alright
S: Today we are going to play a game and here are the rules
S: I'm a very famous person and you need to guess my name you can ask whatever questions you want of me except for my name directly
S: You have at most ten questions and then you get to guess my name exactly once
S: So you can ask whatever questions you want but then if you want to guess my name you only get one try
S: If you get my name correct you win if you get my name incorrect or choose to pass then you lose and then we'll move on to the next round
S: Do you understand and are comfortable with the rules?
P: Yeah yeah
P: So the name is kind of a famous person
P: Okay
P: I'm not sure how good am I in this area
S: Yes
S: I am a famous person and I am male
P: Okay okay good
S: Alright
S: And what is your first question?
P: What is the first question
P: What do you do?
S: I am a leader
P: A leader
P: What is your nationality?
S: I am American
P: Are you alive?
S: I am not alive
P: Are you leading a company?
S: I am not leading a company
P: okay
P: You're not a company leader
P: When are you born?
S: I was born on February twenty second seventeen thirty two
P: Seventeen thirty two
P: Ok
P: Eehm
P: Are a politician?
S: I am a politician
P: Okay
P: So then it is not my area but I will try to guess
P: When were you in the government?
S: Uhm
S: Let's see
S: I retired from the presidency in seventeen ninety seven
P: Ninety seven
P: George Washington
S: Is that your final guess?
P: Yes, Washington
S: Very good, excellent job!
S: Congratulations!

From Visual Prototypes of Action to Metaphors
Extending the IMAGACT Ontology of Action to Secondary Meanings

Susan Windisch Brown
University of Florence
Piazza Savanarola, 1, Florence, Italy
E-mail: susanwbrown@att.net

Abstract

This paper describes an infrastructure that has been designed to deal with corpus-based variations that do not fall within the primary, physical variation of action verbs. We have first established three main categories of secondary variation--*metaphor*, *metonymy* and *idiom*--and criteria for creating types within these categories for each verb. The criteria rely heavily on the images that compose the IMAGACT ontology of action and on widely accepted processes of meaning extension in linguistics. Although figurative language is known for its amorphous, subjective nature, we have endeavoured to create a standard, justifiable process for determining figurative language types for individual verbs. We specifically highlight the benefits that IMAGACT's representation of the primary meanings through videos brings to the understanding and annotation of secondary meanings.

Keywords: semantic annotation, metaphor, metonymy

1. Introduction

IMAGACT is a cross-linguistic ontology of action concepts that are represented with prototypic 3D animations or brief films. This format makes use of the universal language of images to identify action types, avoiding the under-determinacy of semantic definitions. This ontology has been induced from the references to physical actions found in English and Italian spoken corpora (Moneglia et al. 2012) and gives a picture of the variety of activities that are prominent in our everyday life, specifying the language used to express each one in ordinary communication.

IMAGACT uses prototypic scenes to represent the range of variations that natural language verbs can record in a language and maps different languages onto the same ontology of visually represented concepts. Each verb can express one or more concepts, while each concept can refer to one or more verbs. (Moneglia in press).

For example, the verb *to cross* ranges over four main action types (Figure 1), identified in corpus occurrences, some of which can be equivalently identified by other verbs *(pass, climb)*. The specific way of categorizing actions by the verb *to cross* does not find direct correspondence in other languages. For instance, in Italian only type 1 and 3 can be in the extension of the direct translation *(attraversare)* while 2 and 4 respectively require other Italian verbs *(incrociare, superare).*

The IMAGACT ontology has been developed through annotation of English and Italian spoken corpora, in which reference to actions is frequent. Working in their native languages, linguists identified the variation of action-oriented lexicons across different action concepts. 521 Italian verbs and 550 English verbs (i.e., the high-frequency verbal lexicon most likely to be used when referring to action) have been processed (Moneglia et al. 2012).

The corpus-based strategy relied on an induction process that separated the metaphorical and phraseological usages from those strictly referring to physical actions. IMAGACT only specifies the various possible interpretations of verbs with respect to physical actions, while ignoring the other interpretations. Therefore the possible interpretations of verbs beyond physical actions are not considered and are not represented in the ontology.

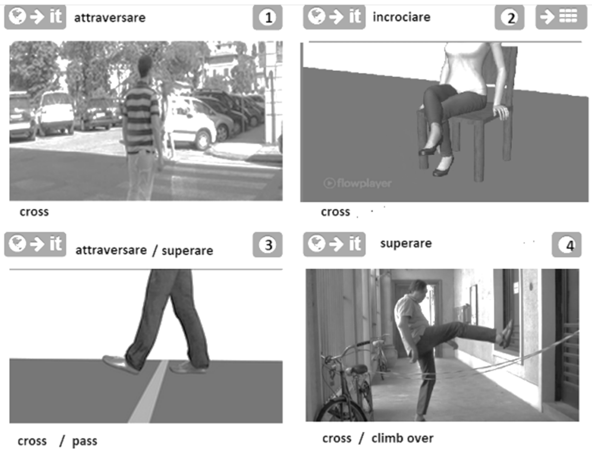


Figure 1. The four action types of the verb *to cross*

The unique visual format of the ontology makes the representation of abstract concepts difficult or impossible. This limitation, however, also constitutes an important added value, which can benefit our knowledge of action verbs in their abstract interpretations and the identification of these meanings within ontologies, as we will show in later sections of this paper.

The capacity to refer to many different physical activities with a single verb belongs to the core of the semantic competence of a language, which has been achieved by mother-tongue speakers during the early phases of their first language acquisition. A speaker cannot assert knowledge of the meaning of *cross* if he is not able to judge that the above events can be the object of its application. At the same time, despite the difference between the different actions represented in each concept, he will also be able to judge that none of them represents the meaning of the verb better than the others and that the verb is applied in its own meaning in all cases (primary meanings). This is not the case for metaphors, phraseology and abstract meanings.

For instance, the semantic competence of the speaker is not affected if she does not understand the meaning of "John crossed wires with Mary" (idiom) or "John needs to cross to another account" (metaphor). Competent speakers are, on the contrary, able to judge that in these cases the verb is not used in its physical meaning (marked meanings). Nonetheless, roughly half of corpus occurrences of action verbs are not used in their primary, physical meanings, and the use of verbal predication extended from physical meanings is one of the more productive means of reference in natural languages.

This paper describes the infrastructure that has been designed to deal with variations that do not fall within the primary, physical variation of an action verb. It will specifically highlight the benefits that IMAGACT's representation of the primary meanings through videos brings to the understanding and annotation of secondary meanings.

2. Processing Corpus Occurrences in IMAGACT and the Selection of Marked Variation

The construction of IMAGACT requires the examination and interpretation of verb occurrences in an oral context, which is frequently fragmented and may not provide enough semantic evidence for an immediate interpretation. To this end, the annotation infrastructure allows the annotator to read the context of the verbal occurrence in order to grasp the meaning. The annotator represents the referred meaning with a simple sentence in a standard form for easy processing. This sentence is positively formed, in the third person, present tense, active voice, with the essential arguments of the verb filled. Crucially, along with the standardization, the annotator assigns the occurrence to a "variation class", either PRIMARY or MARKED (Moneglia et al.2012).

The decision concerning the status of the occurrence makes use of an operational test roughly derived from Wittgenstein (1953). The occurrence is judged PRIMARY if it is possible to say to somebody who does not know the meaning of the verb V that "the referred action and similar events are what we intend with V"; otherwise the occurrence is MARKED. For instance, the occurrences standardized in "John crosses the finish line"; John crosses the street" and "John crosses his legs" are assigned to PRIMARY variation, since all can be pointed to explain "what cross means".

Conversely, the instances standardized as "a thought crossed John's mind" are not what one uses to instantiate the meaning of *to cross* and therefore have been tagged as MARKED. The annotation of primary versus marked variation has been evaluated at 9.5 K-Cohen agreement (Gagliardi 2014).

The positive selection of occurrences in which verbs refer in their own meaning to physical actions preceded the annotation of action concepts. Only occurrences assigned to the PRIMARY variation class make up the set of Action Types stored in the ontology. To this end, the standard IMAGACT infrastructure allows clustering of occurrences under prototypes representing the various action concepts, keeping granularity to its minimal level (8.2 K agreement [Gagliardi 2014]). The full annotation process can be found in Moneglia et al. 2012.

Concepts are represented using the universal language of images, which allows the reconciliation, in the IMAGACT ontology, of the types derived from the annotation of different language corpora. 1010 distinct action concepts have been identified and visually represented with prototypical scenes, either animated or filmed (Frontini et al. 2012; Moneglia et al. 2012). The cross-linguistic correspondences of those actions with the verbs that can refer to them in English and Italian have been established in a MYQL database.

38,462 occurrences have been processed in the English corpus and 42,723 in the Italian corpus. Respectively 19,229 and 16,210 (50% and 38%) have been considered marked.

3. Marked Variation Categories

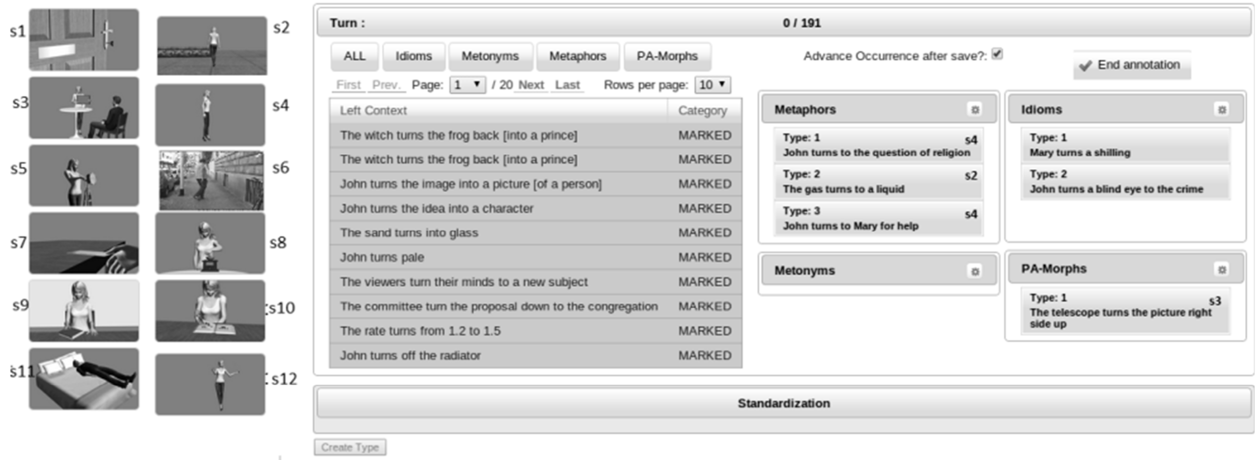
We have established three main categories of marked variation--*metaphor*, *metonymy* and *idiom*--and criteria for creating types within these categories for each verb. The criteria rely heavily on the images that compose the IMAGACT ontology of action and on widely accepted processes of meaning extension in linguistics. Although figurative language is known for its amorphous, subjective nature, we have endeavored to create a standard, justifiable process for determining figurative language types for individual verbs, that we will show in the following sections on the basis of the verbs *to turn* and *to close.*

3.1 Metaphor

The process for identifying a metaphoric type for a verb involves several steps and satisfying several related criteria. First we list all the occurrences of a verb that were labeled as "marked" during the initial corpus annotation of the IMAGACT project. We then use a standard lexicographic procedure of gathering similar usages together. For each group of occurrences that is a potential metaphor, we look for an image or "family" of related images from the IMAGACT ontology to which the occurrences are related. For example, the following list is a sample of one group of corpus occurrences for the verb *to turn:*

John turns to the question of religion
The presenter turns to [the subject of] the book
The colleagues turn to the report
The host turns to the other issues

We have linked this group to the S4 animated video from the IMAGACT ontology shown in Figure 2. The action is of a woman facing straight ahead then turning her head to the right.



Left Context	Category
The witch turns the frog back [into a prince]	MARKED
The witch turns the frog back [into a prince]	MARKED
John turns the image into a picture [of a person]	MARKED
John turns the idea into a character	MARKED
The sand turns into glass	MARKED
John turns pale	MARKED
The viewers turn their minds to a new subject	MARKED
The committee turn the proposal down to the congregation	MARKED
The rate turns from 1.2 to 1.5	MARKED
John turns off the radiator	MARKED

Metaphors

Type: 1 s4
John turns to the question of religion

Type: 2 s2
The gas turns to a liquid

Type: 3 s4
John turns to Mary for help

Idioms

Type: 1
Mary turns a shilling

Type: 2
John turns a blind eye to the crime

Metonyms

PA-Morphs

Type: 1 s3
The telescope turns the picture right side up

Figure 2. Interface sample for *to turn*

The next step is to identify the property of the action that affords the extension of the verb to a more abstract domain. In this video, the actor turning her head now sees whatever is to her right rather than whatever is directly in front of her. This physical turning of her head can indicate a change in the focus of her attention, say, from a street in front of her to a dog barking on her right. With a metaphorical extension, *to turn* can be used to indicate a change in the focus of one's attention to abstract things as well, such as the question of religion.

One of the most influential theories of metaphor has been that of conceptual metaphor (Lakoff 1987), which posits that a fundamental mechanism of human cognition is the use of a concrete, physical domain to understand a more abstract one. These conceptual metaphors are often revealed in a group of related lexical metaphors. For example, the conceptual metaphor Life is a Journey can be seen in the sentences "Mary needs to move on after her divorce" and "the governor ran into a political road block." Where it is possible, we identify the general conceptual metaphor that supports the specific linguistic metaphor in question. Using the list of conceptual metaphors maintained by the University of California, Berkeley, we linked the *turn* metaphor just described to the conceptual metaphors *Change is Motion* and *Ideas are Locations.* Thus, a person facing one location (idea) can turn to face another, indicating a change in her attention from one idea to another.

As with the identification of primary, physical types in IMAGACT, we use equivalent verbs to help distinguish metaphorical types. For the marked variation, we distinguish between equivalent verbs that are used in their primary, or non-figurative, meaning and equivalent verbs that are used in a marked or figurative sense. For example, the verb *shift* has been identified as a verb that can be used in the same situations as *turn* in "John turns to the question of religion." Both of these verbs are used metaphorically in this situation, with the same metaphorical meaning. This match is relevant for an ontology of abstract concepts and corresponds to action concepts in the IMAGACT database.

However, the key means of distinguishing types within the category of metaphor are the links to the action concepts they derive from and the descriptions of the relevant properties that license the metaphorical extensions. Often, links to different action concepts are enough to distinguish two marked types of a verb. For example, "John turns to the question of religion" is linked to type S4, as described above. Another very common metaphor for the verb *to turn* refers to a change of state, such as "the witch turns the frog back into a prince" or "the gas turns to a liquid". The metaphor is linked to the action concept represented by the video in S2. As part of the conceptual metaphor Change of State is Change of Direction, the linguistic metaphor for *turn* in this case uses the property of moving in a new direction from a different action concept and image than the previous *turn* metaphor.

Sometimes two or more metaphors derive from the same action concept but rely on different properties of that concept. Another metaphor of *to turn* links to the S4 image in Figure 2: "John turns to Mary for answers" or "Mary turns to a psychiatrist". In this case, the reorientation of the actor's head indicates an appeal for interaction rather than a change in the focus of his attention. Identifying the prototype related to the metaphor helps in understanding the properties that license the metaphoric extension.

3.2 Metonymy

Metonymy is a less studied phenomenon than metaphor, especially as it pertains to verbs. However, the corpus data we have gathered suggests that it is a necessary category to fully account for the marked variation of certain verbs. For our purposes, we have defined verb metonymy as the use of one action or event to represent a sequence or set of events of which it is a part. For example, many occurrences of *to close* in our English corpus follow the form of "John closed the pub" and "The management closed the factory." This usage of *close* does not follow the process of metaphorical extension, in which an abstract domain is being understood using properties from a physical one. There are actual actions of closing involved in the situations described by these sentences. When John closes the pub, he does indeed close the door. He probably also takes the cash from the register, turns off the lights, and locks the door as he leaves. This is not a physical domain being used to understand an abstract one, but one action in a sequence of events being used to represent the whole sequence (Goossens 1995).

Complicating the situation, the events in such a sequence are not always all physical actions. “The management closed the plant” probably is also meant to include the decision to end production at the plant, as well as the action of closing and locking the doors. For our purposes, as long as part of the whole event can be described using the verb in its physical sense, we have categorized that type as a metonymic one.[1]

For this category, we also link the type to an image from the action ontology. The type of *close* described previously is linked to the video in Figure 3. We also identify one or more equivalent verbs. As with metaphor, where the equivalent verbs are usually other verbs used in a metaphorical way, the equivalent verbs for metonymic types are often other verbs being used metonymically. For example, *shut* is the equivalent verb for this type.

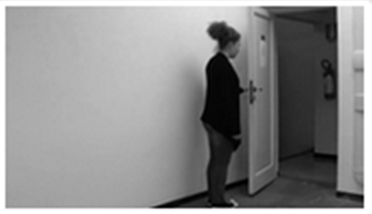


Figure 3. Action type for *to close*

3.3 Idiom

We use a standard definition of idiom: a fixed phrase whose meaning cannot be deduced by combining the meanings of the individual words in the phrase. Because idioms are usually language specific, we have not attempted to link any idioms to the language-independent action concepts in IMAGACT. Instead, we identify an equivalent verb, along with a specific synset in WordNet. For instance, we identify the idiom “turn a deaf ear to” with the equivalent verb *ignore*, connected to the WordNet synset *[neglect, ignore, disregard].*

4. Ongoing Work

We have tested our categories and criteria against the full set of corpus occurrences for five verbs *(turn, cross, pull, close, combine),* creating types to account for all the occurrences. Although this exercise has largely supported the applicability of our schema, it has also raised some questions that we are still in the process of resolving. For some highly frequent verbs, like *to turn,* we find a few, very common marked types. For others, like *to pull,* we find a myriad of different marked types, many of which occur only once or twice in the corpus. How to efficiently account for these rare types remains an open question.

We have also discovered verbs with marked usages that do not seem fit into any of our three categories, such as *Mary received the wire transfer*. In these cases the verbs appear to have the same meaning as one of the primary, physical types for that verb, but to be acting on objects that are not strictly physical. We are in the process of evaluating a fourth type to account for these usages.

We plan to evaluate further our marked categories and methodology for type creation by annotating the full set of corpus occurrences for a larger set of action verbs from the IMAGACT ontology, a set that includes verbs taken from each of the upper level nodes of the ontology. Based on the results, we will finalize the annotation interface, then use it to process all of the marked occurrences identified by the original IMAGACT annotation. We anticipate supplementary annotation to account for thematic roles and the possible regularities among types that they may reveal (Brown & Palmer 2012). We expect this work to lead to a rich study of the relation between the marked and primary types of high-frequency verbs.

5. References

Conceptual Metaphor Home Page. http://www.lang.osaka-u.ac.jp/~sugimoto/MasterMetaphorList/MetaphorHome.html

IMAGACT. http://www.imagact.it

Brown S. W., & Palmer M. (2012). Semantic annotation of metaphorical verbs with VerbNet. Interoperable Semantic Annotation (ISA-8), 8th Joint ISO-ACL/SIGSEM Workshop, Pisa, October.

Frontini, F., De Felice, I., Khan, F., Russo, I., Monachini, M., Gagliardi, G. & Panunzi, A. (2012). Verb interpretation for basic action types: Annotation, ontology induction and creation of prototypical scenes. CogAlex-III Workshop as part of the COLING 2012 conference. Mumbai (India), December.

Gagliardi, G. (2014). Validazione dell’ontologia dell’azione IMAGACT per lo studio e la diagnostica del “mild cognitive impedirment” (MCI). PhD Dissertation, University of Florence.

Goossens, L. (1995). Metaphtonymy: The interaction of metaphor and metonymy in expressions of linguistic action. In L.Goossens, P. Pauwels, B. Rudzka-Ostyn, A. Simon-Vanderbergen & J. Vanparys (eds), *By Word of Mouth.* Amsterdam: John Benjamins, pp. 159--174.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago: University of Chicago Press.

Moneglia, M. (in press). Natural language ontology of action. A gap with huge consequences for natural language understanding and machine translation. In Z. Vetulani, J. Mariani (eds), *Post LTC 2011 LNAI,* Berlin: Springer.

Moneglia, M., Gagliardi, G., Panunzi, A., Frontini, F., Russo, I. & Monachini, M. (2012). IMAGACT: Deriving an action ontology from spoken corpora. 8th Joint ISO-ACL/SIGSEM Workshop, Pisa, October.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford: Blackwell.

[1] In some cases, a metonymic use of a verb seems to have been further extended into a metaphor. Rather than create a complex annotation scheme where categories can interact, we have provisionally decided to treat these as metaphors.

Annotating Cohesion for Multilingual Analysis

Ekaterina Lapshinova-Koltunski*, Kerstin Anna Kunz**

*Saarland University
Universität Campus A2.2,
66123 Saarbrücken
e.lapshinova@mx.uni-saarland.de
**University of Heidelberg
Grabengasse 1, 69117 Heidelberg
kerstin.kunz@iued.uni-heidelberg.de

Abstract

This paper describes a set of procedures used to semi-automatically annotate a multilingual corpus on the level of cohesion, an important linguistic component of effectively organised and meaningful discourse. The annotation categories we operate with base on different degrees of granularity and account for lexico-grammatical and semantic aspects of different types of cohesion. This annotation scheme allows us to compare and differentiate cohesive features across languages, text types and in written and spoken discourse on different levels of abstraction. Our aim is to obtain a fine-grained and highly precise annotation, at the same time avoiding purely manual annotation. Therefore, we decide for corpus-based semi-automatic procedures to identify candidates expressing cohesion in English and in German. The annotated corpus is one of the few existing resources supporting contrastive studies of cohesion.

Keywords: cohesion, discourse relations, annotation, corpora

1. Introduction

Cohesion is an important component of effectively organised and meaningful discourse, as the message being communicated in discourse is not just a set of clauses, but forms a unified, coherent whole. While coherence concerns the cognitive aspects of establishing meaning relations during text processing, cohesion involves explicit linguistic means that signal how clauses and sentences are linked link together to function as a whole. Both concepts have been studied in a range of disciplines, including philology, sociology, philosophy, psychology, computer science and linguistics. The latter analyses inventories of the linguistic markers that are available in a given language, see (Louwerse and Graesser, 2005). Classifications of lexico-grammatical markers and their relational potentials are quite often language specific, cf. (Halliday and Hasan, 1976; De Beaugrande and Dressler, 1981; Brinker, 2005), etc. For multilingual analysis, e.g. contrastive linguistics or translation (both human and machine) studies, it is important to establish categories which enable the comparison of inventories across languages in order to identify commonalities and contrasts. Complex annotations on higher linguistic levels which are geared towards high precision are typically carried out manually and hence, are very time-consuming. To our knowledge, existing resources provide annotations of individual cohesive phenomena only, e.g. pronominal coreference in the BBN Pronoun Coreference and Entity Type Corpus, (Weischedel and Brunstein, 2005), verbal phrase ellipsis in (Bos and Spenader, 2011) or conjunctive relations in PDTB, (Prasad et al., 2008), annotation of (abstract) anaphora in (Dipper and Zinsmeister, 2009) and (Dipper et al., 2012). Most of them are monolingual and apply manual annotation procedures only.

In the present paper, we suggest procedures to semi-automatically identify and annotate cohesive phenomena.

2. Motivation and Theoretical Background

As already mentioned above, cohesion plays an important role in discourse organisation and coherence. Our main interest lies in comparing the realisation of cohesive strategies in different languages and also in written and spoken text types via empirical methods. Therefore, one major challenge is to define categories that enable identification of commonalities and differences in terms of various cohesive aspects.

Our concept of cohesion is based on Halliday and Hasan's definition as relations of meaning that exist within the text, and that define it as a text, see (Halliday and Hasan, 1976). Hence, our long-term focus is on the investigation of the semantic or conceptual relation as such. Cohesive relations, however, require a linguistic trigger, a cohesive device which explicitly signals that there is a relation to another textual expression. These devices can be grammar- or vocabulary-driven. As claimed by (Louwerse and Graesser, 2005), grammar-driven cohesion refers to the semantic reduction of expressions to functional items which are syntactically obligatory, such as proforms. Vocabulary-driven cohesion refers to the lexical vocabulary of the discourse segment. Halliday and Hasan (1976) describe five main types of cohesion in English, for which we adopt for our multilingual analysis: *reference*, *substitution*, *ellipsis*, *conjunction* and *lexical cohesion*. Although their classification claims to be mainly semantic, it is influenced by lexico-grammatical patterns that reflect systemic features of English. We therefore attempt a more conceptual classification which is suitable for the comparison of English and German:

Reference involves identity between instantiated referents/entities, as in example (1). Substitution/ellipsis expresses similarity between different instantiated referents/entities of the same type, see examples (2) and (3) respectively. Conjunction concerns the logico-semantic relations between propositions (e.g. addition, contrast, cause)

– see example (4). Lexical cohesion includes similarity between referents/entities of the same type which bases on sense relations between lexical items (e.g. hypernymy, part-whole relations), as in example (5).

(1) a. Wir arbeiten für Wohlstand und Chancen, weil *das* richtig ist. Wir tun *damit* das Richtige.
b. We work for prosperity and opportunity because *they*'re right. *It*'s the right thing to do.

(2) a. Das war ein Problem. Aber *keins*, mit dem ich mich auseinandersetzen wollte.
b. This was a problem. But not *one* I chose to deal with.

(3) a. Who says that? – My parents ⊗.
b. Wer sagt das? – Meine Eltern ⊗.

(4) a. Sie wollen ein starkes Europa in der Welt. *Deshalb* hat Großbritannien eine europäische Sicherheitspolitik mit auf den Weg gebracht.
b. They want Europe to be strong in the world. *That's why* Britain has helped launch a European security policy.

(5) a. Vor allem müssen die *Entwicklungsbanken* ihre Bestrebungen auf... konzentrieren. Als erstes sollten die *Banken* mehr Ressourcen für die Entwicklung von Humankapital aufwenden.
b. First and foremost, the *development banks* must focus their efforts... To start, the *banks* should devote more resources to the development of human capital.

According to Halliday and Hasan (1976), what distinguishes cohesive relations from other semantic relations is that the lexico-grammatical resources, i.e. the cohesive devices, trigger relations that transcend the boundaries of the clause.

We argue that these semantic relations are realised in both languages under investigation and also across text types including written and spoken discourse. Systemic comparisons, e.g. (Kunz and Steiner, 2012; Kunz and Steiner, 2013; Kunz and Lapshinova-Koltunski, in press), have shown that they differ in terms of lexico-grammatical patterns of realisation, as can be see from the examples above. In addition, we suggest textual contrasts in the frequency of cohesive devices, in types of preferred cohesive relations, in the strength of the cohesive relation, as well as in the breadth of variation.

Starting from these considerations, we formulate subcategories of the five phenomena of cohesion defined above, reflecting the lexico-grammatical and semantic features of the *cohesive devices* that establish these types. Only those categories are defined which are applicable for both English and German, see table 1.

Our analysis also includes *cohesive relations*, which are often described as relations across grammatical domains, e.g. in (Halliday and Hasan, 1976; Eckert and Strube, 2000; Ariel, 2001; Kunz, 2009). Here, we define two categories: *coreference* and *lexical chains*. They both involve a textual relation that is created between linguistic expressions.

reference	
type	**function**
personal	head, modifier, *it*-endophoric
demonstrative	head, modifier, local, temporal, pronominal adverb
comparative	particular, general
conjunction	
synt.type	**sem.type**
connects, subjuncts, adverbials	additive, adversative, causal, temporal, modal
substitution	
nominal, verbal, clausal	
ellipsis	
nominal, verbal, clausal	
lexical cohesion	
general nouns, collocations	

Table 1: Cohesive devices and their functions

The textual relation of coreference evokes a conceptual relation of identity between discourse referents/entities (see above). A coreference relation links at least two coreferring expressions: an antecedent, i.e. a linguistic element introducing a new discourse referent, and a cohesive device of reference which functions as an anaphora (or cataphora, in the case of forward reference) and which points to the same referent again. The cohesive device of reference serves as a linguistic marker which triggers a search instruction to its antecedent e.g. a semantically weak proform or a deictic element. We include all categories defined for reference (see table 1) for the analysis of anaphoras. As several anaphoras may point to the same antecedent, several textual relations may be created for one referent in the same discourse, hence a *coreference chain* is the set of all coreferring expressions which refer to the same antecedent. The same applies to lexical cohesion, although the meaning relation established is a different one (see above): a lexical chain contains at least two lexical expressions in different textual parts which are linked by a semantic relation of hypernymy (e.g. a specific noun linked to a general noun), meronymy, synonymy, etc. or by repetition of the lexical base. Again, the chain may contain more elements and hence also semantic relations, which tie textual referents that belong to the same experiential or semantic domain.

3. Corpus resources

The multilingual corpus we annotate offers a continuum of different text types (registers) from written to spoken discourse. More precisely, it includes English and German texts of ten registers, eight of which represent written discourse and include fictional texts (FICTION), political essays (ESSAY), instruction manuals (INSTR), popular-scientific texts (POPSCI), letters to shareholders (SHARE), prepared political speeches (SPEECH), tourism leaflets (TOU) and corporate websites (WEB). This part was imported from the existing corpus CroCo described in

(Hansen-Schirra et al., 2013). The written texts are saved in two subcorpora according to the language: English written texts (EO), German written texts (GO). The other registers are of spoken discourse and include recorded and transcribed interviews, as well as academic speeches, see (Lapshinova-Koltunski et al., 2012). The spoken texts are also stored in two subcorpora classified according to the language of origin: English spoken texts (EO-SPOKEN) and German spoken texts (GO-SPOKEN). The whole number of words contained in the corpus comprise ca. 730 thousand words (see table 2, although not big, but still provides a usefull data set for annotating and analysing cohesion in both languages.

register	**EO**	**GO**
ACADEMIC	40443	40986
FICTION	36996	36778
ESSAY	34998	35668
INTERVIEW	37901	40198
INSTR	36167	36880
POPSCI	35148	36178
SHARE	35824	35235
SPEECH	35062	35399
TOU	35907	36574
WEB	36119	35779
TOTAL	364565	369675

Table 2: Corpus constellation and size

The corpus is pre-annotated on several levels, which include information on tokens, lemmas, morpho-syntactic features (e.g. case, number, etc.), parts-of-speech, phrase chunks and their grammatical functions, as well as and sentence boundaries. The annotation of the written part was partly imported from CroCo, whereas for the spoken part, we use Stanford POS Tagger (Toutanova et al., 2003) and the Stanford Parser (Klein and Manning, 2003). The corpus is encoded in the CWB format (CWB, 2010) and can be queried with Corpus Query Processor (CQP) (Evert, 2005). These annotation levels allow us to obtain additional information on cohesive phenomena and cohesive relations, i.e. coreference: morpho-syntactic preferences of antecedents and anaphoras, their positions in a clause, the length of chains in terms of elements, diversity of types of antecedents, their parallelism with anaphoras, etc. Furthermore, these annotation levels provide the basis for the semi-automatic procedures described in the present paper.

4. Annotation of Cohesion

4.1. Categories to annotate

In the following, we provide a more detailed description of the annotation scheme based on the categories introduced in 2. above. Note that this mainly concerns the classifications of cohesive devices, which, however, builds the basis for the analysis of cohesive relations (see below). Main categories exist for the main cohesive types reference, conjunction, substitution, ellipsis and conjunction. We distinguish subtypes, which are annotated as 'type' or 'func' feature in the corpus. They reflect general structural groupings of cohesive devices that exist in both languages. These categories, as well as their language realisations (operationalisations) are presented in table 3.

Each subcategory of reference (type) is further subclassified according to grammatical and semantic features of the cohesive device (func). Personal reference includes personal (head) and possessive (modifier) pronouns as well as their morphological variants. For this type, we also annotate reference by *it/es* separately (it-endophoric) due to the ambiguity of their usage in both languages. Demonstrative reference is expressed by means of demonstrative pronouns (head) and determiners (modifier), as well as their morphological variants (in German). Moreover, we include local and temporal relations of identity, which are expressed by adverbs (see table 3) as well as pronominal adverbs (pronadv). These exist in English and German but are employed in German with a higher frequency. Comparative reference is expressed with comparative forms of adjectives, which either trigger a general relation of comparison or a more specific one (particular).

Conjunction is classified in terms of main syntactical types: coordinating conjunctions (connects), subordinating conjunctions (subjuncts) and discourse adverbials (adverbials). They may consist of one or multi-word constructions of conjunctions, e.g. *that is why*, etc. see table 3. For each syntactical subcategory we provide the same semantic subclassifications, according to the main logico-semantic relations that can be established by conjunctive devices.

Both in English and in German, substitution is expressed by indefinite pronouns or other nominal substitutes (nominal), substituting verbs (verbal) and different adverbials, which substitute clausal constructions, such as *so* in English (clausal). Ellipsis can be triggered by different lexico-grammatical means in both languages, and therefore, automatic detection still remains problematic. Nevertheless groupings can be made in terms of which structural elements are mainly omitted in relation to the preceding full textual structure. Again main categories here are nominal, verbal and clausal. Substitution and ellipsis cannot be categorised in terms of other features since their language-specific features do not allow a common subclassification.

For the time being, only two aspects of lexical cohesion are categorised: Textual relations that base on the use of general nouns, for which we use lists of nouns based on those described by (Dipper et al., 2012) and repetitions of lexical bases. We plan to integrate sense relations such as hypernymy and synonymy in the future.

4.2. Annotation of Cohesive Devices

Automatic procedures To annotate the categories presented in 4.1., we elaborate a set of semi-automatic procedures, which involve an iterative extraction-annotation process. We use a method derived from the system used for the YAC chunker, see (Kermes and Evert, 2002; Kermes, 2003). The system is based on the option of the CWB tools to incrementally enhance corpus annotations, as query results deliver not only concordances of the searched structures but also information on their corpus positions. The algorithm makes use of the CWB Perl-Modules to access CQP and the encoding functionality using Perl-scripts as wrapper. Additionally, Perl modules are derived from the

device	type	func	realisation
reference	personal	head	*he/er, she/sie, they/sie*, etc.
		modifier	*her/ihr, his/sein, their/ihr*, etc.
		it-endophoric	*it/es*
	demonstrative	head	*this/dies/das, that/jenes*, etc.
		modifier	*this/diese(r/s), that/jene(r/s)*, etc.
		local	*here/hier, there/da*, etc.
		temporal	*now/jetzt, then/dann*, etc.
		pronadv	*herewith/hiermit, dagegen, damit*, etc.
	comparative	particular	*bigger/grösser, better/besser*
		general	*other/andere, such/solche*
conjunction	connects		*and/und, or/oder*, etc.
	subjuncts		*although/obwohl, where/wo*, etc.
	adverbials		*also/auch, finally/endlich*, etc.
		additive	*and/und, for example/zum Beispiel*, etc.
		adversative	*however/allerdings, in contrast/im Gegensatz*, etc.
		causal	*that is why/weshalb, therefore/deswegen*, etc.
		temporal	*then/dann, first/erstens*, etc.
		modal	*interestingly/interessanterweise, og course/natürlich*, etc.
substitution	nominal		*those/welche, one/eine*, etc.
	verbal		*do/tun*
	clausal		*so/so/dergleichen*
ellipsis	nominal verbal clausal		various triggers
lex. cohesion	general nouns		*problem/Problem, situation/Situation, position/Position*, etc.

Table 3: Annotated categories of cohesion

framework of YAC which facilitate the annotation of information gathered using CQP queries. This permits to import the information on queried data back into the corpus. In this way, our annotation rules are defined in form of CQP queries that allow regular expressions based on string, parts-of-speech, chunk and further constraints.

Each query is applied to the corpus separately. The result is a list of corpus positions indicating the start and the end, and possibly a target position marked within the query. These corpus positions can now be used to extract additional information already encoded in the corpus (e.g., part-of-speech tags, lemma information, sentence position, etc.). If needed, this information can be evaluated against lists in order to classify or exclude them. Finally, the results and possibly additional information can be encoded using the corpus positions as anchors.

In table 4, we demonstrate examples of CQP queries to extract and annotate reference. Query 1 is designed to extract textual instances of local demonstrative reference, whereas query 2 delivers occurrences of demonstrative reference with the grammatical function of a modifier. The results do not need further processing within the annotation process, as the categorisation is encoded in the query itself. The instances are annotated as XML structures with attributes ’type’ and ’func’, where ’type’ is demonstrative and ’func’ is either temporal or modifier respectively. The tags are then imported back into the corpus and saved as CQP structural attributes.

Further two queries are built to extract semantic (query 3) or syntactic (query 4) types of conjunctions. In the final step, we combine both annotations (syntactic and semantic) to exclude non-cohesive occurrences of conjunctions, e.g. in case they link phrases and not clauses, as in the sentence in example (6).

	QP query	example of XML tags
1	<chunk> ([_.chunk_gf=”adv_temp”]+| [word=”then|now” &pos=”rt”])	<reference type=”dem” func=”temporal”> *now* </reference>
2	[lemma=”this|these|those|that” &pos=”dd.*”] [pos=”j.*|n.*|mc|vvg|md”]	<reference type=”dem” func=”modifier”> *this* </reference>
3	[lemma=RE($additive)]	<conj func=”additive”> *in addition*</conj>
4	[_.chunk_gf=”adv.*” pos=”koui|kous|pw.*|appr”]	<conj type=”subjunct”> *although*</conj>

Table 4: Examples of queries and tagged structures in XML

(6) Renewables 2004 will focus on renewables *and* aim at strengthening the political momentum.

Classification of semantic types proceeds directly within the query which includes a simple lexical search – here, we aim to identify all cohesive instances of ’additive’ conjunctions (a closed class of lexical items the members of which we know). The same procedure (based on lexical list) is applied to identify and annotate general nouns.

Manual procedures As our aim is to produce a corpus with highly precise information on cohesive devices in English and German, we integrate a step of manual correction into our procedures. To facilitate this, the annotated corpus (with the structures in XML format as shown in table 4 above) is imported into MMAX2, a tool for manual annotation (Müller and Strube, 2006). Texts are corrected by at least two human annotators with linguistic background. The MMAX2 visualisation allows annotators to decide whether the candidates tagged by the automatic procedures have a cohesive function and belong to the given category. We also add an option to mark the cases as 'problematic' or 'non-problematic' to trace and analyse the reasons for annotators' hesitation in case of a low inter-annotator agreement. This combination of automatic pre-annotation with manual post-correction is less time-consuming for human annotators as annotating raw texts. Moreover, we achieve positive results in the inter-annotator agreement (see below in this section).

Correction by human annotators allows us, on the one hand, to improve both annotations and rules for automatic procedures (rule-based identification of items can be improved on the basis of human annotators' observations), and on the other hand, to evaluate automatic procedures.

Evaluation Our preliminary results show that in the automatic identification of cohesive devices, we are able to achieve a good precision for English (between 76% and 98%) and slightly lower precision for German (between 69% and 73%), shown in table 5. The lower results for German are partially caused by the multi-functionality of the lexico-grammatical means expressing cohesion[1]. In addition, higher flexibility of ordering clausal constituents in German complicates automatic disambiguation of cohesive and non-cohesive forms on the basis of syntactic rules. In terms of recall, we are also able to achieve satisfactory results for English, e.g. 80% for reference and 73% for conjunction, and lower results for German: 60% for reference and 71% for conjunctions.

	EO	GO
reference	0.98	0.73
conjunction	0.76	0.69
substitution	0.84	0.71

Table 5: Precision of automatic procedures

We also calculate the inter-annotator agreement 1) between human annotators (HuHu) and 2) between human annotators and automatic procedures (HuAut), see table 6. The best scores are again observed for English in the agreement between humans and the automatic system. For German, the score is lower. However, the agreement between human annotators is slightly higher in the annotation of German. This can be explained, again, by the complexity of German lexico-grammatical means expressing cohesion.

	HuHu	HuAut
EO	0.74	0.85
GO	0.78	0.66

Table 6: Inter-annotator agreement for reference

Annotation procedures are especially problematic in spoken registers, where cohesive devices are much more frequent as in written registers, see figure 1 in section 4.3. below. Spoken discourse is characterised by numerous repairs, ellipses, unclear sentence breaks and therefore. Cohesive and non-cohesive instances cannot be easily diambiguated as sentences boundaries do not play a role in spoken discourse. This all poses a real challenge for both semi-automatic and manual annotation.

language	register	precision	recall	*F*
EO	**ACADEMIC**	0.88	0.71	0.79
	INTERVIEW	0.81	0.68	0.74
	TOTAL	**0.85**	**0.70**	**0.77**
GO	**ACADEMIC**	0.49	0.60	0.54
	INTERVIEW	0.68	0.62	0.74
	TOTAL	**0.59**	**0.62**	**0.60**

Table 7: Evaluation of procedures in spoken registers

We calculate precision and recall for automatic reference annotation in our spoken data. As seen from table 7, the overall results for English spoken registers are better than for German. Interestingly, the register-specific results differ in both languages. Whereas in English the system performs better on academic speeches (which are mostly monologic), interviews are annotated with less errors than academic speeches in German.

4.3. Annotation of Coreference

For the annotation of coreference, we use the output of the semi-automatic procedures described in 4.2. above and manual annotations produced by humans. To our knowledge, none of the existing automatic procedures can fit our tasks, as most of them operate with a limited set of categories. Moreover, previous works on coreference annotation for spoken discourse, have shown that the available systems can achieve around 60% for written and ca. 50% for spoken texts, see, for instance, (Amoia et al., 2012) for the analyses with Stanford CoreNLP (Lee et al., 2011).

Therefore, we decide for manual annotation of coreference chains, which includes manual identification of antecedents by human annotators, and their linking to the cohesive devices (anaphoras) which were automatically tagged by our system described in 4.2. above, and manually corrected by human annotators. Here again, we use MMAX2 to facilitate the annotations, as this tool allows visualisation of links between two or more elements.

The annotated information is then encoded as an additional attribute of 'mention', which is automatically provided with an identification number (id). Every expression referring to the same antecedent is also assigned with the same id. This information is saved for every text, and then imported into the corpus. The information on the chains can then be extracted with the help of these ids. The information on the type and function of the referring expression is also integrated into this new structure, see figure 1.

In the example presented in figure 1, the items indexed with

[1] See, for example, (Lapshinova-Koltunski and Kunz, in press) for the examples of errors in annotation of conjunctive relations.

```
<mention chain_id="set_1">Dr
Hales</mention> received his M-S
and B-S degrees at Stanford in
nineteen eighty-two.  <mention
chain_id="set_1" type="pers"
func="modifier">his</mention>
<mention chain_id="set_2">PhD at
Princeton in eighty-six</mention>
under the Harold W Dodds Honorific
Fellowship, <mention chain_id="set_1"
type="pers" func="head">he</mention>
<mention chain_id="set_2" type="dem"
func="temporal">then</mention>
went on to the Mathematical
Sciences Research Institute to do
post-doctoral research.  and <mention
chain_id="set_2">then</mention> to
Harvard, where <mention chain_id="set_1"
type="pers" func="head">he</mention>
was an assistant professor for two
years under the National Science
Foundation Fellowship.  <mention
chain_id="set_1" type="pers"
func="head">he</mention> completed
the post-doctoral research fellowship
at the Institute for Advanced Study in
the following year.
```

Figure 1: Annotated coreference chains in the corpus

'set_1' belong to a longer chain. Four anaphoras refer to the same antecedent, which is 'Dr Hales'.

Lexical chains have not yet been annotated in the corpus. However, we aim to use the annotation of general nouns, as well as repetitions of lexical bases, and integrate semantic relations with the help of available resources, e.g. WordNet, see Fellbaum (1998). These automatic annotations will then be corrected in terms of cohesiveness by human annotators.

4.4. Annotation Availability

The annotated corpus is available in XML format and can be queried with CQP. We also provide a CQP-WEB[2] version which is available via CLAIN-D project.

5. Conclusion and future work

In the present paper, we have described semi-automatic corpus-based procedures to annotate cohesive types of (co-)reference, substitution, ellipsis, conjunction and lexical cohesion. These procedures allow both automatic identification of cohesive devices and their automatic annotation, which builds the basis for further annotation of semantic relations. Moreover, the integrated procedure of manual correction enables evaluation and improvement of the automatic procedures. Furthermore, they provide a possibility of consistent annotation on the basis of the pre-defined rules, which cannot be ensured if the entire annotation is of manual character. Our procedures concern two Germanic languages only, which have many common or comparable categories. Therefore, it would be interesting to test the proposed approach on another language pair including languages that belong to different language families. However, this is beyond the scope of the present research project.

The enriched corpus facilitates analysis of German vs. English contrasts, providing information on cohesive phenomena in both languages. Moreover, the availability of spoken material in our corpus allows the analysis of differences which result from differing conditions of speech, such as strong relation to the communication situation, direct interaction of speech participants and constraints on cognitive processing. First findings from our analyses show that mode of production plays an essential role for the grouping of particular registers in the two languages separately, and also across languages. For instance, the spoken registers in both languages exhibit a tendency towards marking important entities, comparing and evaluating them via cohesive relations. Their lexico-grammatical realisations are partially language-specific. Furthermore, we observe greater variation between written and spoken registers than in English, which may find further support in the future, when more spoken registers containing speaker turns are integrated in our corpus.

Such a resource is valuable not only for contrastive linguistics, but also for translation study, including machine translation, as well as further areas of NLP, e.g. automatic coreference resolution. The empirical data obtained from these annotations can be interpreted in terms of various linguistic aspects on different levels of granularity. It can thus be employed for further investigation and interpretation on semantic and conceptual levels of abstraction.

6. References

M. Amoia, K. Kunz, and E. Lapshinova-Koltunski. 2012. Coreference in Spoken vs. Written Text: a Corpus-based Analysis. In *Proceedings of the the 8th international conference on Language Resources and Evaluation*.

M. Ariel. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schliperoord, and Wilbert Spooren, editors, *Text representation*, Human cognitive processing series, pages 29–87. John Benjamins.

J. Bos and J. Spenader. 2011. An annotated corpus for the analysis of VP ellipsis. *Language Resources and Evaluation*, 45(4):463–494.

K. Brinker. 2005. *Linguistische Textanalyse: Eine Einfhrung in Grundbegriffe und Methoden*. Erich Schmidt, Berlin. 6 edition.

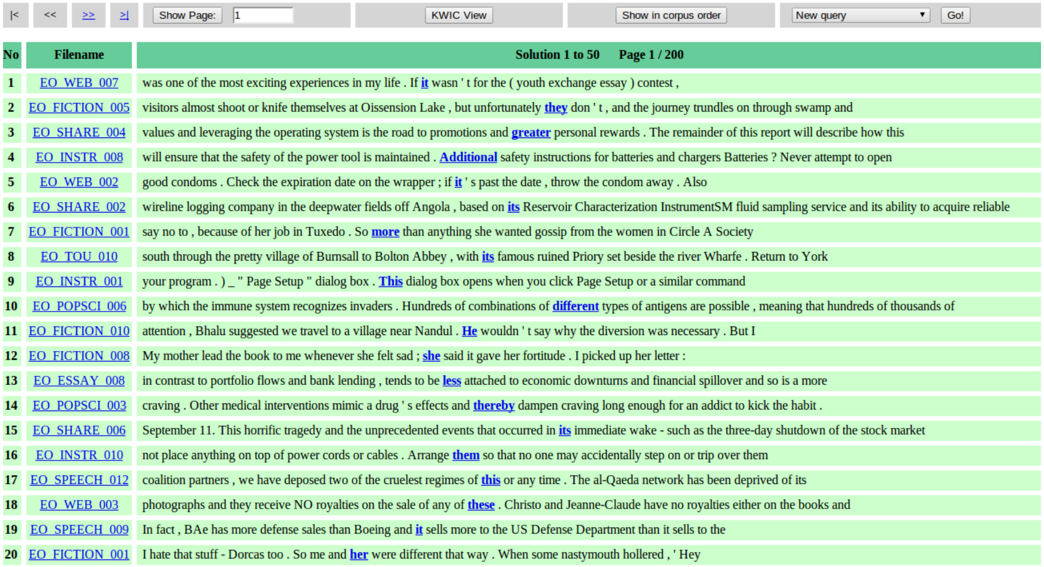
2010. The IMS Open Corpus Workbench. http://www.cwb.sourceforge.net.

R. De Beaugrande and W. U. Dressler. 1981. *Introduction to text linguistics / Robert-Alain de Beaugrande, Wolfgang Ulrich Dressler*. Longman, London ; New York.

S. Dipper and H. Zinsmeister. 2009. Annotating Discourse Anaphora. In *Linguistic Annotation Workshop*, pages 166–169. The Association for Computer Linguistics.

S. Dipper, M. Seiss, and H. Zinsmeister. 2012. The Use of Parallel and Comparable Data for Analysis of Abstract

[2] cf. (Hardie, 2012)



No	Filename	Solution 1 to 50 Page 1 / 200
1	EO_WEB_007	was one of the most exciting experiences in my life . If **it** wasn ' t for the (youth exchange essay) contest ,
2	EO_FICTION_005	visitors almost shoot or knife themselves at Oissension Lake , but unfortunately **they** don ' t , and the journey trundles on through swamp and
3	EO_SHARE_004	values and leveraging the operating system is the road to promotions and **greater** personal rewards . The remainder of this report will describe how this
4	EO_INSTR_008	will ensure that the safety of the power tool is maintained . **Additional** safety instructions for batteries and chargers Batteries ? Never attempt to open
5	EO_WEB_002	good condoms . Check the expiration date on the wrapper ; if **it** ' s past the date , throw the condom away . Also
6	EO_SHARE_002	wireline logging company in the deepwater fields off Angola , based on **its** Reservoir Characterization InstrumentSM fluid sampling service and its ability to acquire reliable
7	EO_FICTION_001	say no to , because of her job in Tuxedo . So **more** than anything she wanted gossip from the women in Circle A Society
8	EO_TOU_010	south through the pretty village of Burnsall to Bolton Abbey , with **its** famous ruined Priory set beside the river Wharfe . Return to York
9	EO_INSTR_001	your program .) _ " Page Setup " dialog box . **This** dialog box opens when you click Page Setup or a similar command
10	EO_POPSCI_006	by which the immune system recognizes invaders . Hundreds of combinations of **different** types of antigens are possible , meaning that hundreds of thousands of
11	EO_FICTION_010	attention , Bhalu suggested we travel to a village near Nandul . **He** wouldn ' t say why the diversion was necessary . But I
12	EO_FICTION_008	My mother lead the book to me whenever she felt sad ; **she** said it gave her fortitude . I picked up her letter :
13	EO_ESSAY_008	in contrast to portfolio flows and bank lending , tends to be **less** attached to economic downturns and financial spillover and so is a more
14	EO_POPSCI_003	craving . Other medical interventions mimic a drug ' s effects and **thereby** dampen craving long enough for an addict to kick the habit .
15	EO_SHARE_006	September 11. This horrific tragedy and the unprecedented events that occurred in **its** immediate wake - such as the three-day shutdown of the stock market
16	EO_INSTR_010	not place anything on top of power cords or cables . Arrange **them** so that no one may accidentally step on or trip over them
17	EO_SPEECH_012	coalition partners , we have deposed two of the cruelest regimes of **this** or any time . The al-Qaeda network has been deprived of its
18	EO_WEB_003	photographs and they receive NO royalties on the sale of any of **these** . Christo and Jeanne-Claude have no royalties either on the books and
19	EO_SPEECH_009	In fact , BAe has more defense sales than Boeing and **it** sells more to the US Defense Department than it sells to the
20	EO_FICTION_001	I hate that stuff - Dorcas too . So me and **her** were different that way . When some nastymouth hollered , ' Hey

Figure 2: Annotated corpus on CQP Web

Anaphora in German and English. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 138–145. European Language Resources Association (ELRA).

M. Eckert and M. Strube. 2000. Dialogue Acts, Synchronizing Units, and Anaphora Resolution. *Journal of Semantics*, 17:51–89.

S. Evert, 2005. *The CQP Query Language Tutorial*. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, April. CWB version 2.2.b90.

Ch. Fellbaum. 1998. *Wordnet. An electronic lexical database*. Mass: MIT Press, Cambridge.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

S. Hansen-Schirra, S. Neumann, and E. Steiner. 2013. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

A. Hardie. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

H. Kermes and S. Evert. 2002. YAC – A Recursive Chunker for Unrestricted German Text. In Manuel Gonzalez Rodriguez and Carmen PazSuarez Araujo, editors, *Proceedings of the Third International Conference on Language Resources and Evaluation.*, pages 1805–1812.

H. Kermes. 2003. *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis, Universitt Stuttgart.

D. Klein and C. D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. Kunz and E. Lapshinova-Koltunski. in press. Cohesive conjunctions in English and German: Systemic contrasts and textual differences. In Kristin Davidse, Caroline Gentens, Caroline Kimps, and Lieven Vandelanotte, editors, *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora*. Rodopi, Amsterdam.

K. Kunz and E. Steiner. 2012. Cohesive substitution in English and German: a contrastive and corpus-based perspective. In Karin Aijmer and Bengt Altenberg, editors, *Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson*, pages 201–231. John Benjamins, Amsterdam.

K. Kunz and E. Steiner. 2013. Towards a comparison of cohesive reference in English and German: System and text. In Maite Taboada, Susana Doval Surez, and Elsa Gonzlez lvarez, editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives*, pages 208–239. Equinox, London.

K. Kunz. 2009. *Variation in English and German Nominal Coreference*. Ph.D. thesis, Saarland University, Frankfurt/Main.

E. Lapshinova-Koltunski and K. Kunz. in press. Conjunctions across Languages, Registers and Modes: semi-automatic extraction and annotation. In A. Diaz Negrillo and F. J. Daz Prez, editors, *Specialisation and Variation in Language Corpora*. Peter Lang. Papers from the CILC2012. Jaen, Spain, March 2012.

E. Lapshinova-Koltunski, K. Kunz, and M. Amoia. 2012. Compiling a Multilingual Spoken Corpus. In Tommaso Raso Heliana Mello, Massimo Pettorino, editor, *Proceedings of the VIIth GSCP International Conference: Speech and corpora*, pages 79–84, Firenze. Firenze University Press.

H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system. In *CoNLL-2011 Shared Task*.

M.M. Louwerse and A.C. Graesser. 2005. Coherence in discourse. In P Strazny, editor, *Encyclopedia of linguistics*, pages 216–218. Fitzroy Dearborn, Chicago.

C. Müller and M. Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *LREC*. European Language Resources Association.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Morristown, NJ, USA. Association for Computational Linguistics.

R. Weischedel and A. Brunstein. 2005. BBN Pronoun Coreference and Entity Type Corpus.

Spatio-temporal grounding of claims made on the web, in PHEME

Leon Derczynski and Kalina Bontcheva

University of Sheffield
S1 4DP, UK
leon,kalina@dcs.shef.ac.uk

Abstract

Social media presents us with a digitally-accessible sample of all human discourse. This sample is full of claims and assertions. While the state of the art in NLP is adapting to the volume, velocity and variety of this sample and the information in it, the accuracy of claims made in social media remain largely unstudied. PHEME, a 36-month EU project started in January 2014, focuses on this fourth challenge: veracity. As a core part of establishing veracity, we need to identify the spatio-temporal context of assertions made on informal websites. This project note introduces the spatio-temporal challenges and planned semantic annotation activities that are part of the PHEME project.

1. Introduction

Social networks are rife with lies and deception, half-truths and facts. Irrespective of an assertion's truthfulness, the rapid spread of such information through social networks and other online media can have rapid and serious consequences. In such cases large amounts of user-generated content need to be analysed quickly, yet it is not currently possible to carry out such complex analyses in real time.

Social media poses three major computational challenges, dubbed the 3Vs of big data: volume, velocity, and variety (Laney, 2001). Content analysis methods have faced additional difficulties, arising from the short, noisy, and strongly contextualised nature of social media. In order to address the 3Vs of social media, new language technologies have emerged, e.g. using locality sensitive hashing to detect breaking news stories from media streams (volume), predicting stock market movements from microblog sentiment (velocity), and recommending blogs and news articles based on user content (variety).

PHEME[1] focuses on a fourth crucial challenge: veracity. It will model, identify, and verify **phemes** – internet memes annotated for truthfulness or deception – as they spread across media, languages, and social networks.

One of the many challenges in determining veracity is the automatic extraction of a claim's context. As well as understanding complex social context, it is critical to know when and where each claim was made, or to when and where it was intended to apply. This project note discusses the role of spatio-temporal information extraction and reasoning in solving this challenge.

2. Motivation

PHEME addresses the spatio-temporal validity of information and historical content to assess contradictions, through means of regional and longitudinal models of users, networks, trust, and influence.

The temporal delimitation of any assertion is of great importance, because the assertion is true only inside these bounds. Specifically, it is possible to extract two truths that seem to contradict (e.g. "The president of the USA is George W Bush" and "The president of the USA is Barack Obama") but are in fact both accurate when the appropriate temporal information is added. In other words, there is something like temporal validity of facts, which needs to be taken into account when detecting contradictions.

Similarly, assertions have spatial constraints, especially when they discuss underspecified entities. For example, we may say "The president is Obama" and "The president is Hollande"; these assertions seem to conflict, but are in fact both true simultaneously – just in distinct spatial regions.

It may not always be possible to ground assertions using single mentions of relations. Assertions may be spread over multiple documents, each mentioning different constraints. However, failing to determine the bounds of assertions – or assigning incorrect dates and places to claims – potentially leads to the rejection of correct information, reducing our overall ability to detect and ground/refute rumours in real-time. Spatio-temporal reasoning and inference offer solutions to these problems, and PHEME seeks to advance spatio-temporal relationship extraction to support measurement of veracity on the web.

3. Background

Unlike traditional news, a notable proportion of social media content posted online is explicitly geotagged (Sadilek et al., 2012), and studies suggest that it is possible to infer the geo-locations of about half of the remaining such content (Rout et al., 2013). Social media messages also have at least a creation time as temporal context. This implicit spatio-temporal (ST) metadata is not currently heavily exploited by modern NLP methods.

Given the constraint that a single entity can only be in any one place at a time, these forms of ST information give a means of determining the truthfulness of statements (Ji and Grishman, 2011; Derczynski and Gaizauskas, 2013).

Temporally, current systems are capable of detecting the publication date of documents (Chambers, 2012) and of grounding some of the time expressions contained therein (Strötgen and Gertz, 2010). Detecting events and assertions and temporally ordering these with regard to times is critical to ST grounding of facts and rumours; the state of the art in event detection is good (Kolya et al., 2012), but ordering events and times relative to each other

[1] The project is named after the Greek goddess Pheme, who was the personification of fame and renown; her favour being notability, her wrath being scandalous rumours.

or across documents remains an active area of novel research with some progress to be made. Fortunately, linking events to times – the most important type of temporal association for PHEME – is the task at which automated systems perform best (Derczynski, 2013).

Spatially, the challenge of grounding the locations in document content is critical to accurate bounding of claims. The state of the art is somewhat less mature than that of temporal context; while many tools can identify a range of named entities, recognition of new families of spatial entities (especially when general nouns are used in a spatial sense) is a subject of active research, e.g. Gaizauskas et al. (2012).

Spatio-temporal annotation in PHEME serves as one component in a complex system, linked together with longitudinal user behaviour modelling, information provenance, network structure, a-priori knowledge, and cross-media links.

4. Digital journalism

Journalists are currently using a plethora of social media applications in order to meet their diverse needs, e.g. Tweetdeck[2] for monitoring the social web; Storify[3] for news aggregation; crowdsourcing tools like Ushahidi's Swiftriver platform,[4] and online content filtering sites like Storyful.[5] The focus of all these tools is on getting the right content to journalists, but not on helping them with interpretation and verification of the authenticity and credibility of that content. Methods and tools vary according to the nature of the journalistic task, however. For example, observations of the Guardian newsroom (Procter et al., 2012) revealed that journalists prefer simple Twitter clients rather than more sophisticated tools such as Tweetdeck in activities such as live blogging. For reliability's sake, journalists prefer to rely on sources that their experience suggests they can trust. This solves the problem of reliability but limits their capacity to exploit social media to its full potential.

Spatio-temporal knowledge plays also an important role in this use case. A key challenge is to identify the regionality of events (e.g., neighbourhood, city, or country level) (Xu et al., 2012). Regionality is important because different events are relevant at varying scales, which impacts their newsworthiness and interestingness to digital journalists and users interested in local content.

5. Content Annotation

The project involves the creation of new language resources. These in turn helps create and evaluate general-purpose tools for projecting spatio-temporal annotations across languages, given parallel texts and re-using existing corpora (e.g. TimeBank, the multilingual TempEval, ACE2 temporal annotations, WikiwarsDE). The resources will be used to develop multilingual temporal annotation tools, based on their state-of-the-art techniques, developed for longer texts.

The project also addresses the problem of geo-locating events mentioned in documents. We intend to go beyond features based on words in the document, and use disambiguated URIs (e.g. against GeoNames[6]) and additional knowledge from the LOD resource (e.g. NUTS subdivisions, latitude/longitude, neighbouring locations).

Regarding annotation schemata, the de facto standards of ISO-TimeML (Pustejovsky et al., 2010) and ISO-Space (Pustejovsky et al., 2011) will be adopted and experimented with. Following Pustejovsky and Stubbs (2011), we intend to use temporal narrative containers for annotating events. In addition, recent adaptations of narrative containers to spatial annotation will be tried (Pustejovsky and Yocum, 2013). Narrative containers promise to lighten human annotator load while still capturing expressive representations of spatio-temporal information.

Standoff annotation may be required in some scenarios, as social media data typically has strict licensing constraints.

Existing standards provide a framework for annotating the factuality of assertions (e.g. Saurí and Pustejovsky (2009)), which can be applied over social media text in order to formalise the strength of assertions made there.

In the scope of rumour detection and analysis, Qazvinian et al. (2011) annotated messages for whether or not they related to a pre-determined rumour. PHEME involves two additional challenges: identifying the rumours in the first place, and then identifying the type of rumour from one of four classes: misinformation, disinformation, controversy and speculation. For the project, a "code frame" system is under development (Procter et al., 2013) for annotating topics and actor types. Code frames are specific to a research question that embodies information demand. Rumour messages are subdivided into categories, which may be related to claims, counter-claims or appeals for information; be with or without evidence; or simply rumour-relevant comments. Streams of related messages are categorised using code frames and annotated accordingly.

6. Project Contribution

PHEME aims to further the state of the art in spatio-temporal annotation and reasoning. In order to spatio-temporally ground asserions, PHEME will adapt existing annotation tools to social media data, through the creation of new training data in this genre. The project will also cover new target languages through lightly-curated annotation porting, taking advantage of the language-independent nature of grounded spatial and temporal data.

Another important benefit of storing and analysing "traditional" and social media content over space time is that these archives enable longitudinal analyses (Derczynski et al., 2013). For instance, longitudinal analyses on the online social graphs can reveal the evolution of social relationships and thus build models of trustworthiness and authority. It is also possible to start building user profiles over time, including previously spread rumours and, in general, what users talked about in the past. Focused on specific events, longitudinal analysis reveals discourse around events, arising from both social and traditional media. Similarly, in journalism and brand and reputation management applications, there is also demand for retrospective analyses of

[2] See http://www.tweetdeck.com/

[3] See http://www.storify.com/

[4] See http://www.ushahidi.com/products/swiftriver-platform

[5] See http://storyful.com/

[6] See http://www.geonames.org/

media content after a significant incident (e.g. to establish whether social media was used to entice more riots).

6.1. Dataset collection

The first phase is a human pilot annotation, of events, times and places in the target genres and languages. This includes annotation of web and social media text for events and times, in order to later temporally bound assertions. It also includes the annotation of locations (both formal and informal), and identification of document creation locations.

Corpora are then extended using cross-linguistic projection. PHEME will develop tools for projecting ST annotations across language (Spreyer and Frank, 2008; Costa and Branco, 2012). This allows the creation of new resources for English, German, Bulgarian, and possibly also the project's minor languages (French, Italian, Swahili).

Following the construction of a dataset, we will build spatial and temporal IE systems in multiple languages. These are aimed at ST grounding. As mentioned in Section 5., we intend to follow the narrative containers scheme. This centres on finding spaces and times within which groups of events are collected, before trying to resolve the specific, hard-to-annotate and potentially low-information individual relations. Finally, for grounding, while documents often come with a document creation date, and document creation location is harder to come by. To address this, the project investigates spatial grounding at both document level (creation location) and at assertion/event level.

Having found spatial and temporal entities in documents, it becomes possible to reason about bounds of assertions. We will apply and extend temporal reasoning and assertion bounding tools, which brings interesting challenges, particularly in the social media domain where one may be faced with many short documents describing different facets of a claim. In particular, cross-document spatio-temporal reasoning is a novel and unexplored research area. The output of these reasoning and bound-finding tools will be used as inputs to trustworthiness assessment systems.

6.2. Spatio-Temporal Information Extraction

Building upon existing resources is important to the advanced, complex tasks that PHEME addresses. Fully-featured ST information extraction pipelines can be built from state-of-the-art tools.

Regarding temporal annotation, we begin with annotation primitives: timexes, events and the relations between them. These are reasonably well-researched problems in newswire, but adapting to short messages which are pushed over networks by humans – i.e. social media messages – presents challenges in terms of the large linguistic variety, and interesting opportunities, from extra information and structure in personal profiles and network connections.

There are existing tools that may provide initial insights into the problem. For timexes, GATE and HeidelTime (Strötgen and Gertz, 2012) offer excellent entity extraction; TIMEN provides an open-source normalisation resource, and the state of the art leads to flexible parsing tools for handling previously-unseen timex formulations (Angeli et al., 2012; Bethard, 2013). Regarding event extraction, while older systems like EVITA are available, fast newer systems like TIPSem and the outcomes of the ARCOMEM project (Demidova et al., 2013) offer better performance. For linking times to events, systems like TIPSem, ClearTK-TimeML and NavyTime may be helpful.

Fewer tools are available for spatial annotation. In terms of locally-accessible location annotation systems, there is ANNIE, LODIE (Damljanovic and Bontcheva, 2012) and tools resulting from SemEval exercises. Developing spatial grounding and annotation systems involves more pioneering work here, beyond adapting existing tools.

Importantly, the project involves the creation of new tools for social media. PHEME couples systems like the above with document grounding and temporal relation annotation systems which operate on new languages and domains. This will involve the creation of new systems and annotations for event co-reference extraction, event-based summarisation, and ST grounding of individual messages. For example, the TimeML <TLINK> tag allows expression of intra-document co-reference and full-interval ordering between events, but need to be extended to handle both uncertain relations and also cross-document links. Similarly, the ability to create ISO-Space <PATH>s, <QSLINK>s between <LOCATION>s in different documents is required – as well as the ability to define common frames of references. We anticipate cross-document co-reference being instrumental in the grounding and subsequent veracity assessment of a significant proportion of claims and messages.

Social media networks present an unconventional kind of discourse, with different uses of reference and anaphora when compared to longer, standalone documents. The investigation of this structure will inform how annotations are used, and then leveraged for reasoning. Cross-document event co-reference is critical in order to group claims together; there is no work on this in social media, but challenges such as TDT generated extensive research on the general topic, e.g. (Bagga and Baldwin, 1999), and general concepts like chains provide a starting point.

Construction of timelines from timexes in messages and events mentioned across the network can then help define temporal bounds for events. Ji and Grishman (2011) excellent work on timelines proposes temporal bounding of assertions using times mentioned in collections of newswire documents, though this is all at day-level granularity. This granularity is suitable for retrospective analysis involving certain types of assertion (e.g. lifetimes), but not sufficient for realtime filtering of all kinds of events. In addition, extracted temporal bounds are likely to be uncertain, and require e.g. a constraint-satisfaction framework to pin down, as well as probabilistic veracity reasoning.

6.3. Evaluation

There are many ways in which PHEME's ST annotation output can be evaluated.

Primarily, we can evaluate against a gold standard (ours, or external ones, e.g. from TempEval). A secondary round of pilot annotations, over non-projected data, provides an opportunity for GS-style evaluation, as well as creating new high-quality language resources. Basic P/R/F1 measures work for spatio-temporal entity extraction – but one also needs to account for entity specificity. This may lead to adopting or creating a new, nuanced evaluation measure. It is difficult to evaluate spatio-temporal reasoning; prior shared tasks in these areas have demonstrated this.

In addition, we can perform extrinsic evaluation using unskilled humans. For assertion grounding, a common-sense check can be applied, asking whether a particular claim (in prose) is intended to apply to certain ST constraints. This could be formulated as dialogue or question answering. A high quality crowdsourcing approach is feasible for this extrinsic evaluation (Sabou et al., 2014).

7. Conclusion

PHEME involves creating the necessary computational apparatus to model, identify, and verify phemes (internet memes with added truthfulness or deception), as they spread across media, languages, and social networks. Doing this raises difficult, interesting and important issues in spatio-temporal annotation of text in a wide variety of situations. PHEME investigates these issues in the context of social media, examining digital journalism and healthcare. Furthering spatio-temporal information extraction research promises a better understanding of the ever-present context that the meaning language relies upon so heavily.

Acknowledgments

This work received funding from the European Unions Seventh Framework Programme (grant No. 611233 PHEME).

8. References

G. Angeli, C. D. Manning, and D. Jurafsky. 2012. Parsing time: Learning to interpret time expressions. In *Proc. NAACL*, pages 446–455. ACL.

A. Bagga and B. Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the Workshop on Coreference and its Applications*, pages 1–8. ACL.

S. Bethard. 2013. A synchronous context free grammar for time normalization. In *Proc. EMNLP*, pages 821–826.

N. Chambers. 2012. Labeling documents with timestamps: Learning from their time expressions. In *Proc. ACL*.

F. Costa and A. Branco. 2012. TimeBankPT: A TimeML Annotated Corpus of Portuguese. In *Proc. LREC*, pages 3727–3734.

D. Damljanovic and K. Bontcheva. 2012. Named Entity Disambiguation using Linked Data. In *Proceedings of the 9th Extended Semantic Web Conference*.

E. Demidova, D. Maynard, N. Tahmasebi, Y. Stavrakas, V. Plachouras, J. Hare, D. Dupplaw, and A. Funk. 2013. Extraction and Enrichment. Deliverable D3.3, ARCOMEM.

L. Derczynski and R. Gaizauskas. 2013. Information retrieval for temporal bounding. In *Proc. ICTIR*.

L. Derczynski, B. Yang, and C. Jensen. 2013. Towards Context-Aware Search and Analysis on Social Media Data. In *Proceedings of the 16th Conference on Extending Database Technology*. ACM.

L. Derczynski. 2013. *Determining the Types of Temporal Relations in Discourse*. Ph.D. thesis, University of Sheffield.

R. Gaizauskas, E. Barker, C. Chang, L. Derczynski, M. Phiri, and C. Peng. 2012. Applying ISO-Space to Healthcare Facility Design Evaluation Reports. In *Proc. ISA*, pages 31–38.

H. Ji and R. Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proc. of ACL'2011*, pages 1148–1158.

A. K. Kolya, D. Das, A. Ekbal, and S. Bandyaopadhyay. 2012. Roles of event actors and sentiment holders in identifying event-sentiment association. In *Computational Linguistics and Intelligent Text Processing*.

D. Laney. 2001. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6.

R. Procter, A. Voss, and P. Brooket. 2012. A study of using social media in journalism. Technical report, University of Warwick.

R. Procter, J. Crump, S. Karstedt, A. Voss, and M. Cantijoch. 2013. Reading the riots: What were the police doing on twitter? *Policing and society*, 23(4):413–436.

J. Pustejovsky and A. Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proc. LAW*, pages 152–160. ACL.

J. Pustejovsky and Z. Yocum. 2013. Capturing Motion in ISO-SpaceBank. In *Proc. ISA*, pages 25–33.

J. Pustejovsky, K. Lee, H. Bunt, and L. Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proc. LREC*.

J. Pustejovsky, J. L. Moszkowicz, and M. Verhagen. 2011. ISO-Space: The annotation of spatial information in language. In *Proc. ISA*, pages 1–9.

V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proc. EMNLP*, pages 1589–1599. ACL.

D. Rout, D. Preotiuc-Pietro, K. Bontcheva, and T. Cohn. 2013. Where's @wally? a classification approach to geolocating users based on their social ties. In *Proc. Hypertext*.

M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proc. LREC*.

A. Sadilek, H. Kautz, and V. Silenzio. 2012. Modeling spread of disease from social interactions. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 322–329. AAAI.

R. Saurí and J. Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *JLRE*, 43(3):227–268.

K. Spreyer and A. Frank. 2008. Projection-based acquisition of a temporal labeller. In *Proc. IJCNLP*.

J. Strötgen and M. Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proc. SemEval*, pages 321–324.

J. Strötgen and M. Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proc. LREC*, pages 3746–3753. ELRA.

J.-M. Xu, A. Bhargava, R. Nowak, and X. Zhu. 2012. Socioscope: Spatio-temporal signal recovery from social media. In *Machine Learning and Knowledge Discovery in Databases*, pages 644–659. Springer.

How to exploit paralinguistic features to identify acronyms in texts?

Mathieu Roche

UMR TETIS, Cirad, Irstea, AgroParisTech,
500, rue J.F. Breton, 34093 Montpellier Cedex 5, France
Mathieu.Roche@cirad.fr

LIRMM, CNRS, Univ. Montpellier 2,
161, rue Ada, 34000 Montpellier, France
Mathieu.Roche@lirmm.fr

Abstract

This paper addresses the issue of acronym dictionary building. The first step of the process identifies acronym/definition candidates, the second one selects candidates based on a letter alignment method. This approach has two advantages because it enables (1) to annotate documents, (2) to build specific dictionaries. More precisely, this paper discusses the use of a specific linguistic concept, *the gloss*, in order to identify candidates. The proposed method based on paralinguistic markers is independent of languages.

Keywords: text mining, acronym expansion

1. Introduction

Acronyms are numerous in specialized domain, e.g. biomedical and agronomy documents (Chang et al., 2002). An acronym is a set of characters corresponding to the first letters of a group of words, for instance, the acronym *FAO* is associated with the definition *Food and Agriculture Organization*. This paper summarizes a method to identify acronyms and expansions in documents. This automatic recognition enables to annotate these elements in texts.
This work deals with the use of paralinguistic features in order to identify acronym/definition couples.

After the description of related work in the following section, Section 3. describes our approach based on 2 steps: Extraction of acronym/expansion candidates (Section 3.1.), Filtering of candidates (Section 3.2.). Finally, before Discussion and Conclusion sections, experiments of our approach are detailed in Section 4.

2. Related work

Among the several existing methods for acronym extraction in the literature, some significant work need to be mentioned. The acronym detection involves recognizing a character chain as an acronym and not as an unknown or misspelled word. Most acronym detecting methods rely on using specific linguistic markers.

Yates' method (Yeates, 1999) involves the following steps: First, separating sentences by segments using specific markers (brackets, points) as frontiers. Then the acronym/expansion couples are tested. The acronym/definition candidates are accepted if the acronym characters correspond to the first letters of the potential definitions words. The last step uses specific heuristics to select the relevant candidates. These heuristics rely on the fact that acronyms length is smaller than their expansion length, that they appear in upper case, and that long expansions of acronyms tend to use stop-words such as determiners, prepositions, suffixes and so forth. Therefore, the pair ”FAO/Food and Agriculture Organization” is valid according to these heuristics.

Other studies (Chang et al., 2002; Larkey et al., 2000) use similar methods based on the presence of markers associated with linguistic and/or statistical heuristics. In this context (Okazaki and Ananiadou, 2006) propose statistical measurements from the terminology extraction area. Okazaki and Ananiadou apply the C-value measure (Frantzi et al., 2000; Nenadic et al., 2003) initially used to extract terminology. It favors a candidate term that doesn't appear often in a longer term. For instance, in a specialized corpus (i.e. Ophthalmology), the authors discovered that the term ”soft contact” was irrelevant, while the frequent and longer term ”soft contact lens” is relevant. An advantage of C-value measure is its independence from characters alignment (actually, a lot of acronyms/definitions are relevant while the letters are in a different order, e.g. ”AW / water activity”).

Other approaches based on supervised learning methods consist of selecting relevant expansions. In (Xu and Huang, 2007), the authors use SVM approaches (Support Vector Machines) with features based on acronym/expansion information (e.g. length, presence of special characters, context, etc). (Torii et al., 2007) present a comparative study of the main approaches (supervised learning methods, rules-based approaches) by combining domain-knowledge.

Larkey *et al.*'s method (Larkey et al., 2000) uses a search engine to enhance an initial corpus of Web pages useful for acronym detection. To do so, starting from a list of given acronyms, queries are built and submitted to the AltaVista[1] search engine. Query results are Web pages which URLs are explored, and possibly added to the corpus.

3. Acronym/expansion recognition

Our method of construction of acronym dictionaries is based on two steps detailed in the following subsections.

3.1. Step 1: Extraction of candidates

First, specific punctuation and character markers are taken into account in order to identify acronym/definition pairs (see Figure 1). In this paper, we investigate the extraction of candidates by exploiting the "glosses" of words and paralinguistic markers (i.e. brackets, punctuations, etc.) to detect acronym/definition candidates.

Glosses are spontaneous descriptions identifiable with specific markers (for example, *called*, *i.e.*, and so forth). These ones highlight lexical semantic relationships, e.g. equivalence, specification of the meaning, nomination, hyponomy, hyperonomy.

The abstract pattern of glosses is given by the structure X *marker* $Y_1, Y_2 ... Y_n$ where X and Y_i can be acronyms and/or definitions. The identification and selection of glosses are based on the use of patterns and Web-mining approaches (Mela et al., 2012).

In this paper, we extract candidates based on the gloss markers "(" and ")":

- **Local Pattern 1 [X=acronym, Y_1=definition]:** The first pattern detects Y_1 (definition), between "(" and ")" following the acronym (X). For example, the sentence *"relation empirique entre l'indice de végétation NDVI (Normalized Difference Vegetation Index), mesuré au maximum ..."* allows to extract X = *NDVI* and Y_1 = *Normalized Difference Vegetation Index*.
- **Local Pattern 2 [X=definition, Y_1=acronym]:** The second pattern detects Y_1 (acronym), between "(" and ")" following the definition (X). The beginning of the definition is recognized with the first word of the phrase in upper case. For example, the sentence *" ... la mesure Normalized Difference Vegetation Index (NDVI)"* allows to extract X = *Normalized Difference Vegetation Index* and Y_1 = *NDVI*.

Note that these patterns are independent of languages because the method is based on paralinguistic markers (i.e., brackets in this work). This is very important when languages are mixed, for instance in specialized domains. The example of Figure 1 shows a definition in English (expansion of "NDVI") in an abstract written in French.

In this situation, we are 4 different cases of results:

- **Case 1:** the relevant definition is returned (like previous examples),
- **Case 2:** the extracted phrase contains the relevant definition (i.e. partially relevant, but too large),
- **Case 3:** the extracted phrase is a part of the relevant definition (i.e. partially relevant, but too specific),
- **Case 4:** the extracted phrase is irrelevant.

Both proposed patterns will be evaluated in Section 4. of this paper.

3.2. Step 2: Filtering of candidates

The second step aims at removing irrelevant acronym/definition pairs and deleting irrelevant word(s) from candidate definitions. For this process, we propose to align the acronym letters with the potential definition words, by mapping each acronym letter with the first character of each definition word, respecting the order of words. If the first letter of the candidate definition word can not be aligned with the acronym corresponding character, the following characters (of the word) are taken into account. For instance, this method allows to find that "Extraction Itérative de la Terminologie" is a possible definition of the French acronym EXIT.

4. Evaluation

This paper focuses on the study of a corpus of 2000 paper abstracts provided by Cirad[2]: French research centre working with developing countries to tackle international agricultural and development issues. Table 1 shows that better results are given with the second local pattern. But a lot of cases are partially relevant (i.e. $\sim$ 40%), so we have to improve and enrich this pattern approach.

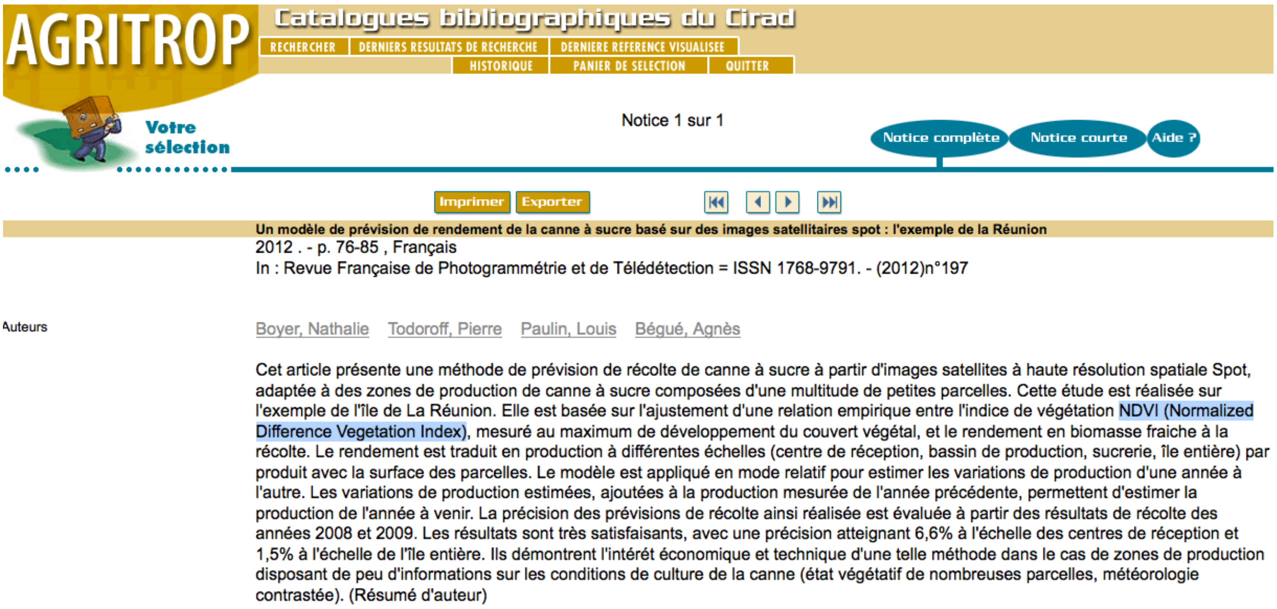
Patterns	Local pattern 1	Local pattern 2
Number of extracted definitions	78	64
Case 1 (relevant)	31 *(39.7%)*	28 *(43.7%)*
Case 2 (partially relevant)	3 *(3.8%)*	6 *(9.3%)*
Case 3 (partially relevant)	1 *(1.3%)*	18 *(28.1%)*
Case 4 (irrelevant)	43 *(55.1%)*	12 *(18.7%)*

Table 1: Evaluation of extracted definition with patterns.

The evaluation of the acronym/expansion extraction method is conducted on a corpus (general domain) having a reasonable size (7465 words). The experiments based on standard evaluation measures of data-mining domain highlight acceptable results (i.e. Precision: 66.7%, Recall:

[1] www.altavista.com/

[2] http://www.cirad.fr/en/home-page



Notice 1 sur 1

Un modèle de prévision de rendement de la canne à sucre basé sur des images satellitaires spot : l'exemple de la Réunion
2012 . - p. 76-85 , Français
In : Revue Française de Photogrammétrie et de Télédétection = ISSN 1768-9791. - (2012)n°197

Auteurs Boyer, Nathalie Todoroff, Pierre Paulin, Louis Bégué, Agnès

Cet article présente une méthode de prévision de récolte de canne à sucre à partir d'images satellites à haute résolution spatiale Spot, adaptée à des zones de production de canne à sucre composées d'une multitude de petites parcelles. Cette étude est réalisée sur l'exemple de l'île de La Réunion. Elle est basée sur l'ajustement d'une relation empirique entre l'indice de végétation NDVI (Normalized Difference Vegetation Index), mesuré au maximum de développement du couvert végétal, et le rendement en biomasse fraiche à la récolte. Le rendement est traduit en production à différentes échelles (centre de réception, bassin de production, sucrerie, île entière) par produit avec la surface des parcelles. Le modèle est appliqué en mode relatif pour estimer les variations de production d'une année à l'autre. Les variations de production estimées, ajoutées à la production mesurée de l'année précédente, permettent d'estimer la production de l'année à venir. La précision des prévisions de récolte ainsi réalisée est évaluée à partir des résultats de récolte des années 2008 et 2009. Les résultats sont très satisfaisants, avec une précision atteignant 6,6% à l'échelle des centres de réception et 1,5% à l'échelle de l'île entière. Ils démontrent l'intérêt économique et technique d'une telle méthode dans le cas de zones de production disposant de peu d'informations sur les conditions de culture de la canne (état végétatif de nombreuses parcelles, météorologie contrastée). (Résumé d'auteur)

Figure 1: Recognition of the couple *NDVI / Normalized Difference Vegetation Index* in AGRITROP database.

Examples of extracted with Local pattern 1	
NRPS	NonRibosomal Peptide Synthetase
VLE	Virtual Laboratory Environment
BMR	Bois Massif Reconstitué
ATPSM	Agricultural Trade Policy Simulation Model
ASA	Articulation du Semi-aride
CLF	Corynespora Leaf Fall
BASIC	Brésil, Afrique du Sud, Inde, Chine
Examples of extracted with Local pattern 2	
CIAT	Centro international de agricultura tropical
BSV	Banana streak virus
ER	Ehrlichia ruminantium
CSSV	Cacao swollen shoot virus
MAE	Mesures agrienvironnementales
ACMV	African cassava mosaic virus
TYLCV	Tomato yellow leaf curl virus

Table 2: Examples of acronyms/definitions.

80%, F-measure: 72.7%) (Roche and Prince, 2008). We plan to apply the second step of the process (see Section 3.2.) with the pattern approach described in Section 3.1. on the Cirad corpus.

Note that our previous work (Roche and Prince, 2008) uses more global patterns ; then a lot of noise is returned. The pattern approach described in this paper is more specific with better results in term of precision ($\sim$ 40% in this current work vs. 15% in our previous work).

5. Discussion: Towards a Web-mining approach

In this section, we propose to integrate Web-mining measures in order to automatically validate results returned by our approach (Turney, 2001; Mela et al., 2012). For instance, we can query a search engine with the acronym ”BSV” and its possible definition to check on the Web if this association exists. This query should be a disjunction (i.e. OR operator) of the acronym and its possible definition returned with our process (i.e. Banana streak virus). This one returns a larger amount of documents. The conjunction of the acronym and the expansion (i.e. AND operator) enables to return a lower number of documents. But the returned documents are more relevant (i.e. the precision is improved).

In our case, we choose to consider the ”hits” given by Google[3] on the examples of Table 2 (i.e. number of pages returned by the search engine based on conjunction). For instance, we have tested the query *”BSV” AND ”Banana streak virus”* that returns 7580 pages[4]. All the results (i.e. hits) are given in Table 3. This table shows that hits have generally very high values, this allows us to automatically validate acronym/definition couples. Note that hits of irrelevant couples return lower values (for instance, with the couples ”ETM”/”environ 5.000 m3.ha-1”, ”SIPSA”/”indicateurs, documents, cartes”, and so on).

Moreover, we can integrate this kind of information in classical similarity measures, e.g. Dice measure (Smadja et al., 1996). Dice measure can be used to compute a sort of relationship between an acronym (i.e. $acro$) and a definition (i.e. def). In our context, Dice measure (formula (1)) is based on the number of Web pages given by search engines (i.e. hits).

$$WebDice(acro, def) \quad = \quad \frac{2 \times \text{hits}(acro, def)}{\text{hits}(acro) + \text{hits}(def)} \quad (1)$$

[3] http://www.google.fr/

[4] Queries performed on the 20th of March 2014.

Acronym	Possible definition	Hits (Google)
NRPS	NonRibosomal Peptide Synthetase	230000
VLE	Virtual Laboratory Environment	36900
BMR	Bois Massif Reconstitué	9270
ATPSM	Agricultural Trade Policy Simulation Model	27700
ASA	Articulation du Semi-aride	663
CLF	Corynespora Leaf Fall	22800
BASIC	Brésil, Afrique du Sud, Inde, Chine	21100
CIAT	Centro international de agricultura tropical	75000
BSV	Banana streak virus	7580
ER	Ehrlichia ruminantium	121000
CSSV	Cacao swollen shoot virus	2040
MAE	Mesures agrienvironnementales	951
ACMV	African cassava mosaic virus	90200
TYLCV	Tomato yellow leaf curl virus	354000

Table 3: Examples of acronym/definition and hits scores.

This measure returns the following result with the previous example:

$$Web_{Dice}(BSV, Banana\ streak\ virus)$$

$$= \frac{2\times\text{hits}(''BSV''\ AND\ ''Banana\ streak\ virus'')}{\text{hits}(''BSV'')+\text{hits}(''Banana\ streak\ virus'')}$$

$$= \frac{2\times 7580}{2840000+15400}$$

$$= 0.0053$$

Web_{Dice} can be applied in order to rank couples (see Table 4). This enables to detect relevant acronym/definition pairs (i.e. couples with high Web_{Dice} values).

Acronym	Possible definition	Web_{Dice}
ATPSM	Agricultural Trade Policy Simulation Model	1.3014
TYLCV	Tomato yellow leaf curl virus	0.7167
NRPS	NonRibosomal Peptide Synthetase	0.4423
CIAT	Centro international de agricultura tropical	0.1408
ACMV	African cassava mosaic virus	0.0970
CSSV	Cacao swollen shoot virus	0.0245
VLE	Virtual Laboratory Environment	0.0222
CLF	Corynespora Leaf Fall	0.0208
BSV	Banana streak virus	0.0053
BMR	Bois Massif Reconstitu	0.0046
ER	Ehrlichia ruminantium	0.0004
BASIC	Brsil, Afrique du Sud, Inde, Chine	0.0001
ASA	Articulation du Semi-aride	0
MAE	Mesures agrienvironnementales	0

Table 4: Acronym/definition couples ranked with Web_{Dice}.

6. Conclusion

The process described in this paper is based on the use of specific linguistic markers to detect acronyms. In future work we plan to integrate statistical information and Web-mining approaches in order to improve our methods based on linguistic rules.

Our text-mining system allows us to enrich specialized thesaurus (e.g. MeSH[5], Agrovoc[6]). These thesaurus are useful to automatically annotate texts.

Moreover we plan to investigate a contrastive analysis of English/French corpora in order to give a new point of view of the phenomenon of spontaneous descriptions. A first study on aligned English/French texts reveals frequent regularities of glosses in a multilingual context. The alignment enables to improve the multilingual lexical acquisition of new words and their translations.

7. Acknowledgements

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University Montpellier 2 and CNRS.

8. References

Chang, J., Schtze, H., and Altman, R. (2002). Creating an online dictionary of abbreviations from medline. *Journal of the American Medical Informatics Association*, 9:612–620.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Larkey, L. S., Ogilvie, P., Price, M. A., and Tamilio, B. (2000). Acrophile: An automated acronym extractor and server. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pages 205–214.

Mela, A., Roche, M., and el Amine Bekhtaoui, M. (2012). Lexical knowledge acquisition using spontaneous descriptions in texts. In *Proceedings of Natural Language Processing and Information Systems Conference (NLDB)*, pages 366–371.

Nenadic, G., Spasic, I., and Ananiadou, S. (2003). Terminology-Driven Mining of Biomedical Literature. *Bioinformatics*, 19(8):938–943.

Okazaki, N. and Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–3095.

Roche, M. and Prince, V. (2008). Managing the acronym/expansion identification process for text-mining applications. *Int. J. Software and Informatics*, 2(2):163–179.

Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Torii, M., Hu, Z., Song, M., Wu, C., and Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*.

Turney, P. (2001). Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Proceedings of ECML, Lecture Notes in Computer Science*, pages 491–502.

Xu, J. and Huang, Y. (2007). Using SVM to extract acronyms from text. *Soft Comput.*, 11(4):369–373.

Yeates, S. (1999). Automatic extraction of acronyms from text. In *New Zealand Computer Science Research Students' Conference*, pages 117–124.

[5] http://www.nlm.nih.gov/mesh/

[6] http://aims.fao.org/standards/agrovoc/about

Lessons Learned from Manual Evaluation of Named Entity Recognition Results by Domain Experts

Sungho Shin[1], Hanmin Jung[1], Inga Hannemann[2], Mun Yong Yi[3]
[1]Korea Institute of Science and Technology Information, Daejeon, South Korea
[2]University of Hildesheim, Hildesheim, Germany
[3]Korea Advanced Institution of Science and Technology, Daejeon, South Korea
{maximus74, jhm}@kisti.re.kr, inga@outlook.de, munyi@kaist.ac.kr

Abstract

Recently, NER (Named Entity Recognition) has been adopted in many practical areas. People with smartphones may prefer services that manage their schedules automatically through scheduling applications that contain NER engines for extracting events from messages and emails. Diversifying the application of NER technology to various fields requires the accuracy of the technology. However, there is still a significant difference between NER results in laboratories and in real fields. For example, the F-score of our NER system is 0.75 in the laboratory and 0.22 in practice. In order to overcome this issue, NER evaluation should be performed manually such that developers and researchers can define the problems that can occur in practical environments with their current NER engines; this facilitates improvements in future versions. This paper addresses the extraction results of NER engines. We approach the problem by hiring domain experts to evaluate the extraction results. Certain problems that are not expected to be extracted by machines are presented; moreover, feedback from the problems is provided in order to improve the NER engine.

Keywords: NER, NER Evaluation, Domain Expert, F-measure, Manual evaluation

1. Introduction

With the rapid changes in today's world, it is important to stay current and up-to-date. Large amounts of information that can help with this task may be found on the Internet. However, people often lack sufficient time to read and understand such information. For convenience, most individuals prefer using machines to accomplish this task. The system discussed in this paper helps with such a task by extracting named entities from large texts. The extraction results are fairly acceptable for research purposes. For example, we obtained an F-score of 0.75 in laboratory tests. However, the goal of our system is to successfully apply the named entities to an intelligent service that we have developed. This means that the extraction results need to be acceptable for practical environments as well. In order to estimate such results in terms of the application, the results must be evaluated manually. Through manual evaluation, we can determine the real extraction results of the system. The F-score we obtained in practical environments is merely 0.22. As can be seen, there is a significant difference between the two F-scores. This means there are certain mistakes in the system that can be discovered only through manual evaluation. When the problems in the system have been identified, strategies to resolve such problems can be developed. This paper discusses the topic of domain experts evaluating extracted named entities. We demonstrate a few problems that are not expected to be extracted through machines and discuss feedback from the problems. Such feedback can contribute to improve the engine.

2. Related Works

A significant number of papers have been published on various approaches to named entity extraction. The evaluation of NER (Name Entity Recognition) systems is an important step for the improvement of such systems. Many approaches have been proposed to rank systems based on their annotating capability. MUC (Message Understanding Conference) events cover the correct type and text. The final MUC score is presented as the micro-averaged F-measure. It is a sort of harmonic mean of precision and recall calculated over all entity slots (Nadeau and Sekine, 2007). ACE (Automatic Content Extraction) evaluation is more complex than the F-measure. ACE considers methods for addressing various evaluation factors (such as partial match, and wrong type). The final score named EDR (Entity Detection and Recognition Value) is 100% minus the error rate (penalties).

Another approach that evaluates NER is Rizzo and Troncy's NERD (Named Entity Recognition and Disambiguation). NERD is an application that human evaluators can use to evaluate certain named entity extractors on the web (Rizzo and Troncy, 2011). A framework is also suggested to evaluate NER systems that do not participate in large evaluation conferences for different reasons, but still meet certain demands to qualify as NER systems (Marrero et al., 2009). Marrero et al. compare several systems based on functions, results, and other factors.

Another evaluation for NER results was performed by Santos et al. in 2006. Santos et al. propose HAREM (*HAREM-Avaliação de sistemas de Reconhecimiento de Entidades Mencionadas* — HAREM-Evaluation of NER Systems), which is a contest for Portuguese NER. HAREM consists of three tasks: identification, semantic classification, and morphological classification (Santos et al., 2006).

Typically, precision, recall, and F-measure are used as evaluation measures, and they obtain similar values for most systems with only a few exceptions. Such measures also compare the number of entities a system can recognize. In this paper, we adopt precision, recall, and F-measure for the evaluation of our NER system because these measures have been used widely.

3. NER System and Its Data

Our NER system executes in a distributed and parallel environment and uses dictionaries, rules, and machine learning for data extraction. This combination causes our system to execute much faster than the existing systems (Shin et al., 2014). Our system takes advantage of the machine learning method, especially the structured SVM (Support Vector Machine) that uses a variant of the Pegasos algorithm to recognize the required named entities. The Pegasos algorithm is faster and more accurate than the standard SVM training algorithms for structured SVM (Lee and Jang, 2009). There are 7 subtypes of named entities. *Institution*, *University*, and *Corporation* are different types of organizations that have been established to obtain specific roles in a society to achieve certain goals, to educate, or to produce and sell products or technologies; *Nation* describes the locations where the organizations operate; *Technology* in this context describes methods of tools, materials, and machine development or production processes and required necessary products; *Person* is the subtype for the persons who work for the organizations and conduct research, or who are otherwise related to the technologies or products; *Product* indicates the article, such as a model or series, that is produced in corporations using technologies (Shin et al., 2013). The resources that should be addressed for extraction are web articles, papers, and patents. The system extracts named entities from the titles and abstracts of papers and patents. For web articles, only the body is used for extraction. The number of documents is 4 millions for papers, 7 millions for patents, and 5 millions for web articles respectively. 16 million documents are totally processed in extraction.

Sub-Type	Precision			Recall		
	Collect	Total	Score	Collect	Total	Score
Person	2051	2708	0.76	2051	2421	0.85
Nation	519	669	0.78	519	636	0.82
University	3656	4011	0.91	3656	4042	0.90
Corporation	6623	8022	0.83	6623	8530	0.78
Institution	607	819	0.74	607	818	0.74
Technology	5581	7431	0.75	5581	9678	0.58
Product	3069	4247	0.72	3069	5227	0.59
Total	22106	27907	0.79	22106	31352	0.71

Table 1: Automated Evaluation Result

To build the structural SVM model, we made the tagged corpus which has 25,297 sentences with 687,632 named entities and their subtype. Training and evaluation are performed at a time. We gave –c 1000 as a parameter for the cost of the structured SVM and set 200 iterations with the Pegasos algorithm in training. We used the ten-fold validation for cross-validation. The cross-fold validation repeats the entire process multiple times with different random samples and decides on a fixed number of folds. Each fold in turn is used for testing, and the remainder for training. Ten error estimates are averaged. The precision, recall, and F-score are 0.79, 0.70, and 0.75, respectively (Table 1).

F1 = 2PR / (P + R) = 2 * 0.79 * 0.71 / (0.79 + 0.71) = 0.75

4. Manual Evaluation

A total of 10,000 sentences were randomly selected from entire sentences for evaluation. The evaluation was conducted by two domain experts. Each expert was provided with these sentences with instances of named entities that the system had extracted. Based on their availability, one expert evaluated 6,000 sentences for seven weeks for a total of 16 hours per week. The other expert reviewed 4,000 sentences for seven days for a total of approximately 57 hours. The actual amount per hour varied strongly depending on the complexity of the sentences. A total of 690 sentences were excluded because, for various reasons, it could not be determined whether the sentences were correct. We used 9,310 sentences for analysis. The number of sentences was 989 from titles, 2,614 from abstracts, and 5,707 from text bodies.

4.1 Process of Evaluation

The data that was used for evaluation was given in EXCEL files. The structure of the data that was evaluated contains three units: the sentence where the named entity was found, the named entity, and the subtype of the named entity. If more than one named entity was found in a sentence, it was listed again for each new entity. Each sentence was manually evaluated. Subsequently, it was determined whether a named entity was extracted correctly or incorrectly. A named entity was deemed to be correct if both the borders of the named entity and the subtype were correct. Finally, all the named entities that should have been extracted but were not extracted, which are false negative, were counted for recall. This process is illustrated in Figure 1.

4.2 Evaluation Measure

Micro-average precision and recall were used as the evaluation measure. In such a measure, the true positives are all the entities that were extracted correctly and the false positives are all the entities that were extracted incorrectly. False negatives are all the entities that should have been extracted but were not, thus constituting the amount of all recall errors. Additionally, it is important to add that a strict measure was used; extremely small mistakes lead to an extraction being classified as an error, even when such mistakes would not necessarily compromise the effectiveness of the system.

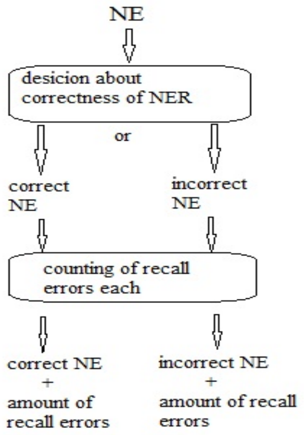

NE
desicion about correctness of NER
or
correct NE
incorrect NE
counting of recall errors each
correct NE + amount of recall errors
incorrect NE + amount of recall errors

Figure 1: Process of Manual Evaluation

4.3 Considerations

Although manual evaluation is more thorough and accurate, certain problems became apparent during the process. Evaluators found it difficult to read the data, which was parsed to enhance its accessibility for the system. Moreover, information that was vital for the understanding of a sentence and for differentiating between possible options was omitted occasionally. In some cases, such an omission poses a problem; certain sentences appear distorted when devoid of context, because the context might be required to understand the meaning of specific words in the extracted sentence, such as abbreviations. This limitation is somewhat unavoidable because of random selection. The data for manual evaluation is selected not from documents, but from parsed sentences for equality in number by year and resources. An issue that is linked to content more than format is the domain of the sentences. Such sentences are extracted from texts in the technology and natural science domains; therefore, these sentences can sometimes be extremely complicated to understand. In particular, sentences related to natural science often cannot be understood without a significant knowledge of the specific topic. Examples of such sentences are: 'synthesis of Azaheterocycles from Aryl Ketone - Acetyl Oximes and Internal Alkynes by Cu-; Rh Bimetallic Relay Catalysts'. Nevertheless, in some cases it is possible to determine without context that an entity has been incorrectly extracted because common sense dictates that a certain subtype cannot be correct. In cases where the lack of knowledge about a subject makes it unclear whether an entity has been correctly extracted, it is advisable to search for additional information on the Internet. We eliminated 690 such incomprehensible sentences from evaluation.

5. Manual Evaluation Results

5.1 Overview

The outcome of the evaluation shows that 5,695 named entities from a total of 9,310 were extracted incorrectly, and that 3,615 were extracted correctly (Figure 2).

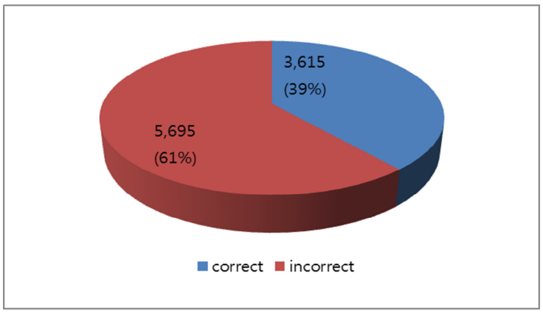


Figure 2: Accuracy of Named Entity Extraction

Figure 3 and Table 2 show the distribution of the number of recall errors in terms of correctness and incorrectness. The occurrence of recall errors in sentences with correctly extracted named entities peaks at three recall errors. The amount of correctly extracted named entities in sentences with no recall errors is considerably small because most sentences contain more than one relevant named entity.

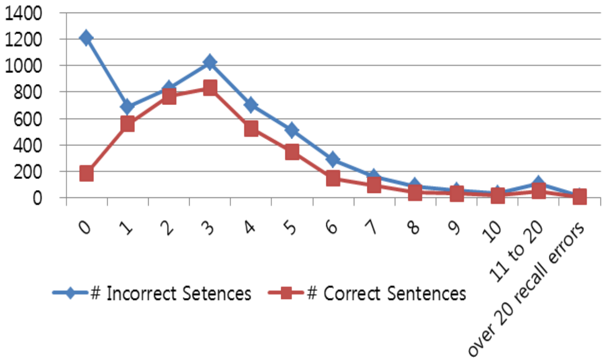


Figure 3: Distribution of the Number of Recall Errors

Number of Recall Errors	Number of Sentences among Incorrect Extractions	Number of Sentences among Correct Extractions
0	1,205	186
1	687	560
2	826	769
3	1,022	832
4	704	524
5	509	351
6	285	148
7	159	97
8	90	41
9	57	32
10	34	17
11 - 20	105	51
> 20	12	9
Total	5,695	3,615

Table 2: Number of Correctness and Incorrectness by Number of Recall Errors

In certain sentences, the number of recall errors is exceptionally high: 11 to 20 recall errors, or over 20 recall errors. These cases are sentences that list named entities, for example: 'The bus state controller includes control registers such as wait controllers, and controls the interface of various semiconductor memory (ROM), burst ROM, SRAM, PSRAM, DRAM, and synchronous RAM, and PC cards (memory and I/O cards in parallel)' or 'System builder and reseller offering Tesla-based Personal Supercomputers include Amax, Armari, Asustek, Azken Muga, Boxx, CAD2, CADNetwork, Carri, Colfax, Comptronic, Concordia, Connoisseur, Dell, Dospara, E-Quattro, JRTI, Lenovo, Littlebit, Meijin, Microway, Sprinx, Sysgen, Transtec, Tycrid, Unitcom, Ustar, Viglen, and Western Scientific'. Considering the F-measure, the precision value is 0.38 and the recall value is 0.15. F-score is calculated as the equation below shows.

$$F1 = 2PR / (P + R) = 2 * 0.38 * 0.15 / (0.38 + 0.15) = 0.22$$

5.2 Error Analysis

When performing the evaluation, different types of errors were observed. The error categories listed in this section either contain information about the subtype or the extracted entity, or a combination thereof. It can be expected that more than the categories listed here can be found.

(1) The entity does not belong to any of the subtypes and the subtype does not appear in the sentence.

In this category, both the named entity and the subtype are wrong and there is no visible connection between the sentence and the subtype, and between the extracted entity and the named entities.

<Example>
Sentence: the following virus which be list in order of the overall abundance within the tested sample be detect: Tobacco streak virus (TSV), Tomato spotted wilt virus (TSWV), Tobacco etch virus (TEV), Tobacco ring spot virus (TRSV), Potato virus Y (PVY), Cucumber mosaic virus (CMV) and Tobacco mosaic virus (TMV).
NE extracted by the system: Tomato
Subtype: $Person

(2) Wrong part of sentence was extracted.

In this case, the extracted entity is not a named entity or part of a named entity, but the subtype is relevant for another named entity in the sentence that has not been extracted.

<Example>
Sentence: SURFACE COATED CUTTING TOOL MADE of CEMENT HAVING PROPERTY-MODIFIED ALPHA TYPE ai203 layer of HARD COATING LAYER.
NE extracted by the system: HAVING
Subtype: $Technology

(3) The entity has the wrong subtype.

This can be the only error, or it can be combined with the errors explained in (4) and (6). If it is the only error, the named entity has been correctly found, but it has been matched with the wrong subtype.

<Example>
Sentence: spin around the first offering from IBM since its personal computing division be acquire by China's Lenovo be the ThinkPad X41 Tablet.
NE extracted by the system: Lenovo
Subtype: $Nation

(4) Only parts of the entity have been extracted.

There are parts that belong to the named entity, but the parts have not been extracted.

<Example>
Sentence: subscribe to anti-virus software, such as Norton AntiVirus, McAfee VirusScan or ZoneAlarm Security Suite.
NE extracted by the system: ZoneAlarm Security
Subtype: $Product

(5) Wrong subtype and only parts of the entity have been extracted.

This error is a combination of the errors explained in (3) and (4).

<Example>
Sentence: SYSTEM and METHOD for ESTABLISHING PEER TO PEER connection BETWEEN PCS and SMART PHONES USING network with obstacle.
NE extracted by the system: PEER
Subtype: $Institution

(6) More than what belongs to the entity has been extracted.

Parts that do not belong to the entity have been extracted, most often articles or conjunctions.

<Example>
Sentence: Symantec Corp. formed the same year Skrenta unleashed 'Elk Cloner', but it dabbled in non-security software before releasing an anti-virus product for Apple's Macintosh in 1989.
NE extracted by the system: Symantec Corp. formed
Subtype: $Corporation

	Number of occurrences	Percentage
(1) Wrong entity and subtype	85	42.5%
(2) Wrong part of sentence	11	5.5%
(3) Wrong subtype	36	18.0%
(4) Entity only partially extracted	25	12.5%
(5) Wrong subtype and entity only partially extracted	34	17.0%
(6) More than is part of the entity was extracted	9	4.5%

Table 2: Rate of Each Type of Errors

To compare the importance of each type of error, the rate of each type of error is calculated from a subset of 200 sentences that contain incorrect extractions (Table 2).

6. Lessons Learned

When evaluating the extractions, some observations can be made. There are certain specific errors that occur in the same patterns throughout the extracted results. Some of these patterns are domain-specific and some are general phenomena. Furthermore, there are some errors that the NER system generates independently.

First, the system has narrowly missed the correct entity.

The system extracts a word or phrase that can be, but is not necessarily, a named entity itself and assigns a wrong subtype. This is different from regular subtype errors, because the correctly named entity for the subtype is immediately adjacent to what was extracted. However, subtype errors in general are a frequent problem. The number of title-type sentences is 989. The precision errors are 906 out of 989. The rate is almost 92%. This problem originates from a mistake in the structural analysis of sentences, especially in the titles of papers and patents. The structure of article titles is different from normal sentences. Titles can be regarded as phrases that do not follow the form of sentences: titles typically do not have verbs. To resolve this problem, the system needs to be equipped with a structural analysis module that can address this type of phrases and accept them as sentences.

Second, errors are related to the nature of the natural science domain.

In the natural science domain, many procedures, scales, instruments, etc. are given the name of the person who invented or discovered them. This leads to the fact that there are many names in the sentences from the natural science domain that do not refer to a specific person. In most cases, the sentences referring to names belonging to specific people include titles such as Mr, Mrs, or Dr before the name, or the names are followed by the word *say* or a synonym; names that do not refer to specific people usually are surrounded by adjectives and nouns, and sometimes have an article in front of the name. In reality, names are extremely important in the natural science domain because they emerge in texts with the inventions such that the names provide information on who invented the technology. However, such names are not full names and are not identified with other similar names. The extraction can be compared to the head and tail of a coin. Therefore, researchers of NER systems need to decide, before extracting, whether a person's name linked with procedures, scales, and instruments from the natural science domain should be extracted.

Third, sentences containing many sequentially named entities produce many recall errors.

These types of sentences are uncommon, though they produce a significant effect on recall errors, as mentioned in Section 5. There were more than 20 named entities in a sentence that should have been extracted. For laboratory evaluation, this means merely one recall error. However, many useful named entities can be missed in practice. To solve this problem, we need to approach with a micro-solution. A NER system that uses machine learning methods is incapable of extracting these types of named entities unless training data do not include sentences or paragraphs that have these types of named entities. Therefore, we should consider various types of sentences while developing training data. The type of sentence is different among sources such as social networking services, journal articles, web articles, patents, emails, and essays. The sentences used for training should be selected based on the type of target sentences that are to be extracted. If the target sentences are from academic papers, the training sentences must include similar types of academic papers.

Fourth, certain named entities were only partially extracted.

The cases of (4) and (5) in Section 5.2 match this problem. The error rate is almost 20%, which is significant. In fact, this is not a precision error, but a recall error because the extracted named entity is correct and there is another named entity that should have been extracted. These two named entities overlap and share common tokens in the sentence. The example shown in Section 5.2 explains that the system is correct because ZoneAlarm Security is surely a named entity that we want to extract. The only mistake the system incurred is that it did not extract the other named entity in the sentence, ZoneAlarm Security Suite. In order to solve this problem, we need to tag both named entities in the training data.

7. Conclusion

The aim of the evaluation discussed in this paper is to determine the ability of the proposed system to manage the extraction of named entities, and to determine possible improvements to the system. Some concerns were found in relation to domain, in addition to other general concerns. First, the proposed system missed some correct entities that were immediately adjacent to the incorrect extraction; this problem can be solved by developing more capable structural analyses that can manage phrases. Second, a specific guideline for NER needs to be established before extraction because named entities can be included in jargons in a natural science domain. Third, the training sentences should be selected based on the type of target sentences in order to extract many sequentially named entities within a sentence. Fourth, both named entities should be tagged in the training data to ensure that the problem of partial extraction of some named entities is addressed.

8. References

Nadeau, D. and Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Linguisticae Investigations*. 30. pp. 3--26.

Rizzo, G. and Troncy, R. (2011). NERD: Evaluating Named Entity Recognition Tools in the Web of Dat. *In Proceedings of the Workshop on Web Scale Knowledge Extraction*.

Morrero, M., Sánchez-Cuadrado, S., Lara J. M., and Andreadakis, G. (2009). Evaluation of Named Entity Extraction System. *Research in Computing Science*, 41. pp. 47--58.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). HAREM: An Advanced NER Evaluation Contest for Portuguese. In Proceedings of Language Resources and Evaluation Conference.

Shin, S., Um, J., Choi, S. P., Jung, H., Yi, M. Y., and Lee, S. (2013). Platform to build the Knowledge Base by Combining Sensor Data and Context Data. *International Journal of Distributed Sensor Networks* (Online published).

Lee, C. and Jang, M. (2009). Large-Margine Training of Dependency Parsers Using Pegasos Algorithm. *ETRI Journal*. 32(3). pp. 486--489.

Shin, S., Jeong, C. H., Seo, D., Choi, S. P., Jung, H. (2013). Improvement of the Performance in Rule-Based Knowledge Extraction by Modifying Rules. *In Proceedings of the 2nd International Workshop on Semantic Web-based Computer Intelligence with Big-data*.

Annotating Discourse Zones in Medical Interviews

Milan Tofiloski, Evan Zhang, Fred Popowich
Simon Fraser University
8888 University Drive
Burnaby, BC, Canada
V5A 1S6
mta45@sfu.ca, evanz@sfu.ca, popowich@sfu.ca

Abstract

We improve a visual analytics workflow for analyzing medical interviews by introducing a discourse annotation scheme for creating an effective multi-document visualization that also facilitates inter-document comparisons. We introduce the concept of *discourse zones* for bringing together the many disparate terms and concepts used in various research areas. The zones are generalized for usage toward any institutional dyad setting (attorney-witness, teacher-student, physician-patient, etc.), including emergency hotlines and switchboards. Our task involves visually identifying the medical problems, their solutions, and contexts in medical encounters (i.e., dialogue-based conversations and interviews). The corpora consists of medical interviews between clinicians and caregivers of children with Fetal Alcohol Spectrum Disorder (FASD).

Keywords: medical terminology, discourse, interviews

1. Introduction

Medical practitioners and researchers examine document collections to understand patient behaviors, the situations surrounding the behavior, and strategies associated with patient care. However, the creation of interview scripts and the manual processing and analysis of interview content to identify situations, behaviors and strategies can be time consuming. Using a collection of semi-structured interviews between researchers and caregivers concerning the care and challenging behaviors of children with Fetal Alcohol Spectrum Disorder (FASD), we provide an analysis that detects patterns across the study population which also can be used for identifying new relationships. Our system assists analysis by automatically labelling the situations, behaviors, and strategies experienced by caregivers which is then used to provide a high-level, structural visualization of the document collection.

Analyzing the text of a conversational dialogue can be cognitively demanding due to the difficulties in viewing the entire conversation all at once (Tat and Carpendale, 2002). One of the objectives involves providing a view of an entire conversation in order to reduce cognitive load. A more formidable objective enables the viewing of multiple conversations simultaneously in order to discover patterns or hidden associations, specifically for identifying effective and ineffective caregiver strategies.

The goal is to segment the clinical conversations for improving an analyst's workflow in identifying patient behaviors as well as possible strategies associated with patient care. Comparisons between clinical interviews are also important in identifying patterns.

2. Annotation of Medical Interviews

The *Beginning-Middle-End* sequence is a familiar narrative structure and can be found in films, sports, literature, etc. Clinical interviews with caregivers can similarly be segmented according to whether the various interview sections are concerned with discussing Situations, Behaviors, or Strategies used to cope and manage.

For an analyst looking into clinical interviews to understand the handling of various scenarios while also being able to learn from said situations, the interview can be structurally characterized into sections such as *Problem*, *Cause*, or *Solution*, with a fourth label *None* that captures all other content occurring in a medical encounter not of interest to the analyst. The three zones we choose as the foundation for annotating the medical interviews are:

- the **Behavior** of the child
- the **Situation** or conditions which led to or caused said behavior
- and the **Strategy** employed by the caregiver to help the child (i.e., resolving the issue), as well as methods used in coping and managing the situation

This structure and its identification forms the basis of the visualization system. A sequence could begin with a Situation that then leads to a Behavior, which then results in a Strategy being employed by the caregiver to remedy said Behavior (assuming difficult behavior was being expressed). Thus, we annotate a medical interview into structural sections for creating a model of the sequences that occur, as well as using such annotations to improve a visual analytics workflow where many interviews are manually pored over.

The general template of the interviews between the clinician (DR) and caregiver/patient (P) were as follows:

> DR: Can you describe any scenarios? (...)
>
> P: Yes, one time at school...
>
> DR: How did you handle that situation? (...)
>
> P: Well, what we tried was...
>
> DR: And were there any other scenarios? (...)

After the discourse zones are applied to the interview, one of three colors is assigned to each zone, which provide a high-level visual overview consisting of possibly multiple interviews to be compared and viewed simultaneously. The sections annotated with the *None* label have no color applied.

3. Related Work

Much study on discourse zones has been performed across various research areas. Unfortunately, each area has constructed its own terminology in isolation, making it difficult to understand what differences exist (if any) between the various research domains. Some of the terms used to refer to *discourse zones* are: Phases of Action (Ten Have, 1989), Document Zoning (Varga et al., 2012), Rhetorical Zones (Mullen et al., 2005), Evidence Based Medicine such as PICO (Amini et al., 2012), Stages (of film reviews) (Taboada, 2011), Information Structure (Jindal, 2014; Guo et al., 2010), Argumentative Zoning (Guo et al., 2011; Guo et al., 2012; Teufel et al., 2009), Section Classification (Li et al., 2010), and Conceptualization Zones (Liakata et al., 2012). With this paper's focus being the biomedical domain (i.e., medical encounters), a comprehensive review of the discourse structure of other domains (e.g., scientific articles, film reviews, literature) is outside the current scope, and a generalized discourse zone framework is left to future work.

Mullen et al. (2005) perform an experiment to identify the document structure of biomedical texts by developing a supervised approach to classify sentences according to one of the rhetorical zones *Introduction*, *Method*, *Result*, and *Conclusion*. For discourse visualization, Guo et al. (2012) create a tool for visually displaying argumentative zones in biomedical articles. Their tool is close to our work both in objective (visual analytics) as well as implementation (use of color to distinguish between discourse zones). The objective of their visualization is to display the information structure (e.g., *Background, the Research Problem, Method, Result, Conclusion, Connection, Difference*, and *Future-work*) of biomedical articles (i.e., text), whereas our task is to visually identify medical problems, their solutions, and contexts in medical encounters (i.e., dialogue-based conversations and interviews).

Significant attention has been placed on the discourse of medical encounters, primarily for the study of doctor-patient power relationships, and how such power manifests (Wilce, 2009; Ainsworth-Vaughn, 2003). The discourse zones, referred to as *phases* or *phases of action* in medical discourse research (Byrne and Long, 1976; Ten Have, 1989), segment the dialogue of the medical encounter similar to the way scientific articles can be structured into a *Introduction-Method-Result-Conclusion* sequence. Various phase sequences (e.g., 1-2-3-5-3 can be found to be indicative of problematic encounters). One early study of medical discourse that has been influential was conducted by Byrne and Long (1976), who segment the medical encounter into six phases of actions:

1. relating to the patient
2. discovering the reason for attendance
3. conducting a verbal or physical examination or both
4. consideration of the patients condition
5. detailing treatment or further investigation
6. terminating

In terms of the Situation-Behavior-Strategy zones, Byrne and Long's typology can be reduced and mapped to our annotation scheme by combining phase 1 with 2 (Behavior), phase 3 with 4 (Situation), and phase 5 with 6 (Strategy).

4. Corpora

Our corpora consists of medical interviews between clinicians and caregivers of children with FASD. Some key properties of the corpora to note are:

- spoken dialogue between two people
- "questions & answers" interview structure
- dialogue turns are brief
- fairly unvarying in format
- entire consultation is fairly brief

These properties can be commonly found in medical encounter discourse. In contrast to conversational discourse, medical encounters in general can be characterized by a more restricted turn-taking system resulting in less interruptions (Ainsworth-Vaughn, 2003). This semi-structured framework allows medical encounters to be processed into discourse zones, unlike normal conversational discourse.

The dataset consists of 60 interviews between healthcare workers and caregivers of children with FASD. 34 have been automatically transcribed from speech to text, with 10 of the transcribed documents annotated with our discourse zone scheme consisting of the three discourse zone labels (Situation, Behavior, or Strategy), as well as a fourth label of *None* that captures dialogue not considered of import to the analyst's objective. A speaker's entire *turn* in a conversation was the basic discourse unit. Thus, each turn in the conversation was annotated with one of the four labels.

Since only one label is applied to each utterance, the issue of what label to assign in cases where one, two, or even all three of the zones are applicable to a single utterance arises. From our initial test annotations, we decided upon ranking the class labels (where zones are ranked as Behavior, Situation, and Strategy) and assigning the higher ranking zone label as the true label, and leave multi-class evaluation for future work. A subset of the FASD interview document collection is currently in the process of being manually annotated by two annotators as a pilot in order to further refine and tighten the annotation guidelines and handle any unforeseen issues.

5. Discussion

We have introduced a general annotation scheme for discourse zones in medical interviews that label each participant's turn in the interview as *Situation*, *Behavior*, and *Strategy*. The generalized scheme is very broad for application in medical interviews and potentially useful for a wide range of clinical and counselling contexts as well institutional dyad settings (attorney-witness, teacher-student, physician-patient, etc.).

Multi-class evaluation (where a discourse utterance can possibly be labelled with more than one of the Situation, Behavior, and Strategy tags) and its analysis is left for future work.

Also of interest is how well the annotation schema can be applied to other institutional dyads (attorney-witness, teacher-student, physician-patient, etc.). Further, the annotation scheme will be applied on marital conflict data, where interviews of couples are also to be analyzed in terms of the Situations, Behaviors, and Strategies that arise between them. Other corpora of interest which our annotation scheme can be applied toward are switchboard/911 phone calls.

6. References

Ainsworth-Vaughn, N. (2003). 23 the discourse of medical encounters. *The handbook of discourse analysis*, 18:453.

Amini, I., Martinez, D., and Molla, D. (2012). Overview of the ALTA 2012 Shared Task. *Population*, 7(5.6):7.9.

Byrne, P. S. and Long, B. E. (1976). Doctors talking to patients: A study of the verbal behavior of general practitioners consulting in their surgeries. *London: HSMO, Royal College of General Practitioners*.

Guo, Y., Korhonen, A., Liakata, M., Silins, I., Sun, L., and Stenius, U. (2010). Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes. pages 1–9, June.

Guo, Y., Korhonen, A., and Poibeau, T. (2011). A weakly-supervised approach to argumentative zoning of scientific documents. pages 273–283.

Guo, Y., Silins, I., Reichart, R., and Korhonen, A. (2012). CRAB Reader: A Tool for Analysis and Visualization of Argumentative Zones in Scientific Literature. pages 183–190.

Jindal, P. (2014). Information extraction for clinical narratives.

Li, Y., Lipsky Gorman, S., and Elhadad, N. (2010). Section classification in clinical notes using supervised hidden markov model. pages 744–750.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000, March.

Mullen, T., Mizuta, Y., and Collier, N. (2005). A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *ACM SIGKDD Explorations Newsletter*, 7(1):52–58.

Taboada, M. (2011). Stages in an online review genre. *Text & Talk - An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 31(2):247–269, March.

Tat, A. and Carpendale, M. S. T. (2002). Visualising human dialog. In *Information Visualisation, 2002. Proceedings. Sixth International Conference on*, pages 16–21.

Ten Have, P. (1989). The consultation as a genre. *Text and talk as social practice*, pages 115–135.

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. pages 1493–1502.

Varga, A., Preotiuc-Pietro, D., and Ciravegna, F. (2012). Unsupervised document zone identification using probabilistic graphical models. pages 1610–1617.

Wilce, J. M. (2009). Medical Discourse. *Annual Review of Anthropology*, 38(1):199–215, October.

Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese

Yu Jie Seah and Francis Bond
Linguistics and Multilingual Studies
Nanyang Technological University
yjseah1@e.ntu.edu.sg, bond@ieee.org

Abstract

A qualitative and quantitative approach was used in this study to examine the distribution of pronouns in three languages, English, Mandarin Chinese and Japanese based on the parallel NTU Multilingual Corpus (NTU-MC). The pronouns are annotated with a componential analysis that allows them to be easily linked across languages. A single text (The Adventure of the Speckled Band, a short story featuring Sherlock Holmes) in three languages is tagged, annotated and linked in the corpus. The results show that although English has the highest number of pronouns, Mandarin Chinese has the highest proportion of contentful pronouns in our corpus. Also, English has more translated counterparts in Mandarin Chinese as compared to Japanese. We attributed this to the difference in usage of pronouns in the languages. Depronominalisation, surprisingly, was even for both corpora. Findings from this study can shed some light concerning translation issues on pronoun usage for learners of the languages and also contribute to improving machine translation of pronouns.

Keywords: pronoun, Chinese, English, Japanese

1. Introduction

Pronouns exist in all the world languages, although there is considerable variation in how they are used. In this paper, we offer a componential analysis of pronouns that is extended into three language (English, Mandarin Chinese and Japanese) from three totally different language families (Indo-European, Sino-Tibetan and Japonic), The way they are employed in different languages is interesting to many linguists. Furthermore, in such a globalized world like today, languages are always translated into other languages. Other than translation of content words, how pronouns are translated from language to language can allow one to learn a lot about the language and its translation. English, being the world's most globalized language, has been translated into many different languages. Comparing its translation to Mandarin Chinese and to Japanese can shed light on the usage of pronouns in each language.

There have been few corpus based studies on differences in pronoun use among languages. According to Kim (2009), there exist qualitative and quantitative differences in the usage of the second person and first person plural pronouns in texts he examined from English and Korean newspapers. In general, English uses pronouns more often, with the notable exception of the first person plural, which was more common in Korean. Our research is part of a wider study of conceptual differences between the languages (Bond et al., 2013). For this reason, we did no restrict ourselves to personal pronouns, but also considered indefinite pronouns, demonstratives and interrogative pronouns.

2. Approach

We proceeded in four steps:

1. Identify pronouns used in the corpus
2. Analyze them in terms of components
3. Tag the pronouns monolingualy in each language
4. Analyze the distribution cross lingually

2.1. Identify the pronouns

We started off by examining words tagged as pronouns in the NTU Multilingual Corpus (NTU-MC) (Tan and Bond, 2012). The NTU-MC exploits the linguistic diversity available in Singapore for the collection of a vast variety of texts from different languages. The current version is an annotated collection of around 6,000 sentences (595,000 words) in 7 languages (Arabic, English, Mandarin Chinese, Japanese, Korean, Indonesian and Vietnamese) from 7 language families (Afro-Asiatic, Indo-European, Sino-Tibetan, Japonic, Korean (language isolate), Austronesian and Austro-Asiatic). Two kinds of annotation are applied in the NTU-MC –monolingual annotation where texts are tagged for parts of speech (POS) and sense; and crosslingual annotation where texts are aligned across sentences (Bond et al., 2013; Wang and Bond, 2014).

Pronouns from the three languages (English, Mandarin Chinese and Japanese) were extracted from four data sets in the NTU-MC. They are two short stories from Sherlock Holmes –*The Adventure of the Speckled Band* and *The Adventure of the Dancing Men* (Conan Doyle, 1892; Conan Doyle, 1905), an essay named *The Cathedral and the Bazaar* (Raymond, 1999) and on-line articles about Singapore tourism (Singapore Tourist Board, 2012). In each set, English is the source language while Mandarin Chinese and Japanese translation texts are aligned to it at the sentence level. The texts have been tokenized and automatically POS tagged.

We took as pronouns anything marked as a pronoun by the part of speech tagger.[1]. This includes personal pronouns, indefinite pronouns and interrogative pronouns. Each language had slightly different part-of-speech tags, with slightly different coverage. We ended up with 60 different English pronouns, 54 Chinese and 69 Japanese. The greater number of Japanese types reflects the greater orthographic variation: the same pronoun can be written in Chinese characters (彼 *kare* "he") or using hiragana (かれ *kare* "he").

[1] PN, WB, WRB, PRP, PRP$, WP, WP$, 名詞-代名詞-一般

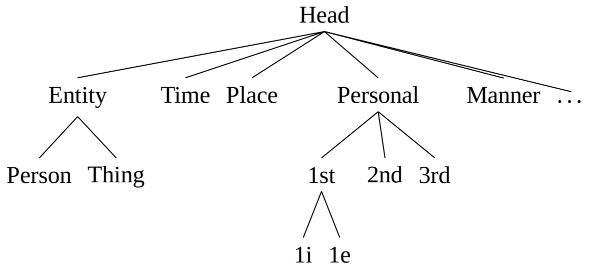


Figure 1: Head Types

2.2. Classify the Pronouns

The next stage was to analyze them componentially. The pronouns were separated into eight categories: Head, Number, Gender, Case, Type, Formality, Politeness, and Distance from Speaker. The features chosen are in line with other research and reference grammars (Backhouse, 1993; Li and Thompson, 1989; Huddleston, 1988). The purpose of this componential analysis is to code the pronouns so that we can compare and contrast them across languages. This also allows the auto-tagging programme to recognize and link the pronouns by their code. This stage took around two weeks due to the detailed componential analysis of every pronoun in the four subcorpora and analyzing ambiguous forms particularly in Japanese. The different features under each heading are shown in Table 1.

2.2.1. Head

In the first column - Head, there are altogether nine components. These are Entity, Time, Manner, Person, Place, Reason, Thing, Personal and Quantifier. This feature restricts the kind of the referent, or says that the pronoun is a quantifier and thus has no restriction. We show them in Figure 1.

Every pronoun extracted will be tagged with one of these features. For example, demonstrative pronouns such as *this* and *that* are Thing (effectively with the semantics “this thing” and “that thing”) while Entity is used for pronouns that do not have a specific category of referent, as it can refer to both person and object. Such pronouns are *all*, 俩 *lia3* “both” and いくつ *ikutsu* “some”. *when*, *how*, *why* and *where* are examples of pronouns labeled under Time, Manner, Reason and Place respectively. For English pronouns, words that end with *-thing* are easily grouped under Thing (*something, anything, nothing, . . .* , while for Mandarin Chinese and Japanese pronouns, they are not so clear-cut. Lastly personal pronouns and pronouns that talk about people like *everybody* and 自己*zi4ji3* “self” are categorized under Person.

Personal pronouns are further divided into 1st, 2nd and 3rd person. 1st person is then divided into exclusive and inclusive (used by Chinese and Indonesian, which make this distinction).

Although strictly speaking not pronouns, determiners and adjectives that are closely related to pronouns (such as *both* and *many*)were also analyzed and labeled as Quantifiers. For example, we annotate *both* in both (1) and (2) of the following two sentences. They share many of the other features, so it makes sense to analyze them together. We did not attempt to cover all determiners, only those that shared some characteristics with the pronouns.

(1) *I talked to both* Entity

(2) *I talked to both authors* Quantifier

2.2.2. Number

For this feature, we identified three kinds of number –Dual, Plural and Singular. *both* is an example of Dual, *those* for Plural and 这 *zhe4* “this” for Singular. Many pronouns are not specified for number.

2.2.3. Gender

For the third column - Gender, three features were identified as well –Masculine, Feminine and Neuter. *it* in English is a neuter pronoun while 她 *ta1* “she” in Chinese is Feminine. Most pronouns are not marked for gender.

2.2.4. Case

Only English marks case. We distinguish Subjective (nominative), Objective (accusative) and Possessive pronouns: e.g. *I, me, my*. Extending to other languages may require further distinctions.

2.2.5. Type

Type differentiates the pronouns by Assertive Existential (*somebody*), Elective Existential (*anybody*), Negative (*nobody*), Reflexive (*myself*), Reciprocal (*each other*), Universal (*everybody*), Interrogative (*who*) or Other (anything else). In a decomposed semantics, we would treat all but Reflexive and Reciprocal as quantifiers: *anybody* thus becomes the equivalent of the quantifier *any* and the noun *person*.

2.2.6. Formality

The sixth column shows Formality, whether the pronouns are informal or formal. This is mainly for the Japanese pronouns, which mark for formality: 僕*boku* “I” is informal whereas 私*watashi* “I” is formal.

2.2.7. Politeness

Japanese and Chinese also encode how respected the referent is, which we call Politeness. 您 *nin3* “you” in Chinese is used to refer to high status people.

Note that Formality and Politeness are somewhat different from the T-V distinctions made in European languages which typically only mark second person, and show the relation between speaker and hearer (historically a power difference, now more often a difference in familiarity between the speakers). Japanese pronouns encode a more absolute level of respect for their referent.

2.2.8. Proximity

The final feature is meant for pronouns that mark for Proximal, Medial or Distal distance from the speaker. These pronouns are used for demonstratives (*this* “proximal thing”, *that* “distal thing”) and by extension Place pronouns such as あそこ *asoko* “there: distal place” and time pronouns (*then* “distal time”). Chinese and English only have a two way distinction (proximal and distal: *this* and *that*). Japanese has a three way system: これ *kore* “this:proximal”,それ *sore* “that:medial” and あれ *are* “that over there:distal”.

Head	Number	Gender	Case	Type	Formality	Politeness	Proximity
Quantifier	Dual	Feminine	Objective	Assertive	Formal	Polite	Distal
Entity	Plural	Masculine	Possessive	Elective	Informal		Medial
Time	Singular	Neuter	Subjective	Negative			Proximal
Manner				Other			
Person				Reciprocal			
Place				Universal			
Reason				Interrogative			
Thing				Reflexive			
Personal (1e, 1i, 2, 3)							

Table 1: The 8 types of pronoun features

2.2.9. Summary

The features are used to define a concept, which we treat as a wordnet synset (Fellbaum, 1998). A single synset may have multiple lemmas associated with it: for example, the synset with features (Person, Assertive) has two English lemmas *someone* and *somebody*. We also linked the types to appropriate wordnet senses (for example Person is $\texttt{person}_{n:1}$, Place is $\texttt{location}_{n:1}$). The other components were kept as a separate table, linked using the wordnet synset IDs. We ended up with 107 different synsets for the 60 English, 54 Chinese and 69 Japanese pronouns.

2.3. Monolingual Tagging

After analyzing the pronouns by their different components we added them to our local wordnets' sense inventories (14 were already there, mainly interrogatives and indefinite pronouns). For English we use the Princeton Wordnet (Fellbaum, 1998), for Chinese the Chinese Open Wordnet (Wang and Bond, 2013) and for Japanese the Japanese Wordnet (Isahara et al., 2008). Treating the pronouns as synsets enabled us to use our existing wordnet tagging tools.

We carried out tagging on a single subcorpus: *The Adventure of the Speckled Band* and its Chinese and Japanese translations. We chose it as it had more (reported) speech than the other genres, and was thus had a greater variety of pronouns. We show the numbers of pronouns found in each language in Table 2. This includes all types, including quantifiers.

The main issue in the monolingual tagging was distinguishing what we shall call contentful pronouns (such as those described above) from purely structural pronouns such as dummy *it*, existential *there*, relative pronouns (*the dog who barked*) and pronouns in idiomatic expressions (*Oh My God!*). We expect contentful pronouns would introduce a quantifier into a formal semantic representation, while the structural ones would not.

In addition, there were some tokenization errors, mainly in the Chinese and Japanese corpora. These we fixed as we carried out the annotation.

2.4. Cross-lingual tagging

In the initial annotation, each pronoun was linked to the pronoun in the corresponding translation with the best feature match. If there was a tie, the leftmost pronoun pair was linked first, then the next and so on. The annotator then went through the bilingual corpus and checked each pair. At this stage they checked both whether they are tagged as pronouns correctly by the auto-tagging programme and whether the concept links between the source language and target language are accurate. This was done several times to ensure accuracy. This stage took around four weeks to complete both English-Chinese and English-Japanese corpora, with a longer time needed for the English-Chinese one due to the greater number of pronouns present there. On average, three to four sentences can be done every hour. An example of matching pronouns is given in (3) where the English is followed by Chinese. The first two English pronouns match the Chinese, the third has no equivalent.

Language	English	Chinese	Japanese
Contentful	1,370	1,177	463
Other	75	19	51
Total	1,445	1,196	514
Sentences	599	620	702
Words	11,628	12,433	13,902

Table 2: Number of pronouns found in the corpora

(3) a. You see that we have been as good as our word

b. 你 瞧， 我们 是 说到做到
ni3 qiao2, wo3men shi4 shuo1dao4zuo4dao4
的
de4

'You see, we do what (we) say'

3. Results

Having linked and tagged the relationships between words, we proceeded to count the number of pronouns in each language and their links. The number of contentful and non-contentful (structural or segmentation errors) are shown in Table 2. Differences in word and sentence tokenization give different numbers of words and sentences for the three languages, even though the content is basically the same. Even allowing for these light differences, English has more pronouns than Chinese which has far more than Japanese. The non-contentful pronouns are mainly structural for English, while they are mainly tokenization errors for Chinese and Japanese.

The results for the linkage of the pronouns are separated into two parts for better understanding — the first part being the results for the English-Chinese corpus (Table 3) and the second part for the results found from the English-Japanese corpus (Table 4).

	Linked Pronouns						Non-linked Pronouns	
	# Matching Features					Pronoun	English	Chinese
	5	6	7	8	9	to Noun		
# Pronouns	5	19	54	789	58	134	369	215

Table 3: English-Chinese pronoun translation

	Linked Pronouns						Non-linked Pronouns	
	# Matching Features					Pronoun	English	Japanese
	5	6	7	8	9	to Noun		
# Pronouns	15	120	114	37	32	139	943	109

Table 4: English-Japanese pronoun translation

There are in total 925 English to Chinese pronouns linked to each other, with 0.5% of them having only 5 pronoun features match, 2.1% having 6 pronoun features match, 5.8% having 7 pronoun features match, 85.3% having 8 features match and 6.3% having 9 pronoun features match where 9 is the maximum match. Most pronouns match everything except Case. Those that matched exactly were mainly indefinite pronouns, which don't show case.

There are also 134 pronouns that are linked to non-pronouns. 76 of them are English pronouns while 58 of them are Chinese pronouns. These typically linked to common nouns.

Out of the 1,370 contentful English pronouns, 26.9% of them are not linked. For the Chinese contentful pronouns, only 18.2% were not linked to anything.

For English and Japanese, far fewer pronouns were linked. There are in total 318 linked English to Japanese pronouns. Out of these, 4.7% have 5 matched features, 37.7% have 6 matched features, 35.8% have 7 matched features, 11.6% have 8 matched features and 10% have 9 matched features. The majority of the linked English-Japanese pronouns, unlike the English-Chinese corpus, have around 6 to 7 matched features. This is because they typically mismatch on both Case in English and Politeness or Proximity in Japanese.

Similar to the English-Chinese corpus, there are 139 pronouns in the English-Japanese corpus that are linked to non-pronouns. 109 of the pronouns are English pronouns and the other 30 are Japanese pronouns. In contrast to the English-Chinese corpus, most (68.8%) of the English contentful pronouns do not link to anything at all. Surprisingly, for the Japanese pronouns, 23.5% of them are not linked to any English words in the English source text.

4. Discussion

English has the most pronouns, followed by Mandarin Chinese and lastly Japanese. If we include non-contentful pronouns (such as dummy *it*, existential *there* and also complementizers like *that* and *which*), this becomes even more pronounced. Also, in English, many pronouns can also double up as determiners (Collins COBUILD, 2005). Determiners share many common words with pronouns such as *this*, *that* and indefinite ones such as *all* and *some*. In contrast, Chinese almost only uses contentful pronouns, and Japanese tends to drop pronouns altogether.

English personal pronouns have more different forms: Subjective, Accusative and Possessive. English also has other categories of pronouns that both Mandarin Chinese and Japanese do not have. For example, for the component Negative, English has *none* and *nothing* which do not have identical correspondents in Mandarin Chinese and Japanese: which do not negate inside noun phrases. This is because both languages tend to use verbs to express negativity instead of marking it in the pronoun like in (4) where the English is followed by Chinese and Japanese.

(4) a. . . . but none commonplace
b. 但是　却　没有　一 例 是
Dan4shi4 que4 mei2you3 yi1 li4 shi4
平淡无奇　的
ping2dan4wu2qi2 de
'But, there is not one case that is featureless.'
c. どれ も 尋常では ない事件 である
Dore mo jinjode wa nai jiken dearu
'There is not any unusual case.'

In addition, Mandarin Chinese and Japanese are topic-prominent languages (Li and Thompson, 1989; Obana, 2000). Once the topic is established, sentences following it omit any pronouns, as there is no need for them to refer back as the readers can infer from contextual knowledge the subject of the sentence.

Furthermore, out of the three languages, only Japanese marks politeness and some evidentiality on the verb (Backhouse, 1993), making the use of pronouns rather unnecessary and this seems to play an important role in reducing the numbers of pronouns found in the corpus as compared to the English source text and Chinese translation text, resulting in the low rate of links to the English pronouns in the original text. One example can be seen below in (5), with English and Japanese:

(5) a. I have heard of you, Mr. Holmes
b. あなたの ことは、以前からお聞きして
Anata no koto wa, izen kara o kiki shite
います。
imasu .
'About you, (I) humbly heard previously.'

Between the English-Chinese corpus and the English-Japanese corpus, another major difference is the number of corresponding features that majority of the linked pronouns have. For the English-Chinese corpus, majority of the linked pronouns have 8 matching pronoun features while for the English-Japanese corpus, majority of the linked pronouns have around 6 to 7 matching pronoun features. This is most likely due to Japanese language having different speech levels (Obana, 2000). The different speech levels cause a differentiation between the pronouns, resulting in Japanese having a few different words for the same pronoun. For example, for the first person pronoun, in Japanese there are variations such as わし*washi* which also marks for masculine speaker and informal and 私*watashi* which marks for formal and politeness. These features do not exist in English but from the perspective of semantics, they should be linked to the first person pronouns in English. This problem does not exist in Mandarin Chinese, as there is no such differentiation in speech levels in Mandarin Chinese. Therefore, more features can be matched.

Also, from the linking of the pronouns, there were many cases where English pronouns were linked to Mandarin Chinese and Japanese pronouns that are different in meaning such as the third person pronoun *it* in the English text to the demonstrative pronoun それ*sore* "that" or even to そこ*there* "there" in Japanese. Although this happens in the English-Chinese corpus as well, they are less frequent, thus resulting in more of the pronouns linked have more matched features as compared to those in the English-Japanese corpus. We give an example of this in (6), with English and Chinese, where *it* is linked to 这 *zhe4* "this".

(6) a. It is a swamp adder!

b. 这 是 一 条 沼地 蝰蛇！
Zhe4 shi4 yi1 tiao2 zhao3di4 kui2she2
'This is a swamp adder!'

Depronominalisation (a pronoun linking to a noun) occurs almost evenly in both the English-Chinese and English-Japanese corpora As seen in the results, the number of pronouns matched to non-pronouns in the English-Chinese corpus is around the same. This result is not expected as depronominalisation was predicted to occur much more frequently in the English-Japanese corpus than in the English-Chinese corpus. It could be a case of the source language effecting the translation: although native speakers said the translations were good, they almost certainly have more pronouns than texts written originally in Chinese or Japanese.

From the tagging of the pronouns and their concept links, there were a few interesting cases that were found. In the English source text, we realized that pronouns often exist in idiomatic phrases. However, these pronouns do not actually have any particular antecedent to refer to as they are almost always used in the same way regardless of its environment and this means that they cannot be linked.

(7) a. My God!

b. 天哪 ！
Tian1na
'Heaven!'

c. なんてこったい！
Nan te kottai
'What the heck'

We see in (7) that *my* is used here as a pronoun in an idiomatic phrase and after translation, no pronouns were seen. In both the Mandarin Chinese and Japanese text, the idiomatic translation has no pronoun in it. We give another example in (8).

(8) a. It is very kind of you.

b. 非常 感谢！
Fei1chang2 gan3xie4
'Very grateful'

c. 感謝 しているよ
Kansha shite iru yo
'(I) am grateful.'

We give one final example in (9). The Chinese translation here again choses to take the figurative meaning of *I am in your hands* and translated it to "I will obey all your instructions". However, in the Japanese text, this is literally translated, possibly because the phrase is commonly used in translating prayers and is thus somewhat established.

(9) a. I assure you that I am in your hands.

b. 我 向 你 保证， 我 一切
Wo3 xiang4 ni3 bao3zheng4, wo3 yi2qie4
听从 你 的 吩咐
ting1cong2 ni3 de fen1fu4.
'I promise you, I will obey all your instructions'

c. あなたの 手にすべてをおゆだねします
Anata no te ni subete o o yudane shimasu
わ
wa
'I will leave everything in your hands'

Another interesting note was that other than pronouns, both Mandarin Chinese and Japanese tend to use classifier phrases anaphorically. Numeral classifiers (like the *head* in *two head of cattle*) are used for most nouns in Japanese. The classifier can combine with numerals, interrogatives and in Chinese determiners. The resulting phrase can be used anaphorically: for example 那*na4jian1* "that room (CLASSIFIER)" which can mean '"hat house/room". Without the need of the proper noun in Mandarin Chinese, the determiner+classifier word can be used to refer to a certain room, thus acting like a pronoun. Although classifiers are not as widely used in English as in Mandarin Chinese and Japanese, numerals in English can sometimes take on anaphoric roles as well.

The annotation scheme we use here has two parts: the lexicon, which in this case is richly structured with components, and the corpus, which allows annotation of concepts and links between them. The two have to be kept synchronized.

5. Future Work

We would like to extend the annotation in a few ways. One is to tag more texts in the NTU-MC. The pronoun distributions in this paper are solely extracted from one story and thus we cannot generalize the results across genres.[2] The second is to add more languages to the pronoun analysis: our next language will be Indonesian, again from a different language family. We also want to extend the componential analysis to related words such as terms of address and numeral classifiers. Chinese, Japanese and Indonesian all use kinship terms to refer to non-kin: you may address a stranger as *uncle* or *older sister*.

We would also like to examine further the cases of pronouns linking to different pronouns and non-pronouns: Are the synsets always compatible? and what cues drive the choice of pronoun or demonstrative or common noun phrase? We hope that the crosslingual analysis will give some insights into the different strategies employed in the different languages.

Our distinction between contentful and structural pronouns is still only informally described. We would like to sharpen this distinction.

Finally, our analysis is compatible with (and partly inspired by) the decompositional analysis of pronouns in the English Resource Grammar (ERG), an HPSG implementation of English (Flickinger, 2000). We would like to check that all our pronouns are in the ERG and add them to the corresponding grammars of Chinese, Indonesian and Japanese. The HPSG grammars distinguish clearly between contentful and structural pronouns, and could be used to help in the monolingual annotation.

The annotated corpora and extended wordnets will be made available from the NTU-MC website: `http://compling.ntu.edu.sg/ntumc`. The corpus is licensed with the Creative Commons Attribution Only License (CC BY)[3], and the wordnets under their respective (open) licenses.

6. Conclusions

In this paper we introduced an annotation scheme for pronouns based on a componential analysis. It was tested on three languages, and used to tag a Chinese, English and Japanese tritext. The results show that pronouns, though universal, are used differently across languages, resulting in a difference in distribution among the three languages and a difference in the concept links between the English-Chinese corpus and English-Japanese corpus. We have began to account for these differences and presented examples of some interesting cases.

With this study, we hope that translation issues regarding pronoun usage would be useful and clearer to those who are learning the language and that the material from this study can contribute to pronoun translation across languages.

Acknowledgments:

This work was supported in part by the Singapore Ministry of Education (Academic Research Fund Tier 2: MOE2013-T2-1-084). Thanks to Shan Wang, David Moeljadi and an anonymous reviewer for their helpful comments.

7. References

Anthony E. Backhouse. 1993. *The Japanese Language: An Introduction*. Oxford University Press, Oxford.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013)*, pages 149–158, Sofia.

Collins COBUILD. 2005. *English grammar*. Harper Collins, 2 edition.

Arthur Conan Doyle. 1892. *The Adventures of Sherlock Homes*. George Newnes, London.

Arthur Conan Doyle. 1905. *The Return of Sherlock Homes*. George Newnes, London. Project Gutenberg `www.gutenberg.org/files/108/108-h/108-h.htm`.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1):15–28.

Rodney Huddleston. 1988. *English grammar: an outline*. Cambridge University Press, Cambridge.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Chul-Kyu Kim. 2009. Personal pronouns in English and Korean texts: A corpus-based study in terms of textual interaction. *Journal of Pragmatics*, 41:2086–2099.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

Yasuko Obana. 2000. *Understanding Japanese: A handbook for learners and teachers.* Kurusio Publishers.

Eric S. Raymond. 1999. *The Cathedral & the Bazaar*. O'Reilly.

Singapore Tourist Board. 2012. Your Singapore. Online: `http://www.yoursingapore.com`. [Accessed 2012].

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18, Nagoya.

Shan Wang and Francis Bond. 2014. Building sense-tagged multilingual corpora. In *Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik.

[2] Ideally, we would like to also annotate text with Chinese and Japanese as their source to control for translationese.

[3] `creativecommons.org/licenses/by/3.0/`

Enriching ‘Senso Comune’ with Semantic Role Sets

E. Jezek[1], L. Vieu[2], F.M. Zanzotto[3], G. Vetere[4], A. Oltramari[5],
A. Gangemi[6], and R. Varvara[7]
(1) Università di Pavia, (2) IRIT-CNRS - Université Toulouse III,
(3) Università di Roma “Tor Vergata”, (4) IBM Center for Advanced Studies,
(5) Carnegie Mellon University, (6) ISCT - CNR, (7) Università di Trento
jezek@unipv.it, vieu@irit.fr, fabio.massimo.zanzotto@uniroma2.it, gvetere@it.ibm.com,
aoltrama@andrew.cmu.edu, aldo.gangemi@cnr.it, rossella.varvara@unitn.it

Abstract

The paper describes the design and the results of a manual annotation methodology devoted to enrich the *Senso Comune* resource with semantic role sets for predicates. The main issues encountered in applying the annotation criteria to a corpus of Italian language are discussed together with the choice of anchoring the semantic annotation layer to the underlying dependency syntactic structure. We describe the two experiments we carried to verify the reliability of the annotation methodology and to release the annotation scheme. Finally, we discuss the results of the linguistic analysis of the annotated data and report about ongoing work.

1. Introduction

Large-scale linguistic resources that provide relational information about predicates and their arguments are indispensable tools for a wide range of NLP applications, where the participants of a certain event expressed by a predicate need to be detected. In particular, hand-annotated corpora combining semantic and syntactic information constitute the backbone for the development of probabilistic models that automatically identify the semantic relationships conveyed by sentential constituents in text, as in the case of Semantic Role Labeling (Gildea and Jurafsky, 2002). In addition, annotated corpora enable the quantitative and qualitative study of various linguistic phenomena at the syntax-semantics interface and the development of data-driven models for lexical semantics.

The LIRICS (Linguistic Infrastructure for Interoperable ResourCes and Systems) project has recently evaluated several approaches for semantic role annotation (PropBank, VerbNet, FrameNet, among others) and proposed an ISO (International Organization for Standardization) ratified standard for semantic role representation that enables the exchange and reuse of (multilingual) language resources. The standard comprises 29 ‘high level’ (coarse-grained) roles identified using an entailment-based methodology (Petukhova and Bunt, 2008; Gotsoulia 2011). This set has been mapped (*inter alia*) onto VerbNet roles and organized hierarchically (Bonial et al. 2011 a, b). Similar lexicons/annotation efforts include the German SALSA project (Burchardt et al. 2006), the Czech dependency treebank and its PDT-Vallex valency lexicon.

In this paper we present the design and the results of a manual annotation methodology based on the ISO-semantic roles, aiming at enriching the *Senso Comune* knowledge base of the Italian language (henceforth SC) with semantic role sets for predicates, to be used for linguistic research and NLP applications. In SC semantic roles sets are not assigned to predicates axiomatically but they are induced by the annotation of the usage examples associated with the *sensi fondamentali* (word meanings which are predominant in terms of use among the most frequent 2000 words in the language, cf. De Mauro, 1999) of the verb lemmas. The methodology encompasses annotation of the role played by participants in the event described by the predicate (intentional agent, affected entity, created entity and so on) as well as annotation of their inherent semantic properties, expressed in the form of ontological categories (person, substance, artifact, and so forth).

In the rest of the paper, we first present an overview of the SC resource, then introduce the annotation scheme and the experimental setting in which the scheme was finalized. Finally, we discuss the results of the annotations in terms of inter-annotator agreements and linguistic generalizations that can be drawn form the analysis of the data. We conclude by observing how interoperability of lexical data can also be supported formally (in the spirit of SC) in a linked data perspective.

2. Resource overview

The SC model features the main structures of standard lexicography (we refer to Vetere et al. 2012 for a general overview). These consist in lexical entries (lemmas) with their linguistic characterization and their senses. Each sense is comprised of a definition (glossa), a number of usage marks, specific grammatical constraints, usage instances, and lexicographic relations. In addition, SC provides substantive senses with ontological annotations, whose labels are taken from a foundational ontology inspired to DOLCE (Gangemi et al. 2002). The idea at the basis of ontological annotations is that linguistic senses (also referred to as *linguistic concepts*) are *tangential to reality*: they are abstract *social* entities whose relationship with extra-linguistic realities is established in the context of human activities. This idea, which comes from semiotics, calls for a formal distinction between two kinds of intensional entities: linguistic concepts (i.e. senses) and ontological categories. In fact, the ontological classification of linguistic concepts is not intended as a direct extensional interpretation over some domain of *real entities*. Instead, we resort on a notion of *ontological commitment*: a word can be used in a certain sense to refer (even vaguely, evocatively, notionally or metaphorically) to entities of some hypothetical kind.

Also, we adopt the distinction between *type* and *token* which comes form classic semiotics (Peirce); the former being abstract sorts, the latter their situated concrete instances. For instance, the Gertrude Stein's verse *a rose is a rose is a rose* counts three *rose* word tokens which instantiate FLOWER-ROSE, i.e. the (single) specific sense of *rose* occurring in the sentence, which, in turn, *commits* to the existence of objects which fall under the NATURAL-OBJECT ontological category. Note that *commits* is not to be read as *logical implication*; on the contrary, senses and ontological categories are logically disjoint, so that lexical relationships (e.g. synonymy) do not imply, nor conflict with, ontological axioms (e.g. equivalence).

3. Annotation scheme and methodology

On approaching the task of providing SC with verbal frames, we decided to start from tokens instead of types. Rather than speculating about predicate structures associated with verbal senses, we focused on annotating usage instances, as registered in the dictionary. The compilation of type-level verbal frames *à la* VerbNet is therefore deferred to a later process of generalization.

To encode the annotation of verbal predicate structures, we opted for a model based on dependencies between shallow syntactic structures, inspired to eXtended Dependency Graphs (XDG) (Basili and Zanzotto, 2002). Basically, the scheme foresees:

- the identification of flat constituents (chunks)
- the identification of the verbal chunk which conveys the exemplified sense
- the annotation of phrases which hold a thematic relation with the verb.

Argumental phrases are annotated according to the following characterization:

- each argumental chunk is given
  - a syntactic role (e.g. SUBJECT)
  - a constituent type (e.g. NP)
  - a semantic role (e.g. AGENT)
  - an ontological category (e.g. HUMAN)
- tokens of the argumental chunk are
  - (automatically) assigned a POS tag and a lemma (lemmatisation)
  - (optionally, and manually) assigned a sense (disambiguation)

Both lemmatisation and disambiguation are based on the SC dictionary. The information structure described above is encoded in a specific annotation data model (Fig. 1). This model is specified in OWL, as part of the ontology underlying the SC knowledge base [1]. Also, we provide a Java implementation which is made persistent and accessible

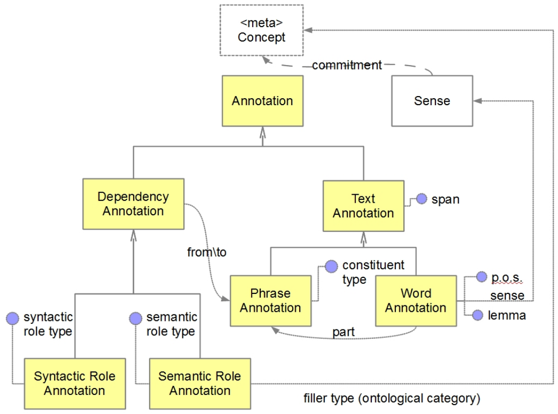


Figure 1: The Annotation Model

on relational databases through an object-relational mapping. Thus, actual annotation data are integrated in the general SC database, which allows issuing conjunctive queries where lemmas, senses, grammatical features and argument structures can be joined to extract relevant patterns.

The induction of type-level verbal frames from usage annotation data will require a process of generalization whose study is included in our future plans. To represent typical verbal frames, we plan to adopt a model in which semantics and syntactics are structurally separated, and yet logically connected. This model aims at preserving the generality of semantic structures as distinct from their syntactic realizations. Our intuition is that, by decoupling semantic and syntactic frames, one could achieve a powerful and concise representation of linguistic data, to better handle and investigate their interplay. For instance, action frames including participants and objects may be rendered in either passive or active forms; still, retrieving the lexical concepts involved in certain actions can abstract from the syntactic unfolding of verbal arguments.

In the following sections we describe the component and tags of the scheme in more detail.

3.1. Constituents and Dependency relations

We choose a light annotation scheme for syntactic dependency relations. Focusing the attentions to the verb dependency relations, we defined three types of relations: Subject (S), Object (O), and other Complement (C). We avoided the distinction, at the syntactic level, between Complement and Adjunct. This distinction is out of the scope of the syntactic phase as it is a target of the overall process of frame annotation.

As the model is inspired to the extended dependency graphs XDG) (Basili and Zanzotto, 2002), the syntactic dependency relations link constituents. We focus on the constituents that may play a role as verb arguments: Nominal Phrases (Sintagma Nominale, SN), Pronoun Phrases (Sintagma Pronominale, Spron), Prepositional Phrases (Sintagma Preposizionale, SPrep), Adverbial Phrases (Sintagma Avverbiale, SAvv), Adjectival Phrases (Sintagma Aggettivale, SAgg), and SubSentence (Sottofrase, SFr). This latter is little tricky as it is defined as a subsentence headed by a verb that is not the target verb. An example for

[1] http://www.sensocomune.org/ontologies/

SC role	LIRICS role
Agente (AG)	Agent, Partner
Causa (CAUSE)	Cause, Reason
Strumento (INSTR)	Instrument, Means
Paziente (PT)	Patient
Tema (TH)	Theme, Pivot
Goal (GOAL)	Goal
Beneficiario (BEN)	Beneficiary
Origine (SOURCE)	Source
Luogo (LOC)	Location, Setting
LuogoFinale (ENDLOC)	EndLocation
LuogoIniziale (INITLOC)	InitialLocation
Percorso (PATH)	Path
Distanza (DIST)	Distance
Tempo (TIME)	Time
TempoFinale (ENDTIME)	EndTime
TempoIniziale (INITTIME)	InitialTime
Durata (DUR)	Duration
Risultato (RESULT)	Result
Quantità (AMOUNT)	Amount
Maniera (MANNER)	Manner, Medium
Esperiente (EXP)	Pivot, Patient
Scopo (PURPOSE)	Purpose
Frequenza (FREQ)	Frequency
Attributo (ATTR)	Attribute

Table 1: Semantic roles set

the two levels of annotations is the following:

Const.	(SN Luca) ha dedicato (SN il libro) (SPrep alla madre)
Dep.	(S Luca) ha dedicato (O il libro) (C alla madre) (Luca dedicated a book to his mother)

where *Luca* and *il libro* (the book) are nominal phrases (SN) and *alla madre* (to his mother) a prepositional phrase (SPred). The three phrases play, respectively, the syntactic role of subject (S), object (O), and other complement (C).

3.2. Semantic Role list

The list of SC roles comprises 24 coarse-grained (high-level) semantic roles based on LIRICS (Petukhova and Bunt 2008) and the on-going attempt to create a unified standard set for the International Standard Initiative with the goal of facilitating mappings between semantic resource of different granularity, including VerbNet (Bonial et al. 2011 a, b). In designing the set, we conflated some LIRICS roles such as Agent and Partner (Co-Agent in VerbNet), and used some classical semantic roles like Experiencer rather than LIRICS's ambiguous Pivot. The final set of categories is given in Table 1, together with the mappings with the ISO roles of LIRICS. Each roles is defined by a gloss and a set of examples, in the LIRICS style.

3.3. Role Taxonomy

To facilitate the understanding of the scheme adopted, in addition to the glosses and the examples, semantic roles are structured into the taxonomic hierarchy of Fig. 2, in a similar way to what is done in (Bonial et al. 2011b) for LIRICS and VerbNet unified roles.

A main difference is that we have added intermediate nodes that do not count as role labels, but, with further glosses, help the annotator in understanding the main discriminating elements between roles. This enabled implementing an ontological distinction between roles that identify event participants proper, and roles that identify elements of the context of the event. As a result, some distinctions that might be difficult to grasp at first, such as Luogo Iniziale (Initial Location) vs. Origine (Source), are made clearer: in this example the first is part of the spatial context of the event, while the second is a proper and non-spatial participant to the event.

3.4. Ontological categories and TMEO methodology

In the context of *Senso Comune* we developed a tutoring system to support collaborative ontology population. As the acronym may suggest to philosophers, TMEO (Tutoring Methodology for the Enrichment of Ontologies) recalls Plato's dialectic methodology of discovering knowledge through reasoning in dialogues (Reale 1990): in this regard, by distilling the key ontological properties of SC into germane questions targeted at users, TMEO plays the role of a 'digital Socrates' in a basic interaction system. For instance, consider the scenario in which a given user is asked to classify the term *shoe*, in the sense of "footwear shaped to fit the foot (below the ankle) with a flexible upper of leather or plastic and a sole and heel of heavier material". TMEO system's interface will submit a series of intuitive conceptual questions to the users in order to disambiguate the intended meaning of the term. The following sequence represents a simplified scenario based on this example:

- TMEO: Can you touch, see, smell, taste, feel **a shoe**? User: Yes

- TMEO: Would you say that "a **shoe** can happen or occur? User: No

- TMEO: In general, does it make sense to use the word **shoe** as answer to the question "when"? User: No

- TMEO: does **shoe** indicate a location? User: No

- TMEO: Can **shoe**s act by intention? User: No

- TMEO: Would you say that **shoe**s are built by someone? User: Yes

- TMEO: **shoe** in the sense of 'footwear shaped to fit the foot (below the ankle) with a flexible upper of leather or plastic and a sole and heel of heavier material' has been classified as ARTIFACT.

As the above-mentioned scenario suggests, TMEO methodology may therefore be adopted not only in the unilateral classification of a given term ('shoe') but also in making related lexical items explicit. This kind of relatedness between terms actually unwraps the inter-categorial relation(s) holding between the corresponding ontological categories (since a detailed presentation of TMEO is out of scope in the current paper, we remand the reader to a more comprehensive publication (Oltramari et al. 2012).

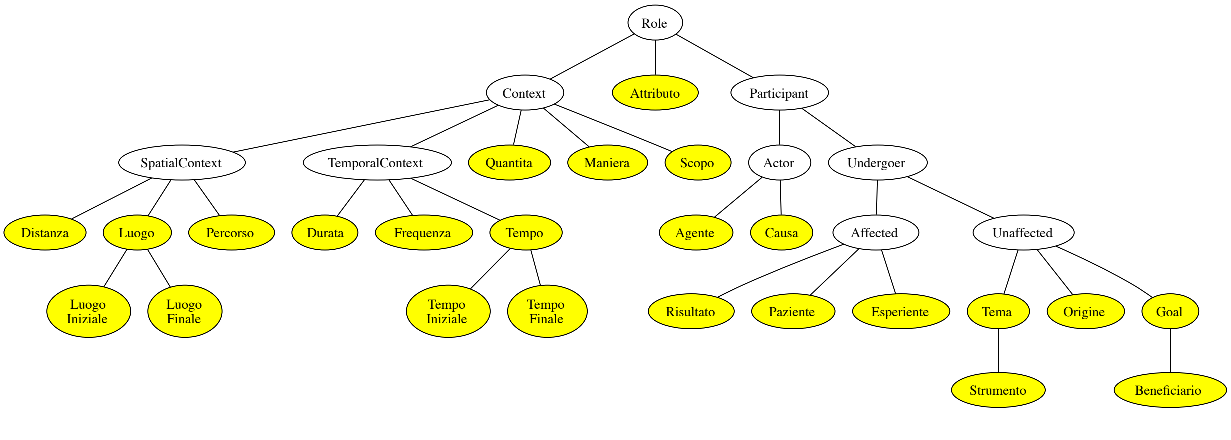


Figure 2: The semantic role taxonomy.

TMEO has been implemented as a finite state machine (FSM): in general, the elaboration process of a FSM begins from one of the states (called a 'start state'), goes through transitions depending on input to different states and must end in any of those available (only the subset of so-called 'accept states' mark a successful flow of operation). In the architectural framework of TMEO, the 'start state' is equivalent to the top-most category ENTITY, the 'transitional states' correspond to disjunctions within ontological categories and 'accept states' are played by the most specific categories of the model, i.e. 'leaves' of the relative taxonomical structure. In this context, queries represent the conceptual means to transition: this means that, when the user answers to questions like the ones presented in the above-mentioned example, the FSM shifts from one state to another according to answers driven by boolean logic[2]). If no more questions are posited to the user, this implies that the system has reached one of the available final 'accept state', corresponding to the level where ontological categories don't have further specializations. TMEO human language interface is very intuitive and comes in the form of a map where *yes/no* options are presented together with the step-by-step questions: figure 3 shows the 'shoe' example in the Italian translation 'scarpa'. In future work we aim at extending the coverage of TMEO's model and improving the scalability of the system towards genuine crowd-based platforms.

The ontological categories underlying the TMEO methodology form a taxonomy as in Fig. 4.

The annotation of ontological categories performed in the context of the work reported here differs from the annotations already present in the SC resource and described in earlier work. Here, instead of a lexical entry with its gloss, annotators were presented a text span in the context of a usage instance. In addition, they were suggested to annotate this text span with multiple categories if this was deemed more adequate than a single one. Such a possibility was introduced to acknowledge the inadequacy of a unique categorization when several interpretations co-exist due to systematic polysemy (e.g. "book" often refers simultaneously to an artifact and to an information object). Finally, the annotators were pushed to distinguish between singular and collective use of such categories. As a result, a text span like "Un ufficio" in the example "Un ufficio che funziona" ('An office that works well') can possibly be annotated POSTO+PERSONA COLLETTIVO+ORGANIZZAZIONE (Place+PersonCollective+Organization).

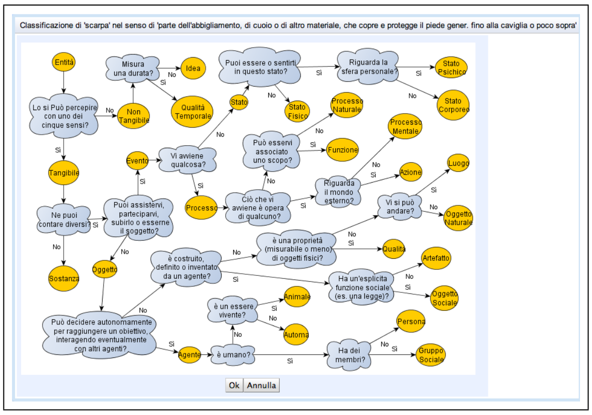

Classificazione di 'scarpa' nel senso di 'parte dell'abbigliamento, di cuoio o di altro materiale, che copre e protegge il piede gener. fino alla caviglia o poco sopra'
Entità
Misura una durata?
No
Idea
Sì
Qualità Temporale
Lo si Può percepire con uno dei cinque sensi?
No
Non Tangibile
Sì
Tangibile
Ne puoi contare diversi?
Sì
No
Sostanza
Puoi assistervi, parteciparvi, subirlo o esserne il soggetto?
Sì
Evento
No
Oggetto
Vi avviene qualcosa?
No
Stato
Sì
Processo
Puoi essere o sentirti in questo stato?
Sì
No
Stato Fisico
Riguarda la sfera personale?
Sì
Stato Psichico
No
Stato Corporeo
Può esservi associato uno scopo?
No
Processo Naturale
Sì
Funzione
Ciò che vi avviene è opera di qualcuno?
No
Sì
Riguarda il mondo esterno?
No
Processo Mentale
Sì
Azione
Vi si può andare?
Sì
Luogo
No
Oggetto Naturale
è una proprietà (misurabile o meno) di oggetti fisici?
No
Sì
Qualità
è costruito, definito o inventato da un agente?
No
Sì
Ha un'esplicita funzione sociale (es. una legge)?
No
Artefatto
Sì
Oggetto Sociale
Può decidere autonomamente per raggiungere un obiettivo, interagendo eventualmente con altri agenti?
No
Sì
Agente
è umano?
No
è un essere vivente?
Sì
Animale
No
Automa
Sì
Ha dei membri?
No
Persona
Sì
Gruppo Sociale
Ok
Annulla

Figure 3: Senso Comune's interface for TMEO

4. Annotation reliability

We verified the reliability of the annotation scheme by comparing annotations carried out by multiple annotators independently. In the following sections we describe the two pilot experiments we carried out, during which the same portion of the corpus was annotated by several participants.

4.1. Annotation experiment

We evaluated the annotation procedure in two experimental settings involving multiple annotators and estimated their agreement on the task. We selected 22 target verbs and performed multiple annotation on a set of 66 non disambiguated examples (3 for each target verb). The annotation

[2] Uncertainty will be included only in future releases of the TMEO system.

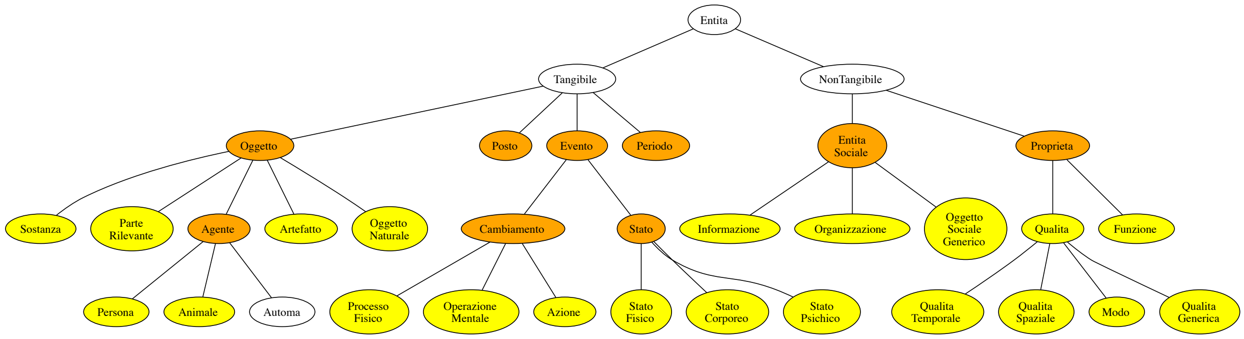


Figure 4: The ontological category taxonomy.

task was split in two subtasks. We first performed syntactic and semantic role annotation; then, we supplied the annotators with the data annotated with the sole syntactic layer, and asked them to annotate the ontological category of the argument fillers. Verbs were selected according to variability in semantic selection (for both roles and ontological categories) and syntactic realization.

4.2. Span detection

Detection the span of the verb arguments is one of the most important activity when annotating. The span of the verb argument define the sentence chunk that has to be syntactically and semantically annotated. Each annotator has to work on the same span in order to make annotations comparable. Even if the annotators decide for the same syntactic and semantic label for a nearly similar chunk of sentence, annotations cannot be compared. Thus, for comparing the annotations we assessed a gold standard, that is the most voted span for each argument.

5. Results

5.1. Interannotator agreement

The two annotation experiments were done by 9 annotators each. Among those annotators, we removed a few outliers, 1 in the first experiment and 2 in the second, for obvious misunderstanding of the task, resulting in 8 and 7 annotators respectively. We chose to use average pairwise Cohen's kappa as a measure of inter-annotator agreement, data being particularly skewed (Artstein and Poesio 2008).

For the first experiment, the inter-annotator agreement among the 8 annotators is 0.86 for the subtask on syntactic dependency relations (4 labels: 3 relations + no annotation) and 0.66 for the subtask on semantic roles (25 labels: 24 roles + no annotation). Such values are usually considered respectively as very good and fair, the latter especially so since semantic tasks are notoriously difficult.

Subgroups of annotator apparently achieved a deepest expertise, with pair agreement respectively reaching maximums of 0.91 and 0.88 on each sub-task.

In the second experiment, since we gave annotators the possibility to annotate multiple categories, there were in total 60 different labels (including no annotation). The raw agreement among the 7 annotators is quite low at 0.41. Taking into account partial agreement in the relatively few cases in which annotators used multiple categories (27 occurrences) and/or used the collective tag (36 occurrences), the agreement slightly rises to 0.46, with a pairwise maximum of 0.57. However, taking advantage of the hierarchical organization of the categories into a taxonomy, meaningful aggregation of categories can be proposed. For instance, one can reduce the 30 base-category labels in Fig. 4 actually used (only the coloured nodes have been used in the experiment), a rather large figure, into 9 labels corresponding to the orange-coloured ones on this figure. This forms a more shallow ontology, but still a meaningful discriminating one, and yields 17 different labels (with multiple categories and collectives). With such a reduction of the labels, the overall agreement clearly increases at a reasonable 0.60, with a pairwise peak at 0.79. Further analysis of the data may show where exactly annotators tend to diverge, enabling focusing on specific merges only and keeping a more fine-grained taxonomy.

6. Linguistic analysis of annotations

Besides confirming well-known difficulties in semantic role annotation, such as confusion between PT and TH due to uncertainties in the interpretation of the notions of "modification", the specificity of the annotation scheme allows us to make interesting observations regarding the role played by the semantic context, particularly the ontological category associated with the argument filler, in semantic roles annotation. This can be illustrated by focusing on the annotation of the semantic role of the subject for the 24 cases in our corpus in which there is complete agreement about the inanimate nature of referent of the filler. The first observation is that in these cases there is much more confusion between roles than average (average of kappa = 0,51). In our view this is related to the following aspects (as a reference theoretical framework cf. Pustejovsky 1995):

- there is metonymy between verb and argument in the context
- the noun is inherently polysemous
- the verb exhibits a shift in meaning
- the annotator confuses the inherent properties of the argument filler with its role.

Consider for example the case of disagreement between AG and TH (the most frequent in this set of data), that can be found in examples such as “il treno corre nella pianura a 100 all’ora” (‘the train runs in the plains at 100 Km/h’ 3AG / 5TH). In these cases, the annotator is confused by the fact that the verb in its basic meaning reports an intentional eventuality, whereas the filler in the instance is inanimate. It appears that two solutions are taken in annotation: either the filler is somewhat interpreted metonymically and assigned the AG role, or the verb is interpreted as carrying a meaning which is not the basic agentive meaning, and the subject is tagged TH.

The additional case of “Un ufficio che funziona” (‘An office that works well’ 5 AG / 3 TH) appears to be more complex, due to the inherent polysemy in the noun. In fact, in this case, we register high disagreement not only at the level of roles but also at the level of ontological categories, where *ufficio* is annotated as POSTO (‘place’, 2/7 annotators), ORGANIZZAZIONE (‘organization’, 2/7), PERSONA COLLETTIVO (‘person collective’, 2/7), POSTO+PERSONA COLLETTIVO (‘place+people’, 1/7).

In this case, one can argue that two phenomena are at play simultaneously, which confuse the annotators: the verb disambiguates the polysemous noun in context but at the same time its meaning is redefined by it (from ‘to work properly’ to ‘to perform a task well’).

Among our 24 cases, other significant cases of disagreement can be found with nouns denoting instruments. Consider the examples “la penna scrive nero” ’the pen writes black’ and “forbici che tagliano bene” ‘scissors that cut well’, that have been annotated as INSTR by 3/8 and 4/8 respectively (*pen* was further tagged as TH by 5/8, while *scissors* as TH by 3/8 and AG by 1/8). These subjects (called *Instrument subjects* in literature, see e.g. Alexiadou et Schäfer 2006) refer to entities frequently used as facilitating instruments in everyday life (as expressed in sentences like “I wrote the letter with a fountain pen”, “I used the scissors to open the package”), but in the examples above they are not presented as instruments, but rather as the entity about which the verb predicates something (that is, they have the characteristic of writing and cutting). Nobody uses them to perform an action; hence, they are THs because they are the participants in the condition described by the verb and are not modified by the event. We argue that in these cases annotators who tag them INSTR confuse the ontological type of the entity denoted by the filler with the semantic role the participant plays in the event.

7. Interoperability of Semantic Roles on the Semantic Web

SC has been formally represented in OWL, and this offers an opportunity to make it interoperable at both synset level (through an ongoing alignment to the Italian version of MultiWordNet, which will be part of the Lexical Linked Data Cloud), and at semantic role level, by aligning it to the VerbNet and FrameNet RDF datasets.

Recently, the problem of interoperability between different linguistic ontologies (schemas for representing linguistic data) has entered the Semantic Web and Linked Open Data radar, since there are mutual advantages in creating linguistic data expressed in RDF (the basic language for the Semantic Web): the Web as an integration platform for heterogeneous linguistic data, as well as easier support for lexicalizing ontologies.

In that context, several initiatives are boosting the adoption of good practices for sharing linguistic data, and make them interoperable at a formal level. NLP Interchange Format (NIF) is an RDF/OWL-based format that allows to combine and chain several NLP tools in a flexible, light-weight way. The Linguistic Linked Open Data initiative is linking many linguistic datasets, but it is still missing a tight integration of lexical resources including semantic roles. FrameNet and VerbNet have been ported to RDF and OWL (cf. Nuzzolese et al. 2011 for FrameNet-OWL), including the mapping between FrameNet frames and VerbNet predicates, but this is not yet extended to the respective role structures. The OntoLex W3C Community Group is going to publish a proposal for a standard to describe lexical resources jointly with ontologies and linked datasets (where the basic innovation is to allow for a sense layer distinguished from lexical expressions and ontological entities, which enables intensional semantics of lexical resources to be used in the mostly extensional formal semantics assumed in the Semantic Web).

The potential of the Semantic Web for semantic role labeling (and vice versa) is exemplified by the FRED architecture (Presutti et al. 2012), where VerbNet roles are used to automatically annotate RDF graphs that are extracted from text by means of multiple NLP algorithms (semantic role labeling, frame detection, relation extraction, sense disambiguation, named entity recognition).

FRED allows to link those graphs to linked data resources; it aligns named entities to linked data resources, as well as named concepts (typically derived from disambiguated terms) to WordNet or DBpedia resources. Since RDF resources are usually typed, FRED graphs can be used for investigating the actual coverage of VerbNet roles, with their associated types (à la selectional restrictions). In fact, FRED complements partial coverage of VerbNet with other roles, e.g. directly expressed by prepositions, which can be further investigated.

8. Conclusions

In this paper, we described the design of a manual annotation methodology devoted to enrich the SC resource with semantic role sets for predicates. We discussed the results of the two experiments performed to verify the reliability of the annotation methodology, in terms of inter-annotator agreement and linguistic generalizations that can be drawn form the analysis of the data. For the future, we plan to perform automatic chunking of the data to be annotated and check it manually before annotation; to annotate the ontological category of the argument fillers out of context; to develop a methodology for extraction of semantic roles sets for predicates from the annotated data; to link SC semantic roles sets to other lexical resources for Italian such as T-PAS structures (Jezek et al. 2014).

Acknowledgments

We thank three anonymous reviewers for their useful comments.

References

A. Alexiadou and F. Schäfer. 2006. Instrument Subjects Are Agents or Causers. In D. Baumer, D. Montero, and M. Scanlon (eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics*, 40-48. Somerville, MA: Cascadilla Proceedings Project.

A. Burchardt , E. Katrin , A. Frank , A. Kowalski , S.Padó, M. Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In Proceedings of LREC 2006.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. In *Computational Linguistics*. 34, 4, 555-596.

R. Basili and F.M. Zanzotto. 2002. Parsing engineering and empirical robustness. *Natural Language Engineering*, 8/2-3.

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

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

T. De Mauro. 1999. Introduzione. In De Mauro T. (Ed. in Chief), *Grande Dizionario Italiano dell'Uso (GRADIT)*, 6 voll. + CD-rom, Torino, UTET, vol. I, VI-XLII.

D. Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language* 67.3, 547-619.

A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, L. Schneider. 2002. Sweetening Ontologies with DOLCE. *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 02).*

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

V. Gotsoulia. 2011. An Abstract Schema for Representing Semantic Roles and Modelling the Syntax-Semantic Interface. *Proceedings of the Ninth International Conference on Computational Semantics (IWCS '11)*, Association for Computational Linguistics Stroudsburg, PA, USA, 115-124.

E. Jezek, B. Magnini, A. Feltracco, A. Bianchini and O. Popescu. 2014. T-PAS: A resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing. in *Proceedings of LREC 2014* (to appear).

A.G. Nuzzolese, A. Gangemi, V. Presutti. 2011. Gathering Lexical Linked Data and Knowledge Patterns from FrameNet. In O. Corcho, M. Musen (eds.) *Proceedings of K-CAP 2011 The Sixth International Conference on Knowledge Capture*, ACM.

A. Oltramari, A. Mehler, K. Kühnberger, H. Lobin, H. Lüngen, A. Storrer, A. Witt. 2012. An Introduction to Hybrid Semantics: The Role of Cognition in Semantic Resources. In *Modeling, Learning, and Processing of Text Technological Data Structures*, Berlin / Heidelberg, Springer, vol. 370, 97-109.

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

V. Presutti, F. Draicchio and A Gangemi. 2012. Knowledge Extraction based on Discourse Representation Theory and Linguistic Frames. In A. ten Teije and J. Vlker (eds.) *Proceedings of the Conference on Knowledge Engineering and Knowledge Management* (EKAW2012), LNCS, Springer.

J. Pustejovsky. 1995. *The Generative Lexicon*. Cambridge MA: MIT Press.

G. Reale. 1990. *A History of Ancient Philosophy: Plato and Aristotle*. SUNY Press.

G. Vetere, A. Oltramari, I. Chiari, E. Jezek, L. Vieu, F.M. Zanzotto. 2012. 'Senso Comune': An Open Knowledge Base for Italian. In *Revue TAL (Traitement Automatique des Langues)*, Journal Special Issue on *Free Language Resources* 52.3, 217-43.

Automatic Tagging of Modality: identifying triggers and modal values

Paulo Quaresma[1,4], **Amália Mendes**[2], **Iris Hendrickx**[2,3], **Teresa Gonçalves**[1]

[1] Department of Informatics, University of Évora, Portugal
[2] Center for Linguistics at the University of Lisbon, Portugal
[3] Center for Language Studies, Radboud University Nijmegen, The Netherlands
[4] L2F – Spoken Language Systems Laboratory, INESC-ID, Portugal

Abstract

We present an experiment in the automatic tagging of modality in Portuguese. As we are currently lacking a suitable resource with detailed modal information for Portuguese, we experiment with small sample of 160.000 tokens, manually annotated according to the modality scheme that we previously developed for European Portuguese (Hendrickx et al., 2012). We consider modality as the expression of the speaker (or subject)'s attitude towards the proposition and our modality scheme accounts for seven major modal values, and nine sub values. This experiment focuses on three modal verbs, *poder* 'may/can', *dever* 'shall/might' and *conseguir* 'manage to/ succeed in/ be able to', which may all have more than one modal value. We first report on the task of correctly detecting the modal uses of *poder* and *dever*, since these two verbs may have non modal meanings. For the identification of the modal value of each occurrence of those three verbs, we applied a machine learning approach that takes into consideration all the features available from a syntactic parser's output. We obtained the best performance using SVM with a string kernel and the system improved the baseline for all three verbs, with a maximum F-score of 76.2.

Keywords: modality, annotation scheme, automatic tagging

1. Introduction

As the vast amount of digitally available data keeps growing, so does the demand to automatically extract relevant information. A clear problem for automatic extraction tools is to recognize the factual or non-factual nature of events, and the subjective perspective underlying the texts. In this paper we focus on modality: an important indicator of subjectivity and factuality in text. Modality is usually defined as the expression of the speaker's opinion and of his attitude towards the proposition (Palmer, 1986). It traditionally covers epistemic modality, which is related to the degree of commitment of the speaker to the truth of the proposition (whether the event is perceived as possible, probable or certain), but also deontic modality (obligation or permission), capacity and volition. Modality detection is therefore also clearly linked to the current trend in NLP on sentiment analysis and opinion mining.

This paper presents an experiment in the automatic tagging of modality in Portuguese. Not much related work has been done in this area, certainly not for languages other than English. A prerequisite for building an automatic modality tagger is to have a corpus with labeled examples to train and evaluate such tool. As we are currently lacking a large and suitable corpus, one of the main aims of the study presented here is to create a tagger on a small corpus sample in order to (semi) automatically tag a larger corpus with modality information. For this purpose, we use a corpus of 158.553 tokens, manually annotated with a modality scheme for Portuguese (Hendrickx et al., 2012b). In this paper, we restrict our experiment to three modal verbs: *poder* 'may/can', *dever* 'shall/might' and *conseguir* 'manage to/ succeed in/ be able to'. These three verbs are high frequent words in Portuguese and have different modal meanings, what makes them an excellent study object for our experiments.

The automatic modality tagger that we devised has two objectives: the identification of modal verbs (which we call the modal trigger) and the attribution of a modal value to this trigger. All three verbs have two or more modal meanings: for example, *poder* may be Epistemic, stating that something is possible, as in example (1); Deontic, denoting a permission, as in (2); or it may express an Internal capacity, the fact that someone is able to do something, as in (3). And frequently, a single context may be ambiguous between one and more of these readings.

(1) E é evidente que um jogador que arrisque **pode** vir a ser apanhado mas, sem a certeza do controlo, a minha opinião é de que vai ter tendência para arriscar mais.

'It is obvious that a player that takes risks might be caught but, without the certainty that there will be a control, in my opinion he will tend to take more risks.'

(2) Segundo Cândida Almeida, "os jornalistas não **podem** usar meios que a própria lei veda a polícias e magistrados em nome dos direitos, liberdades e garantias dos cidadãos".

'According to Cândida Almeida, "the journalists can not use means that the law itself forbids to the police and to prosecutors in the name of the citizen's rights, liberties and warranties.

(3) Os deputados portugueses, para serem ouvidos e terem influência, precisam de **poder** comunicar facilmente com os seus colegas, o que implica, num ambiente genuinamente multilinguístico, o domínio de várias línguas estrangeiras (…).

'The Portuguese representatives to the European Parliament, to be heard and to have influence, need to be able to communicate easily with their colleagues, what implies, in a genuinely multilingual environment, the mastery of several foreign languages.'

This polysemy increases the level of difficulty of the automatic annotation task. To create the modality tagger, we first automatically assign POS and syntactic tags, we then automatically identify modal triggers and apply a machine learning approach to attribute a modal value to the triggers, comparing the results with our gold dataset of 158.553 tokens.

The paper is structured as follows: we first revise related work in section 2, before briefly presenting our modality scheme and golden dataset in 3. Our automatic annotation system is described in section 4, the results of trigger identification are presented in 5.1 and the results of automatic attribution of modal value in 5.2, followed by a conclusion in 6.

2. Related work

Several annotation schemes of modality have been proposed in recent years, such as Baker et al. (2010), Matsuyoshi et al. (2010); Saurí et al. (2006), Nirenburg and McShane (2008) and, for Brazilian Portuguese, Ávila and Melo (2012). We will not discuss here in detail the differences between those annotation schemes (see Hendrickx et al. (2012b) and Nissim et al. (2013)) but rather focus on some experiments in the automatic annotation of modality that have been reported, mainly for English. Baker et al. (2010) tested two rule-based modality taggers to identify the modal trigger and its target and report results of 86% precision for tagging of a standard LDC data set. Also, Saurí et al. (2006) report on the automatic identification of events in text, and their characterization with modality features, achieving accuracy values of 97.04 with the EviTA tool. Battistelli and Damiani (2012) aim to annotate textual segments that have enunciative and modal (E_M) features. They use semantic clues to identify modal triggers and a syntactic parser to calculate the length of the E_M segment. However, the implementation of the system is an upcoming work. A specific system for the annotation of belief is reported by Diab et al. (2009). The authors mention that they treat all auxiliary verbs as epistemic, although they are aware of the fact that they may be deontic, and consider that this might be a source of noise in their system (an aspect that we also have to deal with). An extension of this experiment is reported in Prabhakaran et al. (2012), testing the tagging of different modality values (Ability, Effort, Intention, Success and Want). The authors report experiments on MTurk annotations (using only those examples for which at least two Turkers agreed on the modality and the target of the modality) and on a gold dataset, with respectively an overall 79.1 and 41.9 F-measure. It is important to mention that the corpora for both experiments differ greatly: MTurk data is entirely from email threads, whereas Gold data contains sentences from newswire, letters and blogs in addition to emails.

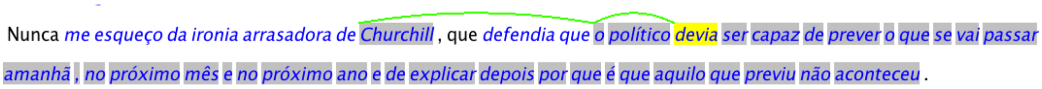
The work of Ruppenhofer and Rehbein (2012) is close to our own objectives in this paper. The authors report an experience to automatically identify five English modal verbs (*can*/*could*, *may*/*might*, *must*, *ought*, *shall*/*should*) in texts and predict their modal value, by training a maximum entropy classifier on features extracted from the training set. The authors manage to improve the baseline for all verbs but *must*, and achieve accuracy numbers between 68.7 and 93.5.

The detection of uncertainty and its linguistic scope was the subject of a shared task at CoNLL2010 (Farkas et al., 2010) focusing on hedging clues, which includes a broader set of lexical and syntactic clues than modality as we contemplate it in this paper. The area of BioNLP includes modality and factuality in the annotation of events: the dimension "level of certainty" is part of the system of meta-knowledge assignment to pre-recognised events described in Miwa et al. (2012), which attains F-measures of 74,9 for "low confidence" and 66,5 for "high but not complete confidence".

3. Annotation Scheme and Corpus

The annotation scheme for Portuguese presented in Hendrickx et al. (2012a) is not restricted to modal verbs and also covers nouns, adjectives and adverbs. Modality is understood as the expression of the speaker's attitude towards the proposition. So, the concept of factuality is not included, contrary to approaches such as Nissim et al. (2013), who accounts for both values but in different layers of the annotation scheme. Furthermore, our annotation scheme does not account for verb tense and mood, although this category is related to modality. The approach is very similar to the OntoSem (Mcshane et al., 2005) annotation scheme for modality (Nirenburg and McShane, 2008).

We include several modal values, based on the modality literature, but also on studies focused on annotation and information extraction (e.g. Palmer (1986); van der Auwera and Plungian (1998); Baker et al. (2010)). Seven main modal values are considered (Epistemic, Deontic, Participant-internal, Volition, Evaluation, Effort and Success), and several sub-values. There are five sub-values for epistemic modality: Knowledge, Belief, Doubt, Possibility and Interrogative. Contexts traditionally considered of the modal type "evidentials" (i.e., supported by evidence) are annotated as Epistemic belief. Two sub values are identified for deontic modality: Deontic obligation and Deontic permission (this includes what is sometimes considered Participant-external modality, as in van der Auwera and Plungian (1998)). Participant-internal modality is subdivided into Necessity and Capacity. Four other values are included: Evaluation, Volition and, following Baker et al. (2010), Effort and Success. We present the list of values and sub values in Table 1, together with their frequency in our golden set.



Nunca me esqueço da ironia arrasadora de Churchill , que defendia que o político devia ser capaz de prever o que se vai passar amanhã , no próximo mês e no próximo ano e de explicar depois por que é que aquilo que previu não aconteceu .

Figure 1: Screenshot of MMAX2 annotation tool

Main modal values	Sub values	Freq	%
Epistemic			
	knowledge belief doubt possibility interrogative	183 161 29 279 87	7,1 6,3 1,1 10,9 3,4
Deontic			
	obligation permission	581 159	22,7 6,2
Participant-internal			
	capacity necessity	126 122	4,9 4,8
Evaluation		159	6,2
Volition		396	15,4
Effort		110	4,3
Success		119	4,6

Table 1: Modal values and frequencies in our golden set

The annotation scheme comprises several components: (a) the trigger, which is the lexical element conveying the modal value; (b) the target; (c) the source of the event mention (speaker or writer) and (d) the source of the modality (agent or experiencer). The trigger receives an attribute *modal value*, while both trigger and target are marked for polarity. An example with the verb *dever* is given in (4)[1]. In fact, the example sentence in (4) contains three other triggers as well. In this particular context, the trigger *esqueço* 'I forget' expresses the modal value Epistemic knowledge, the trigger *defendia* 'argued' expresses Epistemic belief, and the trigger *capaz* 'be able' expresses Participant-internal capacity. In example (4) however we focus on the annotation of the trigger *dever* in more detail.

(1) Nunca me esqueço da ironia arrasadora de Churchill, que defendia que o político devia ser capaz de prever o que se vai passar amanhã, no próximo mês e no próximo ano e de explicar depois por que é que aquilo que previu não aconteceu.

'I never forget the devastating irony of Churchill, who argued that a politician should be able of predicting what is going to happen tomorrow, next month and next year and then explain why what he had predicted didn't happen.'

Trigger: devia
  Modal value: deontic_obligation
  Polarity: positive
Target: o politico@ ser capaz de prever o que se vai passar amanhã, no próximo mês e no próximo ano e de explicar depois por que é que aquilo que previu não aconteceu
Source of the modality: Churchill
Source of the event: writer
Ambiguity: none

This annotation scheme was applied to a corpus sample extracted from the written subpart of the Reference Corpus of Contemporary Portuguese (CRPC) (Généreux et al, 2012). Details about the selection of the sample are provided in Hendrickx et al (2012b). We used the MMAX2 annotation software tool (Müller and Strube, 2006) for our manual annotation task. The MMAX2 software is platform-independent, written in java and can freely be downloaded from http://mmax2.sourceforge.net/. The elements of our annotation consist of markables that are linked to the same modal event, which we call a "set". We present a screenshot of the results in Figure 1. The trigger *devia* and related markables are connected under a single set and are highlighted.

Full details on our annotation scheme and on the results of an inter-annotator experiment are provided in Hendrickx et al. (2012b). An enriched version with the interaction between Focus and Modality, specifically the case of exclusive adverbs, is presented in Mendes et al. (2013).

In the experiments that we present here, we focus on the Trigger component and its attribute *modal value*, and specifically on three semi-auxiliary modal verbs. The frequency of the modal verbs in our data set and their values are presented in Table 2.

The verb *dever* has two modal values in our golden set: Deontic obligation and Epistemic possibility. The value Participant-internal capacity is also possible with this verb but was never selected in our data as the primary meaning, although manual annotators have marked it in the 'Ambiguity' field of our annotation system in several cases. For this experiment, we didn't take into consideration cases marked as ambiguous but this is certainly an important aspect to tackle in future research. Our experiments will therefore focus on five modal values: Deontic obligation, Deontic permission, Epistemic possibility, Participant-internal capacity and Success.

[1] Notice that the discontinuity of the target is marked with the symbol @ in the example, but is encoded in XML in our data set.

Main values	**Sub values**	**Freq.**
dever		**113**
	Deontic obligation	74
	Epistemic possibility	39
poder		**244**
	Deontic permission	43
	Epistemic possibility	158
	Participant-internal capacity	44
conseguir		**84**
	Participant-internal capacity	41
	Success	43

Table 2: Frequency of *dever*, *poder* and *conseguir* in our gold dataset.

4. Modality tagging

Our automatic modality tagger is composed by three modules:

- Syntactic analysis of the corpus;
- Identification of the modal verbs *poder*, *dever*, *conseguir*;
- Labeling of each verb with the appropriate modal value in its specific context.

The syntactic analysis was performed by the PALAVRAS parser (Bick, 1999), and the results were transformed into XML and logical terms (Prolog format) using the tool Xtractor (Gasperin et al., 2003). We then selected the set of parsed sentences that included the modal verbs and distinguished the modal uses of the verbs from the non-modal ones. As we aim to use this tagger to create a larger corpus, this first step of finding the modal triggers needs to be performed with very high accuracy.

We then used SVM, Support Vector Machines (Vapnik, 1998), to classify the modal value of each verb. We evaluated several machine learning algorithms and SVM kernel types with Weka (Hall et al., 2009), and obtained the best performance using SVM with a string kernel (Lodhi et al., 2002). We report the results obtained in two experiments: one using just the original sentences and another using the POS tags and functional and syntactic information extracted from the sentence's parse tree, in a window of 70 characters around the verb. For the evaluation we used a 10-fold stratified cross-validation procedure. Note that this is a challenging task as we only have a few hundred examples to train and test the automatic tagger. We analyze the results in the next section.

5. Results

5.1 Modal verb detection

Here we first discuss to what extent we were able to correctly detect the modal verbs based on the output of the automatic syntactic parser. The verbs *poder* and *dever* may occur with non-modal uses, therefore the task involves the correct identification of contexts that are indeed modal. The case of the verb *conseguir* is different because it always involves one of the modal values contemplated in our annotation system. For this specific verb, the system has to correctly identify sentences containing the lemma in the results of the parser, a much simpler task. Taking this into consideration, we will only discuss the results obtained for the verbs *poder* and *dever*, and compare our system's output with the manually tagged information. This is summarized in Table 3.

	poder	*dever*
total verb occurrences	258	120
modal occurrences	244	113
automatic identification	236	108
false positives	0	0
error rate	3.1	4.2
precision	100	100
recall	96.7	95.6
F-measure	98.3	97.7

Table 3: Results of modal verb detection

Data from Table 3 show that the error rate in the identification of the modal occurrences is quite low: 3.1 for *poder* and 4.2 for *dever*. Precision receives the maximum value and Recall is above 95 for the two verbs. Errors are due to complex Portuguese sentences causing parsing problems, especially contexts where the semi-auxiliary modal verbs and the main verb are distant in the sentence. Another difficulty of the parser is to deal with cases where the semi-auxiliary modal is followed by a pronominal clitic. These issues could be partially dealt with in an additional post-processing step and would possibly result in an improvement of our performance in the future. However, syntactic complexity will remain a difficult challenge for semi-auxiliary detection.

5.2 Attribution of modal value

To identify the modal value, we applied a machine learning approach to the sentences detected by the previous module. Our system takes into consideration all the features available from the PALAVRAS output: lemma and POS of the trigger, left and right syntactic context, and semantic features: predicate argument structure, [±human] nature of arguments. We also computed scores for a baseline system that always assigns the most frequent modal value for each verb.

The results for both experiments (using the sentences and a text linearized format of the parse tree within a window around the verb) are presented in Table 4 (for *dever*), Table 5 (for *poder*) and Table 6 (for *conseguir*). We give results for a baseline and for both experiments (sentences and window parse tree), computing Precision (P), Recall (R) and F-value (F) and the macro-average over the different modal values.

dever		baseline			sentences			window parse tree		
	count	P	R	F	P	R	F	P	R	F
Total/macro-average	108	32.9	50.0	**39.7**	65.6	63.8	64.3	65.7	64.5	**64.9**
deontic obligation	71	65.7	100	79.3	74.4	81.7	**77.9**	75.0	80.3	77.6
epistemic possibility	37	0	0	0	56.7	45.9	50.7	56.3	48.6	**52.2**

Table 4: Results of the automatic modal value attribution for *dever*

poder		baseline			sentences			window parse tree		
	count	P	R	F	P	R	F	P	R	F
total/macro-average	236	21.8	33.3	**26.3**	34.6	33.4	32.2	34.3	34.0	**33.7**
deontic permission	42	0	0	0	23.1	7.1	10.9	18.8	14.3	**16.2**
epistemic possibility	154	65.3	100	79.0	64.6	80.5	**71.7**	65.5	75.3	70.1
participant internal capacity	40	0	0	0	16.1	12.5	14.1	18.5	12.5	**14.9**

Table 5: Results of the automatic modal value attribution for *poder*

conseguir		baseline			sentences			window parse tree		
	count	P	R	F	P	R	F	P	R	F
total/macro-average	84	25.6	50.0	**33.9**	57.1	57.0	56.8	76.3	0,762	**76.2**
participant internal capacity	41	0	0	0	57.1	48.8	52.6	76.9	73.2	**75.0**
success	43	51.2	100	67.7	57.1	65.1	60.9	75.6	79.1	**77.3**

Table 6: Results of the automatic modal value attribution for *conseguir*

The results in Tables 4-6 show that our system was able to improve the baseline for all three verbs: for *dever* it improves the baseline from 39.7 to 64.7 macro-average F-value, for *poder* from 26.3 to 33.7 and for *conseguir* from 33.9 to 76.2. The higher values attained for *conseguir* are tied to the fact that its two modal values have similar frequencies in our gold dataset, making it easier to improve the baseline.

With these experiments we obtained macro-average F-values between 33.7 and 76.2. We obtain better performance measures for *conseguir* and *dever* than for *poder*, possibly because *poder* has three modal values. Obviously, the automatic tagger obtains the best results for the most frequent values.

Comparing the experiments using the sentences and the window parse tree, the results show no significant differences, although the window parse tree experiment generally presents higher results, especially with *conseguir* (F-value 76.2 vs. 56.5).

6. Conclusion

We have presented a system for the automatic tagging of modality in Portuguese, using a manually annotated corpus as training data. The identification of the modal instances of the three auxiliary verbs receives high recall and precision values and could be further improved at the parsing level. The results of the attribution of the modal value reach macro-average F-measures between 33 and 76 % F-value depending on the modal verb and on the modal value. The results are promising, considering that we trained our system on a tiny data set, and suggest that our aim: creating a larger corpus with modal information by a (semi) automatic tagging process based on a small sample seems to be a feasible next step.

In future work we plan to provide a detailed study identifying the individual role of the syntactic and semantic features that play a role in the automatic attribution of the modal value in our system. Another goal is to apply the modality tagger to a larger set of verbs to see whether we can keep a reasonable performance for a more diverse set of verbal triggers. We also aim to compute a learning curve to estimate the amount of manually annotated examples that are needed to get a good performance from the modality tagger.

As we are currently applying a 'word expert' approach and training separate classifiers for different verbal triggers, it is clear that this approach will not be able to handle modal triggers that it has not seen before. As a next step we will study this problem and for example try to train a general modal trigger classifier that is not dependent on the verb itself.

Acknowledgements

This work was partially supported by national funds through FCT – Fundação para a Ciência e Technologia, under project Pest-OE/EEI/LA0021/2013 and project PEst-OE/LIN/UI0214/2013. We would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

Luciana Ávila and Heliana Melo (2013) Challenges in modality annotation in a Brazilian Portuguese Spontaneous Speech Corpus, *Proceedings of IWCS 2013 WAMM Workshop on the Annotation of Modal Meaning in Natural Language*, March 19-20, 2013, Postdam, Germany.

Kathrin Baker, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of LREC'10*, Valletta, Malta. ELRA, 1402-1407.

Delphine Battistelli and Marine Damiani. 2013. Analyzing modal and enunciative discursive heterogeneity: how to combine semantic resources and a syntactic parser analysis. In *Proceedings of WAMM-IWCS2013*, Potsdam, Germany.

Eckhard Bick. 1999. The parsing system PALAVRAS. *Aarhus University Press.*

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, Suntec, Singapore, August. Association for Computational Linguistics, 68–73.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, ACL, 1-12.

Caroline Gasperin, Renata Vieira, Rodrigo Goulart, and Paulo Quaresma. 2003. Extracting XML syntactic chunks from portuguese corpora. In *TALN'2003 - Workshop on Natural Language Processing of Minority Languages and Small Languages of the Conference on "Traitement Automatique des Langues Naturelles"*, Batz-sur-Mer, France, June 2003.

Michel Généreux, Iris Hendrickx, and Amália Mendes. 2012. Introducing the Reference Corpus of Contemporary Portuguese On-Line". In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* - LREC 2012, Istanbul, May 21-27, 2012, 2237-2244.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1.

Iris Hendrickx, Amália Mendes, Silvia Mencarelli, and Agostinho Salgueiro. 2012a. *Modality Annotation Manual*, version 1.0. Centro de Linguística da Universidade de Lisboa, Lisboa, Portugal.

Iris Hendrickx, Amália Mendes, and Silvia Mencarelli. 2012b. Modality in Text: a Proposal for Corpus Annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* - LREC 2012, Istanbul, May 21-27, 2012, 1805-1812.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Christopher J. C. H. Watkins. 2002. Text Classification using String Kernels. *Journal of Machine Learning Research.* 2:419-444.

Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating event mentions in text with modality, focus, and source information. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. ELRA.

Marjorie McShane, Sergei Nirenburg, Stephen Beale, and Thomas O'Hara. 2005. Semantically rich human-aided machine annotation. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II*: Pie in the Sky. ACL, 68-75.

Amália Mendes, Iris Hendrickx, Agostinho Salgueiro, and Luciana Ávila. 2013. Annotating the Interaction between Modality and Focus: the case of exclusive particles. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse (LAW VII).* Association for Computational Linguistics, Sofia, Bulgaria, August 8-9 2013, 228-237.

Makoto Miwa, Paul Thompson, John McNaught, Douglas B Kell and Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics* 13:108.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, 197-214. Peter Lang.

Malvina Nissim, Paola Pietrandrea, Andrea Sansò and Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. In Harry Bunt (ed.) *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation* (isa-9), March 19-20, 2013, Postdam, Germany, 7-14.

Sergei Nirenburg and Marjorie McShane. 2008. Annotating modality. Technical report, University of Maryland, Baltimore County, March 19, 2008.

Frank R. Palmer. 1986. *Mood and Modality*. Cambridge textbooks in linguistics. Cambridge University Press.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM '12).* Association for Computational Linguistics, Stroudsburg, PA, USA, 57-64.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? Annotating the senses of English modal verbs. In *Proceedings of the 8th International Conference on*

Language Resources and Evaluation (LREC), May 24-26, 2012, Istanbul, Turkey, 1538-1545.

Roser Saurí, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of the 19th International FLAIRS Conference*.

Johan Van der Auwera and Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology*, 2(1): 79-124.

Vladimir Vapnik. 1998. Statistical learning theory. Wiley, NY.

Using the Crowd to Annotate Metadiscursive Acts

Rui Correia[1,2], Nuno Mamede[2], Jorge Baptista[2,3], Maxine Eskenazi[1]

[1] LTI - Carnegie Mellon University, Pittsburgh, USA
[2] INESC-ID, Lisbon, Portugal
[3] Universidade do Algarve, Faro, Portugal
rcorreia@cs.cmu.edu, Nuno.Mamede@inesc-id.pt, jbaptis@ualg.pt, max@cs.cmu.edu

Abstract

This paper addresses issues relating to the definition and non-expert understanding of metadiscursive acts. We present existing theory on spoken metadiscourse, focusing on one taxonomy that defines metadiscursive concepts in a functional manner, rather than formally. A crowdsourcing annotation task is set up with two main goals: (a) build a corpus of metadiscourse, and (b) assess the understanding of metadiscursive concepts by non-experts. This initial annotation effort focus on five categories of metadiscourse: INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING, and EMPHASIZING. The crowdsourcing task is described in detail, including instructions and quality insurance mechanisms. We report results in terms of time-on-task, self-reported confidence, requests for additional context, quantity of occurrences and inter-annotator agreement. Results show the crowd is capable of annotating metadiscourse and give insights on the complexity of the different concepts in the taxonomy.

Keywords: Metadiscourse, Crowdsourcing, Non-experts

1. Introduction

Metadiscourse is one of the basic functions of language. Commonly referred to as *discourse about discourse*, it is composed of rhetorical acts and patterns used to make the discourse structure explicit, acting as a way to guide the audience. Crismore et al. (1993) define metadiscourse as "linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organize, interpret and evaluate the information given". Some examples of metadiscursive acts include introductions ("I'm going to talk about…"; "In this paper we present…"), conclusions ("In sum,…"), or emphasis ("The take home message…"; "Please note that…").

This study focuses on the function of metadiscourse in spoken communication. The functional analysis of such phenomena in discourse can contribute to tasks such as simplification or language understanding, and can be used for language learning purposes, such as presentation skill instruction. We describe the task of building a corpus of metadiscursive acts using crowdsourcing to annotate transcripts of presentations. By using non-experts, we expect not only to obtain the annotations of some metadiscursive acts, but also to get feedback on how those acts are perceived.

In this paper we start with background on metadiscursive theory, addressing how existing taxonomies and resources represent it (Section 2). Section 3 focuses on the choice of the material that the crowd will annotate with metadiscursive acts. Section 4 describes a preliminary annotation task aimed at testing the presence of some of the acts taken from our adopted metadiscourse taxonomy. Section 5 focuses on the setup of the crowdsourcing task, considerations regarding instructions, and quality control. The results obtained using the crowd and an ensuing discussion are presented in Sections 6 and 7. In Section 8, we conclude and present future directions.

2. Background

In the literature on discourse analysis we find studies that address function in discourse. For example, the contribution of Miltsakaki et al. (2008) to the Penn Discourse Treebank (PDTB) (Marcus et al., 1993) organized discourse connectives according to their function, considering categories such as giving examples (INSTANTIATION), making reformulations and clarifications (RESTATEMENT), comparing (CONTRAST), or showing cause (REASON). Another example is the RST Discourse Treebank (Marcu, 2000), a semantics-free theoretical framework of discourse relations based on Rhetorical Structure Theory (Mann and Thompson, 1988), which includes categories such as EXAMPLE, DEFINITION, or SUMMARY. Even though these projects explore function in discourse, they focus on written language and do not address the meta aspect of language.

The lack of work on the explicit nature of discourse motivated our decision to build a corpus targeting the function of metadiscourse in spoken communication. To accomplish that, we looked for definitions of metadiscourse.

Luuka (1992) developed a taxonomy for use in both written and spoken academic discourse. This taxonomy is composed of three main categories: TEXTUAL (strategies related to the structuring of discourse), INTERPERSONAL (related to the interaction with the different stakeholders involved in the communication) and CONTEXTUAL (covering references of audiovisual materials). Mauranen (2001), on the other hand, focused only on spoken discourse. This author's taxonomy is also composed of three categories with no further division: MONOLOGIC (similar to TEXTUAL in Lukka's taxonomy), DIALOGIC (similar to INTERPERSONAL in Lukka's taxonomy) and INTERACTIVE (related to question answering and other interactions with the speaker).

Luuka's and Mauranen's taxonomies organize metadiscourse in similar ways. However, both studies focus on

METALINGUISTIC COMMENTS

Repairing
Reformulating
Commenting on Linguistic Form/Meaning
Clarifying
Manage Terminology

DISCOURSE ORGANIZATION

Managing Topic

Introducing Topic
Delimiting Topic
Adding to Topic
Concluding Topic
Marking Asides
Enumerating

Managing Phorics

Endophoric Marking
Previewing
Reviewing
Contextualizing

SPEECH ACT LABELS

Arguing
Exemplifying
Other

REFERENCES TO THE AUDIENCE

Managing Comprehension
Managing Discipline
Anticipating Response
Managing the Message
Imagining Scenarios

Figure 1: Ädel's taxonomy of metadiscourse.

the *form* of metadiscourse (i.e. number of stakeholders involved), not addressing its *function*.

A *functional* approach to metadiscourse can be found in the work of Ädel (2010) who unifies existing taxonomies under a framework that encompasses both spoken and written discourse. This framework was built using two academic-related corpora: MICUSP (Römer and Swales, 2009) – comprised of academic papers – and MICASE (Simpson et al., 2002) – a corpus of university lectures.

The categories and organization of Ädel's taxonomy of metadiscourse (Figure 1) reflect the author's concern about the unification of theories for both written and spoken discourse and the desire to describe metadiscourse in a functional manner. For these reasons, we have decided to adopt this taxonomy as a source of categories of metadiscourse. This taxonomy will be discussed further in Section 4.

3. Corpora

Having adopted a set of metadiscursive acts to annotate, we then needed to select a source of data where these strategies could be found. Two main sources of data were considered: classroom recordings and TED talks[1].

Analysis of the contents of these two sources led us to choose TED talks over classroom recordings. TED talks are consistently good quality presentations from good presenters. Each talk is carefully rehearsed beforehand, conveying one message in a short span of time (from 5 to 20 minutes). This contrasts with classroom recordings which are typically longer and where there is an order in which the classes should be listened too. Even if only self-contained classes are considered, they are targeted at a very specific audience and the topics are advanced and require a significant amount of previous knowledge. Secondly, TED talks are uniform in content. They contain high-quality audio and video material and are available in several languages. They are also updated daily and subtitled, providing a good source of transcribed material. Classroom recordings, on the other hand, are a more heterogeneous resource as far as source and recording conditions are concerned, making them harder to be automatically processed with the least amount of human intervention possible. Even though they are not further addressed in this paper, classroom recordings would be a good resource set to extend our TED findings at a later time.

At the time of the preparation of this annotation task there were 730 TED talks available in English with subtitles, synced at sentence level (a total of 180 hours, approximately).

4. Preliminary Annotation Task

A small preliminary annotation task was carried out to test the suitability of the combination of Ädel's taxonomy and the TED talks. The goal of this annotation was to find which metadiscursive categories are present in the TED talks. Ten TED talks were annotated with the tags from the chosen taxonomy (see Figure 1). The ten talks were randomly chosen, spanning a variety of topics and years. This annotation task was performed by the first author.

The following paragraphs, each named after the 4 main categories of the taxonomy, present the taxonomy itself and, at the same time, describe how each type of metadiscourse is distributed over the sample.

Metalinguistic Comments are composed of 5 metadiscursive acts: REPAIRING, REFORMULATING, COMMENTING ON LINGUISTIC FORM/MEANING, CLARIFYING and MANAGING TERMINOLOGY. Most of these categories are exclusive to spoken discourse. From this set, only CLARIFYING and MANAGING TERMINOLOGY (defining of concepts) were found consistently in the sample. We believe that the fact that the other tags were not found is due to the high degree of preparation of each talk (when compared to academic lectures).

[1] http://www.ted.com/

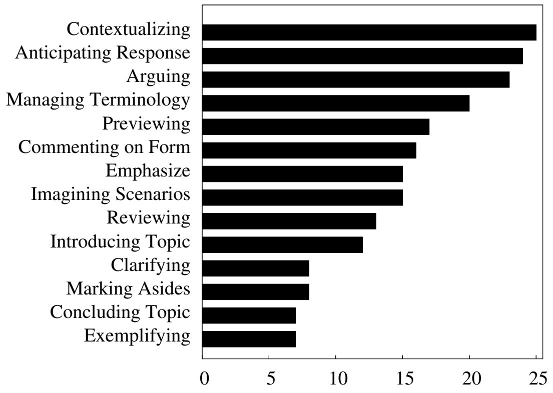


Figure 2: Occurrences of the most frequent tags.

Discourse Organization is divided in two other categories: *Manage Topic* and *Manage Phorics*. In *Manage Topic*, there are 5 metadiscursive acts: INTRODUCING TOPIC, DELIMITING TOPIC, ADDING TO TOPIC, CONCLUDING TOPIC and MARKING ASIDES. These structures were found consistently throughout the sample. The audience comes from a broad set of areas, and the speakers must wisely structure their discourse to convey their message. Additionally, the short time frame that is allotted for each talk demands an efficient use of language. The exceptions in this group were the tags DELIMITING TOPIC and ADDING TO TOPIC. The reason behind this may be the fact that TED talks have well-defined topics. The speakers tend to focus on what they want to talk about, going straight to the relevant points. *Manage Phorics*, the other subcategory under *Discourse Organization*, has four tags. PREVIEWING, REVIEWING and CONTEXTUALIZING are related to pointing to other locations in the current discourse. ENDOPHORIC MARKING contains references to physical elements (such as an image in the presentation), and was not considered in this preliminary task since it involved the integration of elements outside the discourse. The first three categories were well-represented in our sample.

Speech Acts contains 3 metadiscursive acts: ARGUING, EXEMPLIFYING, and OTHER (where the author included acts that were not frequent enough to generate a new tag). The first two tags were found frequently in the sample, and the category *other* was ignored since it did not represent a single concept.

References to the Audience is related to contact with the audience. Unlike in academic lectures, in TED talks the speaker typically does not interact with the audience. The message has to be conveyed without direct interaction, such as questions and checks for understanding. For these reasons, the tags MANAGE COMPREHENSION (check if the audience is in synch with the content of the presentation) and MANAGE DISCIPLINE (adjusting the channel asking for less noise, for example), were not found and therefore were not considered. The remaining 3 tags in this category (ANTICIPATING RESPONSE, MANAGING THE MESSAGE and IMAGINING SCENARIOS), on the other hand, were found frequently throughout the sample.

Figure 2 shows the distribution of the most frequent tags found in the ten talk sample. From the resulting fourteen categories, a small subset was chosen for the initial annotation effort in which we tested the suitability of using non-experts to identify occurrences of metadiscourse. Three criteria dictated the set of tags used in this annotation task. We considered (a) the most frequent concepts in the literature on presentation skills, (b) the concepts that could be best explained to non-experts, and (c) the input from Carnegie Mellon's International Communications Center (entity that holds presentation skills workshops and is responsible for administering tests for non-native speakers applying for teaching assistant positions). The resulting set of five tags are: INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING and MANAGING THE MESSAGE. Additionally, under the category EXEMPLIFYING we decided to collapse both EXEMPLIFYING and IMAGINING SCENARIOS (since they both consist of illustrating an idea). For simplification, MANAGING THE MESSAGE (in Ädel's work, "typically used to emphasize the core message in what is being conveyed") will be referred to as EMPHASIZING.

5. Crowdsourcing

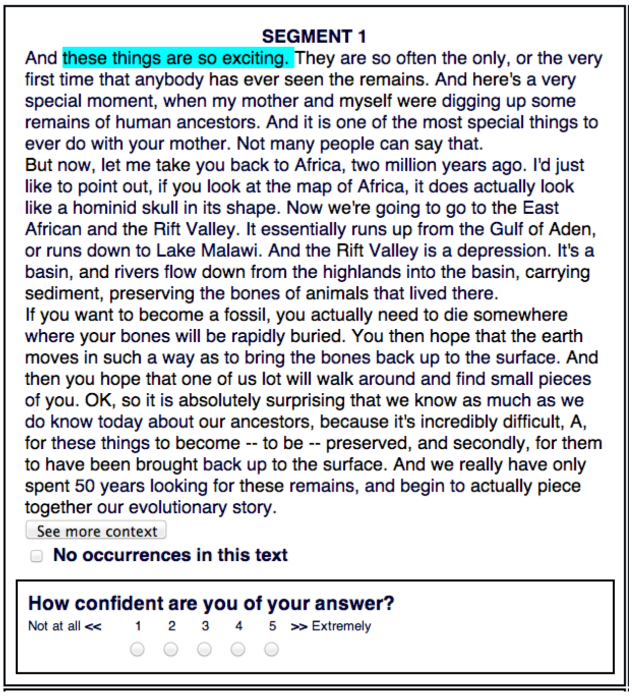
It has been shown that the quality of the crowdsourcing results can approach that of an expert labeler, while requiring less monetary- and time-related resources (Nowak and Rüger, 2010; Zaidan and Callison-Burch, 2011; Eskenazi et al., 2013). However, this advantage comes at a cost. Unlike experts, using the crowd requires setting up training and quality assurance mechanisms to eliminate noise in the answers. Additionally, it is necessary to approach problems in a different way, such as dividing complex jobs in subtasks to reduce cognitive load (Le et al., 2010; Eskenazi et al., 2013).

In our case, the reasons behind using crowdsourcing go beyond time and money. It allows the assessment of the crowd understanding. By designing a task requiring the annotation of metadiscourse, we are building a corpus of the phenomenon and understanding how non-experts comprehend metadiscursive concepts.

In the remainder of this section, we will describe the setup of a crowdsourcing annotation task (run on Amazon Mechanical Turk[2]).

The first decision concerned the amount of text that workers would annotate in each HIT (Human Intelligence Task – the smallest unit of work someone has to complete in order to be paid). Each HIT had to be simple and to allow workers to do it in the fastest way possible. However, metadiscursive phenomena are not local, requiring understanding the context. With this in mind, we decided to use segments of approximately 300 words. This limit was influenced by the design of the interface of the annotation task, taking into consideration that all the text should be visible on the screen without having to scroll down (scrolling increases time-on-task, influencing the answer rate). To make it monetarily worthwhile for a worker to chose our task, we included four segments per HIT, shown in a 2 by 2 matrix. Figure 3 shows

[2] https://www.mturk.com/mturk/welcome



SEGMENT 1

And these things are so exciting. They are so often the only, or the very first time that anybody has ever seen the remains. And here's a very special moment, when my mother and myself were digging up some remains of human ancestors. And it is one of the most special things to ever do with your mother. Not many people can say that.
But now, let me take you back to Africa, two million years ago. I'd just like to point out, if you look at the map of Africa, it does actually look like a hominid skull in its shape. Now we're going to go to the East African and the Rift Valley. It essentially runs up from the Gulf of Aden, or runs down to Lake Malawi. And the Rift Valley is a depression. It's a basin, and rivers flow down from the highlands into the basin, carrying sediment, preserving the bones of animals that lived there.
If you want to become a fossil, you actually need to die somewhere where your bones will be rapidly buried. You then hope that the earth moves in such a way as to bring the bones back up to the surface. And then you hope that one of us lot will walk around and find small pieces of you. OK, so it is absolutely surprising that we know as much as we do know today about our ancestors, because it's incredibly difficult, A, for these things to become -- to be -- preserved, and secondly, for them to have been brought back up to the surface. And we really have only spent 50 years looking for these remains, and begin to actually piece together our evolutionary story.

See more context

□ **No occurrences in this text**

How confident are you of your answer?

Not at all << 1 2 3 4 5 >> Extremely

○ ○ ○ ○ ○

Figure 3: One of the four segments in a HIT

the interface for one of the segments in a HIT. This configuration generated 2,461 HITs (or 9,844 segments) per category. It is also important to notice the presence of the button *See more context* in Figure 3. This feature allowed the workers to see the surrounding text of the segment in the talk (before and after), in case they needed additional context to support their decision.

The second consideration concerns the design of the instructions. Knowing that a metadiscursive act is a complex notion which workers may have never heard of, we decided that each HIT would target only one category, instead of requiring the identification of all five categories in each segment in one single passage. This decision lessens the cognitive load for the workers at each point. The instructions, for the *emphasis* task read as follows:

> *When making a presentation, to guide the audience, we often use strategies that make the structure of our talk explicit. Some strategies are used to announce the topic of the talk ("I'm going to talk about..."; "The topic today will be..."), to conclude a topic or the talk ("In sum,..."; "To conclude,..."), to emphasize ("The take home message is ..."; "Please note that..."), etc. We believe that by explaining and explicitly teaching each of these strategies we can help students improve their presentation skills.*
>
> *In this task, we ask you to focus on the strategies that the speaker uses to EMPHASIZE A POINT. Your job is to identify the words that the speaker uses to give special importance to a given point, to make it stand out, such as "more important", "especially", or "I want to stress that...". The passages you mark will be used on a presentation skills virtual tutor, showing students how professional speakers EMPHASIZE a point.*

Since the idea is to do one pass over all the segments for each one of the tags, we designed different sets of instructions for each one of the five metadiscourse categories. It is important to notice that the first paragraph of the instructions above was only included after some preliminary trials. Its inclusion was intended to reveal a concrete example of the applicability of our work, as a way of motivating workers. The inclusion of this paragraph increased the response rate. After the instructions there is a section with examples and counterexamples derived from the preliminary annotation task. Finally, at the bottom of the page, before the presentation segments, there is a succinct set of steps that explain the interface and how to use it to annotate the passages:

> **STEP 1:** For each of the extracts below, click on EVERY word that the speaker uses to EMPHASIZE A POINT. There may be zero, one ore more instances in each extract.
>
> **STEP 2:** The words you click on will display a light blue background. If you change your mind, you can click on the word again to deselect it.
>
> **STEP 3:** If you need more information to support your decision, you can click "*See more context*" below the segment to see the its surrounding context in the talk.
>
> **STEP 4:** If the speaker does not emphasize any point in the extract, select the "*No occurrences in this text*" checkbox below the text.
>
> **STEP 5:** Click the SUBMIT button once you are finished.

The last set of considerations had to do with quality control. We took advantage of the AMT prerequisites feature to filter out workers who were not native-speakers of English and find those who had a high reliability rate ($\geq$95%). Workers who satisfied the prerequisites and accepted the HIT were then guided through a four-segment training session. The training tested if the worker read the instructions and examples carefully, and was capable of performing this task. Only upon successful completion of the four training segments were the workers allowed to access real HITs in the category they were just certified on.

This training strategy is effective in filtering out *bots*, however it does not prevent malicious workers from giving random answers to the real HITs. For that reason, and in line with what is done in much of the crowdsourcing community, we defined a gold standard for each of the five metadiscursive tags. In every four HITs, at least one segment was compared to an expert annotation. The gold standard segments were very similar to the examples provided, and failing one of them raised a flag for the worker. This information was then checked before accepting or rejecting that annotator's work. Workers also noted their confidence level for each segment on a 5-point Likert scale (see Figure 3).

Category	time (m)	Confidence	Context Requests (%)
ASD	10	3.60	5.52
INT	3.7	3.95	1.32
CONC	3.5	4.00	37.09
EXMPL	6.2	3.94	4.81
EMPH	6.3	3.99	1.14

Table 1: Results in terms of time-on-task, self-reported confidence score and percentage of context expansion requests for MARKING ASIDES (ASD) INTRODUCING TOPIC (INT), CONCLUDING TOPIC (CONC), EXEMPLIFYING (EXMPL) and EMPHASIZING (EMPH).

A final mechanism to assure quality consisted of submitting the same HIT to 3 different workers, using a majority vote scheme.

Prior to publishing all the HITs in each category, we uploaded a small sample of 100 HITs to test the suitability of the instructions and interface. This trial phase allowed us to modify the instructions and examples for each category if necessary, and to test if the workers were able to understand and identify the metadiscourse act.

6. Results

This section presents the results of the annotation for each of the five metadiscursive acts. In Table 1 we report the results in terms of average time-on-task in minutes; self-reported confidence score on a 5-point Likert scale; and percentage of segments in which workers expanded context (by clicking on the *See more context* button). Table 2 indicates the number of occurrences of the metadiscourse tag; and inter-annotator agreement (κ). We used the Fleiss' kappa (Fleiss, 1971) as a measure of annotator agreement. Complete agreement corresponds to $\kappa = 1$, and no agreement (other than chance) corresponds to $\kappa \leq 0$. Herein, annotators agree if the intersection of the words selected by each of them is not empty. For example, two workers agree when one selects "Today, I would like to say that" and the other misses some of the words, selecting "I would like to say".

It is important to notice the absence of the tag MARKING ASIDES in Table 2. All the categories with the exception of MARKING ASIDES produced satisfying results in the trial sample of 100 HITs uploaded prior to submitting the entire set of talks. This fact lead us to discard the asides-related category. This will be discussed in detail in Section 6.1.

6.1. Marking Asides

As mentioned, the annotation of MARKING ASIDES was discontinued due to the inconclusive results obtained during the AMT trial phase. The first indicator of unsuccessful annotations was the slow response rate. The 100 HITs were up for one week during which less than 50% were completed. In the remaining categories, the sample was fully completed in less than two days. This slow response rate could be due to the small amount of HITs that were uploaded (workers tend to focus on tasks that have a significant amount of HITs online, in order to minimize training time and maximize payment). However, the four other categories were also first presented with 100 HITs and completed much faster.

Category	# occurrences	κ
INT	1,159	0.64
CONC	628	0.60
EXMPL	1,327	0.72
EMPH	2,580	0.58

Table 2: Number of occurrences and inter-annotator agreement (Fleiss' kappa) for the completed categories: INTRODUCING TOPIC (INT), CONCLUDING TOPIC (CONC), EXEMPLIFYING (EXMPL) and (EMPH).

We looked for other indicators and decided that the crowd could give us some insight on the understanding of the concept MARKING ASIDES. Workers were spending 10 minutes on average for each HIT, contrasting with the 4 to 6 minutes the other tasks took. Self-reported confidence scores were also the lowest of the five categories: 3.60, as opposed to 4.00 for the category CONCLUDING TOPIC. Finally, the workers wrote comments that clearly showed the task was hard, justifying the slow response rate and lack of confidence. Workers wrote: *"I am nervous that I am not doing these correctly *at all*"*; *"I hope that this is what you are looking for"*; and *"a little difficult"*.

6.2. Introducing a Topic

The task of annotating introductions resulted in an inter-annotator agreement of 0.64. Workers took on average 3.7 minutes to complete each HIT and identified over 1,000 instances of INTRODUCING TOPIC in our set of talks. It is important to note that speakers sometimes introduce several topics throughout a single talk, and therefore there can be more occurrences of INTRODUCING TOPIC than the total number of talks in the set (in this case 730). A final interesting point was the low number of times that workers asked for more context: only in 1.32% of the segments.

6.3. Concluding a Topic

The annotation of conclusions provided results that resembled the previous category: a slightly lower inter-annotator agreement ($\kappa = 0.60$), and similar average time-on-task and self-reported confidence. An important difference comes from the percentage of segments for which annotators asked to see the surrounding context: 37% of the segments. This might be an indication that conclusions are less local, needing a wider context to be identified. It is also important to notice that the number of occurrences of conclusions (628) is lower than the number of talks. This aligns with what we encountered in the preliminary annotation task (7 conclusions over 10 talks) and is related to the fact that the speakers do not always explicitly conclude (particularly true for shorter talks).

6.4. Exemplifying

In this category, workers spent on average two more minutes per HIT than while annotating instances of INTRODUCING TOPIC and CONCLUDING TOPIC. This results from the greater quantity of occurrences detected (1,327).

The more occurrences a category has, the more time workers will spend clicking on them. As previously described, this category collapses two metadiscursive acts as defined in Ädel's taxonomy: EXEMPLIFYING and IMAGINING SCENARIOS. Despite the collapse of tags, in this category annotators reached the highest agreement ($\kappa = 0.72$), which corroborates our decision to combine the two tags.

6.5. Emphasizing

While annotating occurrences of EMPHASIZING, the relationship between average time-on-task and number of instances was similar to the one found for the previous category. Workers spent on average 6.3 minutes per HIT and identified over 2,500 occurrences. While identifying emphasis, workers asked for the lowest amount of additional context amongst the five categories (1.14). EMPHASIZING was also the category where workers achieved the lowest inter-annotator agreement (0.58). This result may be due to the fact that this category is the only one in which there is a scale of intensity related to the concept, i.e., different workers might have different thresholds for considering that the speaker is emphasizing.

7. Discussion

The results obtained in this annotation task show that, once trained, non-experts can understand concepts of metadiscourse and identify them on TED presentations. However, this is not true for all of the categories we proposed to annotate. The category MARKING ASIDES was discarded during the trial phase on AMT since workers manifested signs of not understanding the task.

After the experiment took place, we looked into the instructions for this category to understand why workers were not able to annotate it. One of the counterexamples stressed the difference between MARKING ASIDES (where the speaker digresses to a topic sidetrack, such as in "Just a little side note here…") and ADDING TO TOPIC (where the speaker explicitly adds to the current topic, such as in "Let me add that…"). This distinction may have added to the worker's cognitive load. They were not only asked to be aware of another category in the taxonomy, but also required to focus on a subtle difference. The solution to this problem may be the division of the category in two. This can be done with a first pass collapsing both concepts under a more general notion, such as *adding information*, and a second pass where workers now only see instances that were detected in the first pass and decide if the addition of information is on or off-topic.

Another interesting result from this experiment is the need for additional context in different metadiscursive acts. Workers were able to identify occurrences of INTRODUCING TOPIC and EMPHASIZING in a window of 300 words without requesting for additional context. On the other hand, identifying conclusions was the task where more context was needed. The fact that workers expanded context in 37% of the segments might result from the necessity to first understand which topic is being presented, before deciding on the occurrence of its conclusion.

8. Conclusion and Future Work

In this paper, we have described an annotation task that took place on Amazon Mechanical Turk, where workers focused on a predefined set of metadiscourse categories to annotate text extracted from TED talks. We started from a set of 730 presentations and a taxonomy of metadiscourse and described the considerations for setting up a crowdsourcing annotation task aimed at finding metadiscursive concepts in the talks. The task was successful for four of the five categories that were submitted.

In future work, we plan to continue this annotation effort, extending it to the remaining categories of Ädel's taxonomy, and refining unsuccessful attempts (i.e. MARKING ASIDES) to meet the workers' cognitive load. We plan to extend this analysis to other languages, more precisely to European Portuguese, comparing the use of metadiscourse between the two languages. Finally, we aim at using the resulting annotation as training data for an automatic metadiscourse classifier.

Acknowledgments

This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and by FCT project CMU-PT/HuMach/0053/2008.

9. References

Ädel, Annelie. (2010). Just to give you kind of a map of where we are going: A Taxonomy of Metadiscourse in Spoken and Written Academic English. *Nordic Journal of English Studies*, 9(2):69–97.

Crismore, Avon, Markkanen, Raija, and Steffensen, Margaret S. (1993). Metadiscourse in persuasive writing. *Written communication*, 10(1):39.

Eskenazi, Maxine, Levow, Gina-Anne, Meng, Helen, and Parent, Gabriel. (2013). *Crowdsourcing for Speech Processing*. John Wiley & Sons.

Fleiss, Joseph L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Le, John, Edmonds, Andy, Hester, Vaughn, and Biewald, Lukas. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 21–26.

Luukka, Minna-Riitta. (1992). Metadiscourse in academic texts. In *Conference on Discourse and the Professions. Uppsala, Sweden*, volume 28.

Mann, William C and Thompson, Sandra A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, Daniel. (2000). *The theory and practice of discourse parsing and summarization*. The MIT press.

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

Mauranen, Anna. (2001). Reflexive academic talk: Observations from MICASE. In *Corpus linguistics in North*

America: Selections from the 1999 symposium, pages 165–178.

Miltsakaki, Eleni, Robaldo, Livio, Lee, Alan, and Joshi, Aravind. (2008). Sense annotation in the Penn Discourse Treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286. Springer.

Nowak, Stefanie and Rüger, Stefan. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 557–566. ACM.

Römer, Ute and Swales, John M. (2009). The Michigan Corpus of Upper-level Student Papers (MICUSP). *Journal of English for Academic Purposes*, April.

Simpson, Rita C., Briggs, Sarah L., Ovens, Janine, and Swales, John M. (2002). The Michigan Corpus of Academic Spoken English.

Zaidan, Omar F. and Callison-Burch, Chris. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1220–1229.