Interoperable Annotation in the Australian National Corpus

Steve Cassidy
Department of Computing
Macquarie University
Sydney, Australia
steve.cassidy@mq.edu.au

Abstract

The Australian National Corpus (AusNC) provides a technical infrastructure for collecting and publishing language resources representing Australian language use. As part of the project we have ingested a wide range of resource types into the system, bringing together the different meta-data and annotations into a single interoperable database. This paper describes the initial collections in AusNC and the procedures used to parse a variety of data types into a single unified annotation store.

1 Introduction

The Australian National Corpus (AusNC) is a new project to create a wide ranging resource for research on language in Australia. In contrast to other National Corpora, it is not a new, targeted collection of language data. Instead, the AusNC will manage a range of collections of language use in Australia that will be unified by common meta-data, data and annotation standards and formats. This approach allows us to curate existing important collections and incorporate new collections into a larger whole that may prove more useful than the sum of its parts.

In the long term, AusNC aims to illustrate Australian English in all its variety, situational, social, generational, and ethnic; and to document languages other than English used in Australia, including Aboriginal and Torres-Strait Islander languages, AUSLAN, and the community languages of immigrants. The Corpus also aims to serve a wide range of research disciplines from grammatical and lexical studies to sociolinguistic research and language technology. By including audio and video sources the Corpus hopes to be able to serve researchers interested in acoustics and gesture as well as language technology applications that require this kind of data to train and test computational models.

The pilot project that established the AusNC chose a small number of corpora that were felt to characterise the range of corpora in use by Australian researchers. These include a number of important historical collections that have been used to characterise Australian English in the past. The primary focus of the project was to ingest the corpus text and meta-data into a web accessible form and provide a way of browsing this data and publishing meta-data records to the Research Data Australia directory[1]. However, as a part of the ingestion process, we undertook to parse as much annotation data as possible and convert it to an RDF format (Cassidy, 2010) so that it might be used in a future version of the technical infrastructure.

This paper describes some aspects of the process by which meta-data and annotations were extracted from these corpora and the measures we took to ensure the interoperability of the data in the AusNC platform.

2 Overview of Corpora

The corpora included in the initial collection are drawn from a range of disciplines and contain a varied amount of meta-data and annotation. In summary, the corpora are:

[1] http://researchdata.ands.org.au/australian-national-corpus

- **The Australian Corpus of English (ACE)**: Written language, some simple XML like markup for header, bylines etc.
- **The Australian ICE Corpus**: Written and spoken language, XML like markup following the ICE standards.
- **The Corpus of Oz Early English (COOEE)**: Historical texts with minimal markup.
- **The Monash Corpus of Spoken English**: transcribed audio of conversations in Word format, speaker turn annotation
- **The Griffith Corpus of Australian Spoken English**: transcribed audio of conversations in PDF format with embedded Conversation Analysis markup.
- **The AustLit collection**: TEI formatted samples of Australian fiction.
- **The Mitchell and Delbridge Corpus**: audio recordings with time aligned word and phonetic annotations.
- **The Braided Channels Research Collection**: video recordings with transcriptions in Word format, speaker turn annotations, roughly time aligned with video.

All of these corpora are hand-annotated - the annotation was done as part of the data collection and served the research in a particular discipline. There is clearly scope for adding more machine-generated annotation such as sentence segmentation and POS tagging, but doing so was beyond the scope of the project. The work we report here is about understanding the existing annotation and ingesting it into an interoperable framework.

3 Some End User Goals

The goal of the AusNC is to bring together more collections of Australian language so that researchers can benefit from being able to work with many collections in a uniform way. To illustrate this we will look at two example 'use cases' from the point of view of a Linguistics researcher.

The first case involves a study of utterance final constructions and their effect on the following utterances. Researchers want to identify certain lexical items occurring at the end of a speaker turn (eg. 'is it?', 'can he?'), classify the turns according to the gender of the speaker and then study the turns and those that follow them to look for common patterns.

The second case looks at overlapping speech in dialogue. The researcher is interested in the lexical items that are used in backchannel interjections ('hmm', 'yeah', 'really') and so wants to generate a list of words that occur during overlapping speech ordered by frequency and distinguished by the gender of the speaker.

Each of these tasks can be achieved by researchers on the existing data sets; in fact they are things that have been done already. The main issue is that the variability in the way that meta-data and annotation is represented in the corpora mean that any study that wanted to work over multiple corpora would need to process each one separately with difficult and different manual methods. The three corpora that we'll target in these examples are the Griffith, Monash and ICE-AUS corpora, all of which contain transcriptions of dialogue with some overlap information and which have been identified by researchers as good resources that they would like to be able to make use of.

The two cases are similar in that they both involve identifying speaker turns in dialogue. These are represented differently in the source corpora, with Griffith and Monash using formatting within the Word or PDF document (a line starting with a speaker identifier and a colon) and ICE-AUS using XML like markup in the text. In Griffith and Monash, the end of a speaker turn is implicitly marked as the newline before the start of the next turn and so searching for words at the end of turns is problematic.

Speaker meta-data is available in all three corpora but in very different forms. In ICE-AUS it is in a separate spreadsheet; in Griffith and Monash it is at the head of each transcript in a table. Essentially, finding the gender of each speaker is a manual process of tabulating the available data, except for Monash which encodes gender in the speaker identifier.

The third kind of annotation we need to look at is overlap. This is handled very differently in each

case. Monash and ICE-AUS use explicit markup for regions of overlapped speech - in the case of Monash the text is enclosed in square brackets. Griffith's CA style of annotation uses an open square bracket to mark the start of overlap and vertical alignment to mark the relationship between the two speaker's utterances, but the end of overlap is not marked explicitly. ICE-AUS has an explicit mechanism for linking two overlapping segments but Monash relies on the reader to line up multiple segments. So if we have three speakers:

```
BH4M:    [whats that]
BH4MMo:  [what] did he do?
BH4MFa:  .. well we were going to
         the milkbar on Sunday
BH4MMo:  [oh]
BH4M:    [oh] here we go
```

we need to be very careful to keep track of the overlaps from the start of the discourse to be able to identify what overlaps with what.

A final consideration is document selection. Both the Monash and Griffith corpora represent a single kind of language use - conversation. However, the ICE-AUS corpus contains samples of conversation alongside monologues, newspaper text and fiction. Clearly in carrying out any study over multiple corpora, a researcher needs to be able to select appropriate documents based on their descriptive meta-data.

Based on this review, it is clear that if a researcher is to be able to perform queries on more than one data set, the main thing standing in their way is the diversity of representations of the phenomena that are annotated. In this case, the meaning of the annotations is aligned in each case (speaker turns, overlap) but their realisation is quite distinct. In addition, the link to meta-data about the speaker and the kind of language represented in each document needs to be clear.

4 Technical Architecture

The goal of the project is to establish a unified technical platform that can store the source media (text, audio, video), meta-data and annotations from these different corpora and provide not only online access to the resources but value-added services that make them more useful to the research community. The technical architecture builds on the DADA system (Cassidy, 2010) and integrates separate data stores for the source media, meta-data and annotation behind a web based presentation and analysis layer based on the Plone content management system.

The meta-data and annotation stores are built on an RDF triple store. The use of RDF for meta-data is well understood and our implementation makes use of standard vocabularies as far as possible to describe corpora and their contents. Modelling annotation data as RDF is less well established but our earlier work has shown that the data model and query language are well suited to the task. Among the challenges in this project are managing the scale of data resulting from ingesting annotations from a large number of corpora and dealing with the issues that arise in storing many different corpora in a single annotation store.

4.1 Parsing Annotation

All annotation in the corpus is stored as stand-off annotation, so the source media, be it text, audio or video, is stored separately in a web accessible location that will be referenced by the meta-data and annotation stores. For audio and video resources this is standard practice; for the text based corpora this has meant generating markup-free versions of the text to act as the source media.

To generate the markup-free based versions of the text we have developed a parsing library that is able to handle the variety of markup that we have found in our target corpora. The library, based on the Python `pyparsing`[2] module, is written such that new parsers can be built by chaining together primitive parser elements. The output of the parsing process is twofold – the plain text without markup and a stream of annotation objects that reference character offsets in the plain text stream. An example of calling a simple parsing procedure is shown in Figure 1.

The output from these parsing procedures is combined to produce the plain text version of the document and a collection of annotations that are then converted to RDF.

In the case of the ICE corpus, we drew on earlier work on a validating parser for ICE markup (Wong et al., 2011) which was able to convert the validated ICE markup to a standoff annotation format suitable

[2] `http://pyparsing.wikispaces.com/`

```
>>> markupParser('h', 'heading').parseString("<h>some stuff</h>")
([@(some stuff,[heading:  0 -> 10])], {})
```

Figure 1: An example call to one of the parser procedures, in this case parsing an XML style header from the ACE corpus. The result is a representation of the plain text and the annotation with character offsets.

```
RF3:  [Okay]
BH1M: [Im fifteen] years old.
RF3:  Fifteen?
BH1M: Yes.
RF3:  How do I spell your surname?
```

Figure 2: Sample of the original text from the Monash corpus

```
Okay
Im fifteen years old.
Fifteen?
Yes.
How do I spell your surname?
```

Figure 3: Sample of plain text from the Monash corpus corresponding to the raw text in Figure 2

for ingestion.

As described in earlier papers on the DADA system (Cassidy, 2010), annotations are modelled as RDF and stored on the server in a Sesame triple store. The annotation model used is now closely aligned with the proposed ISO Linguistic Annotation Framework (ISO 24612, 2012) and the intention is that this system is a realisation of that standard as an annotation database, rather than a data exchange format.

4.2 Parsing Speaker Turns and Overlaps

An example of the text version of a document from the Monash corpus is shown in Figure 2; this contains examples of both of the phenomena mentioned in Section 3: speaker turns and overlap. The parsing process removes all markup (in this case, the speaker identifiers and the square bracket overlap notation) and generates the text shown in Figure 3 and a collection of RDF annotations which will be discussed below.

A second part of the ingestion process is to read and normalise the meta-data that is associated with the primary data. This is found in different forms:

```
monash:speaker/BH1M a foaf:Person;
    monashp:role "primary";
    monashp:school "BH";
    foaf:age "15";
    foaf:gender "male" .
```

Figure 4: Part of the meta-data for the sample of Figure 2 describing the speaker BH1M.

spreadsheets, text files and in the case of the Monash and Griffith corpora, in tables at the start of each transcription file. This data is parsed as part of processing the document and normalised to standard vocabularies where possible. Items like speaker identifiers are treated specially to ensure we maintain the link between speaker data and annotations on speaker turns, and that speaker identifiers are unique across the different corpora. Figure 4 shows the description of one speaker which uses the standard `foaf` namespace[3] commonly used to describe individuals. Since the same property names are always used, we can filter speakers by gender or age (where available) irrespective of the corpus they contributed to.

A sample speaker turn annotation is shown in Figure 5 in the RDF format used by the DADA system. This is basically a set of descriptions of objects via attribute-value pairs. In this case, the object `monash:5514A` is an instance of the class `dada:Annotation` and has properties `dada:type` etc. The colon notation denotes namespaced identifiers which can be described by a formal vocabulary (ontology). The RDF descriptions of annotations can reference parts of the meta-data as seen in the `ausnc:speakerid` property in the example which references the speaker described in Figure 4.

The text in Figure 2 also contains an example of overlapping speech marked as square bracketed text. This is also recognised as part of the parsing process

[3] `http://www.foaf-project.org/`

```
monash:5514A a dada:Annotation;
    dada:type ausnc:speaker;
    dada:partof monash:10cdaedc;
    dada:targets monash:5514L;
    ausnc:speakerid monash:speaker/BH1M .

monash:5514L a dada:UTF8Region;
    dada:start 91;
    dada:end 113 .
```

Figure 5: Part of the RDF annotation generated from the raw text in Figure 2. The first part describes the annotation object itself which has a number of properties, this *targets* a locator object described in the second part as a region bounded by UTF8 character offsets. This represents the second line in Figure 2.

and annotations marking this region as overlap are generated. In this case it would be useful to also record the relationship between these two instances of overlap - that 'Okay' is spoken at the same time as 'Im Fifteen'; however, our parser is not yet capable of doing this for the Monash data. We have done this for another corpus, ICE-AUS as part of the work reported in (Wong et al., 2011) but in this case, instances of overlap were numbered to allow the correspondence to be made explicit. However, we found that since the annotators were unable to validate the markup they were writing (it was XML like but didn't conform to any formal system), there were many deviations from the stated rules that needed to be corrected before a useable parse could be completed. We suspect that this will be the case with the Monash data as well.

There are also examples of overlap in the Griffith corpus, marked up with the CA convention of an open square bracket, vertically aligned with the corresponding text from the second speaker. Here's an example:

```
11  H:  [family gen[der book two
12  S:  [can-      [can I borrow
13      that?
```

Given the involvement of vertical alignment and the lack of explicit end markers for the overlap, we've not yet been able to successfully parse this markup, however we are confident that we should be able to recover most of the information here with further work.

```
monash:5513A a dada:Annotation;
    dada:type ausnc:overlap;
    dada:partof monash:10cdaedc;
    dada:targets monash:5513L .
monash:5513L a dada:UTF8Region;
    dada:start 91;
    dada:end 102 .
```

Figure 6: Part of the RDF annotation generated from the raw text in Figure 2 showing an overlap annotation corresponding to the text 'Im Fifteen'

5 Discussion

5.1 Achieving User Goals

In Section 3 we presented two example tasks that users had identified as targets for the work we were doing in building the AusNC. These relied on having a more uniform annotation model that would allow queries over speaker turns and overlapping speech when the source corpora have quite different ways of expressing this markup.

We have described the ingest process for the AusNC which aims to build this uniform representation of annotation. An important part of this is the use of common labels for annotation types such that the same phenomena in different corpora can be identified in the same way. While the examples we chose were quite simple (and not particularly 'semantic'), they illustrate the concept of using standard types to describe kinds of annotation.

The solution that we have describe only goes part of the way towards solving the problems presented in Section 3 however. We've built a model but we need to build the query tools and analysis engines that can make use of the data to answer questions from researchers. We are currently involved in a follow-on project that aims to do just this, adding infrastructure for running tools that will support query and analysis of corpus data from the AusNC as well as generating new annotations by running automatic processes such as parser and POS taggers.

5.2 Annotation Types

Though the annotation data model is standardised across the different corpora, the types and contents of the annotations is different. The `dada:type`

property of each annotation denotes an *annotation type* while the `ausnc:val` property is used to carry a value or label for the annotation. Other feature values can be expressed as additional RDF properties on the annotation node.

The concept of annotation type is not directly expressed in the ISO-LAF standard but is realised in most examples as a non-distinguished property of each annotation or via the AnnotationSpace property. The main point being that there is no *requirement* in ISO-LAF for any kind of type system but that there are a couple of mechanisms by which one could be implemented which would be equivalent to the model used here.

The use of the type system allows us to assert that certain kinds of annotation are semantically equivalent - in this case the speaker turns and overlaps in different corpora. This is a key to the interoperability of annotations because without this we cannot reliably treat the annotations as having the same meaning. The use of RDF makes it natural to use a schema to describe the annotation types, meaning that we can generate schemas to describe different styles of annotation - from transcribed dialogue to Penn Treebank style parse trees.

In order to make any type system useful, the way that it is used needs to be standardised. The DADA vocabulary makes one suggestion that is compatible with the ISO-LAF framework; while there may be other options to consider, it would be an important next step to discuss how this should be realised within the standard.

5.3 Other Annotation Types in AusNC

As the ingest scripts were developed for the different corpora in AusNC, common type names were used for annotations where possible. However, since the focus of the project was on the ingestion of primary data and meta-data, there were only a small number of types that were identified as common over more than one corpus.

In all other cases, annotation type names, values and other properties were derived from the names used in the individual corpora or where appropriate in the documentation for the corpora. A good example is the Griffith corpus which uses Conversational Analysis markup embedded in the text. The documentation for this annotation style was taken from (Lerner, 2004) which contains a glossary of transcription symbols with an informal description of their use and meaning. Table 1 lists the types that we have parsed with some examples of their use (there are a few other types that are used in the corpus that we are still working on parsing correctly).

Type Name	**Example**
micropause	(.)
pause	(1.2)
elongation	fo:r commu:nicating
intonation	if ↑I couldnt bo↓rrow,
latched-utterance	7 H: sexuality=
	8 S: =ah
speaker	5 S: I'm glad I saw you
volume	business °cause° I missed
uncertain	S: (,) this morning,

Table 1: Annotation types and examples from the Griffith corpus

6 Summary

This paper has tried to summarise some of our experiences in taking source data in many different formats and generating a single, interoperable annotation store that can hold annotations on many resources from different collections. The current system is able to present these resources via the web[4] and we are now starting to develop tools to work with the annotated data to help answer research questions for the diverse communities who make use of this data.

References

Steve Cassidy. 2010. An RDF Realisation of LAF in the DADA Annotation Server. In *Proceedings of ISA-5*, Hong Kong, January.

ISO 24612. 2012. Language Resource Management – Linguistic Annotation Framework.

G.H. Lerner. 2004. *Conversation analysis: studies from the first generation*. Pragmatics & beyond. John Benjamins Pub.

Deanna Wong, Steve Cassidy, and Pam Peters. 2011. Updating the ice annotation system: tagging, parsing and validation. *Corpora*, 6(2):115–144.

[4] `http://ausnc.org.au/`

Conceptual and Representational Choices in Defining an ISO Standard for Semantic Role Annotation

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

Martha Palmer
Department of Linguistics
University of Colorado
Boulder, Co.
USA
martha.palmer@colorado.edu

Abstract

This paper presents two elements of the ISO standard for semantic role annotation which is under development (ISO CD 24617-4:2013), namely (a) the metamodel, which describes the types of concepts that may occur in semantic role annotation and their conceptual relations, and (b) an annotation language for expressing semantic role annotations, with its abstract syntax, XML-based concrete syntax, and semantics.

1 Introduction

ISO project 24617-4, Language resource management Semantic annotation framework Part 4: Semantic Roles, has the aim of defining an international standard for the annotation of semantic roles, including an inventory of core semantic roles defined as ISO data categories, and an annotation language with an XML-based representation format and a formal semantics.

Semantic roles are receiving increasing interest in the information processing community because they make explicit the key conceptual relations of participation between a verb and its arguments, i.e., they specify Who did what to whom, and when, where, why, and how. For English alone, there are already several different semantic role frameworks, including FrameNet, VerbNet, LIRICS, EngVallex and PropBank (see Fillmore & Baker, 2004; Kipper-Schuler, 2005; Schiffrin & Bunt, 2007; EngVallex, 2011; and Palmer et al., 2005, respectively). Although these have been developed independently, there are strong underlying compatibilities between these frameworks, and they share a central definition of what a semantic role is, and what its span is, within an individual sentence. In addition to defining key concepts, the ISO standard aims at clarifying and specifying these underlying compatibilities and providing where possible a mapping between similar semantic roles across different frameworks. This mapping illustrates how different semantic role definitions can be linked to each other across frameworks, and presupposes a specification of clearly defined criteria for distinguishing semantic roles.

The specification can be used in two different situations:

- in annotations where the semantic roles are recorded in annotated corpora;
- as a dynamic structure produced by automatic systems; a process typically called semantic role labelling (SRL)

The objectives of this specification are to provide:

- A reference set of data categories defining a structured collection of semantic roles with an explicit semantics.
- A pivot representation based on a framework for defining semantic roles that could facilitate mapping between different formalisms (alternative semantic role representations/syntactic theories/eventually different languages) promoting interoperability.
- Guidelines for creating new resources that would be immediately interoperable with pre-existing resources

The ISO semantic roles project follows a design strategy for semantic annotation projects that includes (a) the design of a conceptual model which contains the key concepts involved in the kind of semantic annotation and which describes how these concepts are related; such a model is called a 'metamodel' (see Bunt & Romary, 2004), and (b) the three-part definition of an annotation language, the parts being (1) an 'abstract syntax', specifying how the basic concepts defined by the metamodel may be combined into set-theoretic structures called annotation structures'; (2) a 'concrete syntax', defining a reference representation format, typically using XML, for representing the annotation structures defined by the abstract syntax, and (3) a formal semantics describing the meaning of annotation structures (see Bunt, 2010; 2013 for a description of this methodology, called the CASCADES methodology). This paper focuses primarily on the metamodel constructed in the project for semantic role annotation (section 2) and the definition of the annotation language (3). For a more detailed description of the frameworks discussed and of semantic roles in general see the ISO document ISO 24617-4:2013, Bonial et al. (2011) and Johnson et al. (2001). The paper concludes with a brief discussion of what has been achieved and what remains to be done.

2 A metamodel for semantic role annotation

2.1 Predicate-argument structures and eventualities

A predicative expression with its arguments can be viewed semantically as describing an actual or hypothetical eventuality with its participants. Associated with the predicate (most prototypically a verb) is a subcategorization frame, describing the participants that are expected in that particular type of eventuality. Each slot in the subcategorization frame can be given a semantic role label which can then be associated with any argument that fills that slot. In the most fine-grained view each individual lexical item can be seen as defining a unique eventuality type with a unique set of possible participants.

Different predicative expressions may share the same or a very similar set of possible participants. Obvious examples are nouns and adjectives that constitute derived forms of the same lexical item (*observe, observance, observer*). Other examples are *buy* and *sell*, and *give* and *receive*. Depending on the desired level of generalization, the grouping of lexical items into shared subcategorization frame classes may stop there (this is one view of the PropBank Frame Files) or may continue to include a small set of items with very closely related semantics (the FrameNet view) or may extend to include items that share specific patterns of argument types but may have a fairly tenuous semantic relation (the VerbNet view). These frameworks take the subcategorization frame as a whole into consideration when determining the choice of individual semantic roles; this is motivated by examples such as *replace*, which can have one participant as the old item being replaced and another participant as the new item replacing it, with an obvious dependency between these two roles.

LIRICS does not use subcategorization frames or any other a priori association of semantic roles, but uses a set of features, like intentionality of the involvement of a participant, to distinguish among individual semantic roles, in the spirit of Dowty (1991). For example, in (1a), the behaviour of 'Martin' is clearly intentional, and he would be assigned the Agent role. In (1b), there is no intentionality involved, and *The lightning* would be assigned the Cause role. Sentence (1c) is ambiguous as to whether Martin's behaviour caused the children to be frightened as an intended or as an unintended effect, and so the semantic role of *Martin's behaviour* is either Agent or Cause.

(1) a. Martin frightened the children by pulling faces at them.

b. The lightning frightened the children.

c. Martin's behaviour frightened the children.

Note that the same word can have multiple senses, each of which might be associated with a distinct event type, and therefore a distinct frame. In this case the word could be represented by several eventuality types, each one associated with a different frame or class. Therefore, for the approaches to semantic role labelling embodied in FrameNet, PropBank, EngVallex and VerbNet, there are three core

elements that must be defined for semantic role labelling:

1. the word sense, or lexical unit, under consideration;
2. the frame associated with that word sense; and
3. specific semantic role labels associated with each slot in that frame that will be assigned to the participants filling the slot.

The more examples that can be provided to illustrate the degree of syntactic variation available to each sense, the better. These examples, or instances, are considered tokens that are each associated with the appropriate type definition.

An additional consideration in defining any semantic role labelling scheme is exactly which constituents are labeled as adjuncts and whether or not a set of general adjunct types is defined. It is notoriously hard to draw a clear line between arguments of a verb and adjuncts, and approaches to semantic role labelling differ in how they draw such a line, or finesse the question by giving individual labels to adjuncts associated with each eventuality type. Finally, frames may include information about likely semantic types of the semantic roles being specified.

The frames associated with a semantic role labelling scheme specify the roles associated with the eventuality types. (For FrameNet they would be the FrameNet Frames, for PropBank and for EngVallex they are the PropBank role sets or framesets, and for VerbNet they are defined in VerbNet classes.) The frames are typically consulted during annotation to guide the decisions and ensure consistency. This makes the specification of the frame a critical step in the path towards an annotated corpus. For each predicate in a language, a meta-level description of the predicate and its arguments needs to be created, with examples, which constitutes the definition of the eventuality type frame.

2.2 Eventualities, participants, types and tokens

Figure 1 visualizes the conceptual view that underlies semantic role annotation according to standard ISO 24617-4 under development. A predicative expression in natural language, in the sense in which it is understood in a given utterance, is viewed as denoting a certain type of eventuality, and the occurrence of the verb form in the utterance as denoting an instance (or 'token') of that type of eventuality. Each eventuality type has a semantic role set or 'frame' defined, which determines the possible choices of individual semantic roles for the participants in an instance of that eventuality type. Eventuality types may further be grouped into classes that have similar role sets, possibly defining hierarchies of event classes/types and the corresponding role sets/frames (not shown in Fig. 1).

Like eventualities, participants also have a semantic type, typically expressed by the lexical item that serves as the nominal head of a noun phrase or that forms the central element in a predicative expression. The metamodel in Fig. 1 indicates that in a given utterance, the semantic roles relate the participants that occurrences of nominal (or adveribial) lexical items refer to, to the eventualities corresponding to an occurrence of a verb (or noun, or other event-denoting predicative expression). Participants and eventualities are both tokens of certain types, which pertain to a semantic type system.

Since annotations add linguistic information to stretches of primary data, the identification of relevant stretches in the data is essential. In standoff format, this realized through pointers to the primary data (the original text) or to elements at another layer of annotation, such as a syntactic parse, where the regions of primary data are identified. Following ISO practice, the term *'markable'* is used to refer to the entities that anchor an annotation directly or indirectly in the primary data. Note that the metamodel stipulates that participants and eventualities are expressed by markables in the original text ('source document'), but that semantic roles are not textually expressed.

3 SemRolesML

3.1 Abstract syntax

The abstract syntax of an annotation language consists of two parts (Bunt, 2010): (a) a specification of the elements from which annotation structures are built up, called a 'conceptual inventory', and (b) a specification of the possible ways of combining these elements in set-theoretical structures, called 'annotation structures'.

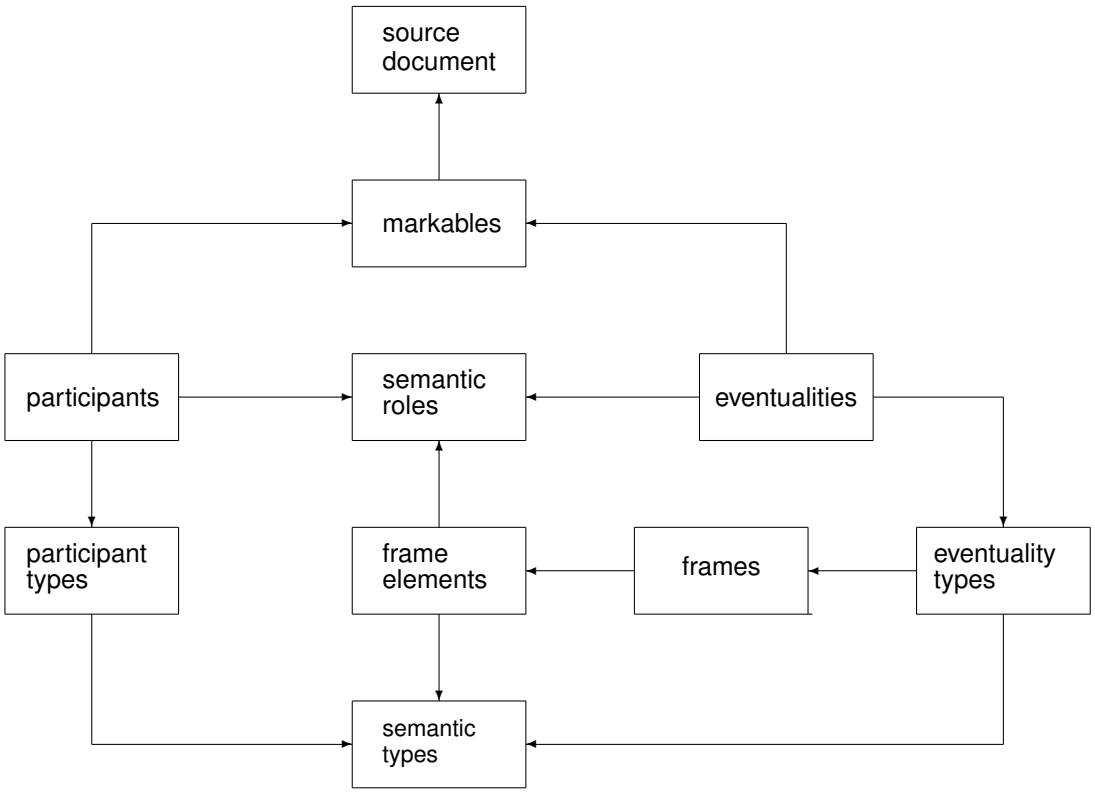

Figure 1: Metamodel for semantic role annotation.

a. Conceptual inventory

The conceptual inventory of the SemRoleML markup language, defined as part of ISO 24617-4, is derived from the metamodel shown in Fig. 1 by identifying among the categories of concepts in the metamodel those which are elementary and those which are composite, the latter being defined in terms of other concepts occurring in the metamodel. The listing of the basic concepts constitutes the conceptual inventory.

Of the ten categories represented in Fig. 1, the 'source document' is present only as a source of the markables and a carrier of possibly relevant metadata. Of the other nine categories, 'participants' and 'eventualities' are tokens of the basic concepts 'participant type' and 'eventuality type', respectively, and are identified by the occurrences of predicates and argument NPs in certain markables; as such they are instances (or 'tokens') of basic concepts, rather than basic concepts themselves. (Technically, they correspond to so-called 'entity structures' in the abstract syntax, see below.)

Concepts from the three categories at the bottom of Fig. 1, 'frames', 'frame elements' and 'semantic types', do not necessarily show up in semantic role annotations (but they often do in FrameNet annotations); they are especially important in the lexical resources supporting semantic role annotation. With respect to our abstract syntax, frames are a composite concept, that include n-tuples of frame elements. Frame elements include pairs of semantic role labels and specifications of the most likely semantic type of a participant playing that role, and are thus also composite concepts. So the five categories of elemenary concepts that form the SemRoleML conceptual inventory are: *markables, semantic roles, participant types, semantic types*, and *eventuality types*.

The specification of the SemRoleML conceptual inventory is thus the following listing of elementary concepts:

1. EV, a finite set of eventuality types, typically corresponding to verbs, nouns and adjectives.

2. RL, a finite set of semantic roles, such as the LIRICS role set (Schiffrin and Bunt, 2007; Petukhova and Bunt, 2007). This set can have a hierarchical organization, such as the unified VerbNet-LIRICS hierarchy presented by Bonial et al. (2011), with lower tiers expressing more fine-grained meanings, however this is not part of the conceptual inventory as such, but follows from the definitions of these roles (cf. Miltsakaki et al., 2008).

3. MA, a finite set of markables to which semantic roles can be attached.

4. PT, a finite set of participant types.

5. ST, a finite set of semantic types. The set PT of participant types and the set EV of eventuality types are subsets of ST.

b. Annotation Structures

An annotation structure is a set of entity structures and link structures. An entity structure is a pair $\langle m, s\rangle$ consisting of a markable (element of MA) and a specification of semantic information about that markable. For semantic role annotation, entity structures describe the eventualities and participants (both at token level) that are related by semantic roles. There are two kinds of entity structures in SemRoleML, those where the component s characterizes an eventuality and those where it characterizes a participant.

A link structure in SemRoleML is a triplet $\langle \epsilon_e, \epsilon_p, \rho\rangle$ consisting of two entity structures ϵ_e and ϵ_p, corresponding to an eventuality and a participant, respectively, and a semantic role specification ρ, which is either simply a semantic role label R or a pair $\langle \phi, R\rangle$, where ϕ is a frame, i.e. a list of frame elements $\phi = \langle \phi_1, \phi_2, \phi_k\rangle$. A frame element is either just a specification of a semantic role, or a pair $\langle R_i, t_i\rangle$ consisting of the specification of a semantic role and a semantic type (expected to subsume the participant type of a participant filling that role).

For the example sentence (2) two entity structures are created, one for the markable *The soprano*, and another one for the markable *sang*, shown in (3):

(2) The soprano sang

(3) a. $\epsilon_1 = \langle the\ soprano, \text{SOPRANO}\rangle$
b. $\epsilon_2 = \langle sang, \text{SING}\rangle$

For easy of readability, the strings *the soprano* and *sang* are used here to indicate markables (i.e. an occurrence of a stretch of text in the source document), SOPRANO is a participant type (an element of PT), and SING is an eventuality type (an element of EV).

A link structure is moreover created consisting of the two entity structures ϵ_1 and ϵ_2 and the semantic role *Agent*. The link structure is thus the triplet:

(4) $L_1 = \langle \epsilon_1, \epsilon_2, Agent\rangle$

The annotation structure for sentence (2) is the pair consisting of these entity structures and link structure(s):

(5) $\alpha = \langle \{\epsilon_1, \epsilon_2\}, \{L_1\}\rangle$

Note that ST, the set of semantic types, can be used to distinguish semantic roles and help determine their applicability. These are specified as selectional preferences by VerbNet, and are often included in the textual descriptions in FrameNet. As with the semantic roles, inheritance relations can hold between semantic types; these can be based on an hierarchical classification such as the hypernyms in WordNet (Miller, 1990; Feelbaum, 1998). In the example *The soprano sang*, the verb *sing* will plausibly have a frame which specifies that the frame element for the Agent slot expects a participant with the semantic type ANIMATE (or maybe HUMAN $\cup$ BIRD, if we agree that only humans and birds sing); since sopranos are humans, the semantic type system should include the knowledge SOPRANO $\subset$ HUMAN, and therefore the participant type is indeed subsumed by the semantic type.

The frames discussed above specify for each eventuality type the associated set of semantic roles, and can be used to guide the annotation process. Each frame consists of an eventuality type, e (an element of EV), and a subset, S_e, of RL with at least one element, such that $e \in EV$, and $r_i \in RL$ for all $r_i \in S_e$. For example, the frame for *sing* as occurring in example (2) above would consist of the eventuality type, SING, and the possible roles, including *Agent* and *Theme*, both of which are members of RL.

3.2 Semantics

The CASCADES design methodology (Bunt, 2013), used in the development of ISO 246171-4, derives a formal semantics for a given abstract syntax through a translation of the components of annotation structures to discourse representation structures (DRSs, Kamp and Reyle, 1994), which are combined by unification operations into a DRS for the annotation structure as a whole.

An entity structure $\langle m, s\rangle$ is interpreted as a DRS which introduces a discourse marker paired with a name of the markable m,[1] and which contains for each component s_i of s a condition of the form $p_i(x, a_i)$, where a_i is the interpretation of the component s_i, p_i is a predicate that indicates the role of a_i, and x is the newly introduced discourse marker. So the entity structures ϵ_1 and ϵ_2 are interpreted as the following DRSs, where m_1 names the markable *the soprano* and m_2 the markable *sang*:

(6) a. $\epsilon_1 \rightsquigarrow$

$\langle m_1, x_1\rangle$
PARTICIP_TYPE$(x_1, soprano)$

b. $\epsilon_2 \rightsquigarrow$

$\langle m_2, e_1\rangle$
EVENT_TYPE$(e_1, sing)$

A link structure $\langle\langle m, s\rangle, \langle m', s'\rangle, \rho\rangle$ is interpreted as a DRS which introduces discourse markers z_1 and z_2, paired with the markables m and m', respectively, and which has a condition of the form $R'(z_1, z_2)$, where R' is the DRS-predicate interpreting the relation ρ.

So the link structure L_1 of (4) is interpreted as the following DRS:

(7) $L_1 \rightsquigarrow$

$\langle m_1, z_1\rangle$, $\langle m_2, z_2\rangle$
AGENT(z_1, z_2)

Merging these interpretations of the entity and link structures results in the following interpretation of the annotation structure (5):

(8) $\alpha \rightsquigarrow$

$\langle m_1, x_1\rangle$,$\langle m_2, e_1\rangle$
PARTICIP_TYPE$(x_1, soprano)$
EVENT_TYPE$(e_1, sing)$
AGENT(e_1, x_1)

Once the DRS-interpretations of the entity structures and link structure have been combined (see footnote 1), the markable names can be deleted, resulting in a DRS of the usual kind.

A classical DRS is semantically equivalent to a formula in first-order logic; in this case the equivalent formula is (9), which says that there exist an eventuality, an eventuality type, a participant, and a participant type, such that the eventuality is a token of the eventuality type, the participant is a token of that participant type, and the participant is the agent of the event.

(9) $\exists e_1.\exists et_1.\exists p_1.\exists pt_1.$ EVENT-TYPE$(e_1, et_1) \wedge$ PART-TYPE$(p_1, pt_1) \wedge$ AGENT(e_1, p_1)

In this semantic representation, AGENT is a first-order predicate constant that expresses the meaning of the semantic role Agent. The hardest part of the semantics of SemRoleML is in fact the formal definition of the logical predicates that express the meanings of the individual semantic roles. Defining these predicates comes down to formalizing the semantic role definitions in ISO CD 24617-4: 2013, Annex A. Figure 1 shows three examples of these definition. The Agent role, for example, is defined as one where a participant initiates and carries out an event intentionally or consciously, and who exists independently of the event. The condition of acting 'intentionally or consciously' distinguishes the Agent role from the Cause role; the existence independently of the event forms one of the distinctions between the Agent and Cause roles on the one hand and the Result role on the other hand (and, more significantly, also distinguishes the Result role from the Theme and Patient roles).

The formalization of such definitions can be used to complete the semantics of semantic role annotations; for example, the interpretation (9) of the

[1] The paring of discourse markers with markable names serves to ensure that, when an annotated text is interpreted which contains more than one occurrence of the same stretch of text, the right occurrences are combined in the semantics. See Bunt (2012) for details.

SemRoleML annotation of the sentence *The soprano sang* can be completed by replacing the predicate AGENT by (10a). Similarly, the semantics of CAUSE can be described by (10b).

(10) a. AGENT = $\lambda e.\lambda x.$ [Intent-Init(x,e) $\vee$ Consc-Init(x,e)] $\wedge$ [Intent-Do(x,e) $\vee$ Consc-Do(x,e)] $\wedge$ Indep-Exist(x,e)

b. CAUSE = $\lambda e.\lambda x.$ Init(e) $\wedge$ $\neg$Intent-Init(x,e) $\wedge$ $\neg$Consc-Init(x,e) $\wedge$ $\neg$Intent-Do(x,e) $\wedge$ Indep-Exist(x,e)

For some frameworks this approach to the semantics of semantic roles could be almost prohibitively burdensome. FrameNet has thousands of frame elements, and while VerbNet has less than 30, the definitions of each one can change subtly from class to class. On the other hand, this is perhaps the only way to semantically make sense of these elements with a formal rigour, required for automatic inferencing.

3.3 Concrete syntax

Following the CASCADES design methodology,a reference representation format for annotation structures, based on XML, can be defined as follows, given an abstract syntax specification.

1. For each element of the conceptual vocabulary define an XML name;
2. For each type of entity structure $\langle m, s\rangle$ define an XML element with the following attributes and values:
   (a) the special attribute `@xml:id`, whose value is an identifier of the entity structure representation;
   (b) the special attribute `@target`, whose value represents the markable m;
   (c) attributes whose values represent the components of s, and which themselves represent the significance of the components;
   (d) if s_i is an elementary concept then it is represented by its name.
3. For each type of link structure $\langle \epsilon_1, \epsilon_2, \rho\rangle$ define an XML element with three attributes, two which have values that refer to the representations of the entity structures ϵ_1 and ϵ_2, the value of the third denoting the semantic relation between them.
4. For each type of auxiliary structure (see below) specify an XML representation.

Applied to the abstract syntax of SemRoleML, this results in the following concrete syntax:

1. The XML elements `<event>` and `<participant>` are defined for representing entity structures corresponding to eventualities and participants, respectively. Both of these elements have the attributes `@xml:id` and `@target`, and additionally they have the attributes `@eventType` and `@participantType`, respectively.
2. XML constants are chosen for the values of the attributes `@eventType` and `@participantType`.
3. The XML element `<srLink>` is defined for representing semantic role link structures; this element has the attributes `@event` and `@participant` whose values refer to the eventuality and the participant that are related by a semantic role, and the attribute `@semRole` whose value represents the semantic role of the participant in the eventuality.
4. For completeness, we mention that it is convenient to introduce auxiliary structures in the abstract syntax for frames and frame elements, which may occur within the relational component ρ of a link structure $\langle \epsilon_e, \epsilon_p, \rho\rangle$; see ISO CD 24617-4 (2013) for more details.

For the example sentence *The soprano sang* this gives us the following representation of the annotation structure (5):

(11)
```
<event xml:id="e1"
 target="#m2"
 eventType="sing"/>
<participant xml:id="x1"
 target="#m1"
 participantType="soprano"/>
<srLink event="#e1"
 participant="#x1"
 semRole="agent"/>
```

/agent/	
Definition	Participant in an event who initiates and carries out the event intentionally or consciously, and who exists independently of the event.
– Source	Adapted from Dowty [1989], EAGLES, SIL, Sowa [2000] and UNL
Explanation	An agent may be animate, or only seemingly, or perceived, as animate; this is so that cases of nonhuman agency such as a robot, or an institution will not be excluded from being able to initiate an event, e.g. "GM offers rebates on its new models".
Example	"John [agent e1] built e1 the house"

/cause/	
Definition	Participant in an event that initiates the event, but that does not act with any intentionality or consciousness; the participant exists independently of the event.
– Source	Adapted from: SIL (Causer) and Sowa [2000] (Effector)
Explanation	Except for the lack of intentionality of the participant, this semantic role is very similar to that of the agent and in fact shares all its other properties. The role of cause can often be identified with verbs of initiation, or causation, such as: to cause, to produce, to start, to originate, to occasion, to generate.
Example	"The wind [cause e1] broke e1 the window" "His talk [cause e1] produced e1 a violent reaction e2 from the crow

/result/	
Definition	Participant in an event that comes into existence through the event. It indicates a terminal point for the event: when it is reached, then the event does not continue.
– Source	Adapted from Sowa [2000]
Explanation	Result is the completed point of a process, and unlike goal is dependent upon the event for its existence.
Example	"(Within the past two months [duration e1]) (a bomb [cause e1]) exploded e1 (in the offices of El Espectador in Bogota [location e1]), (destroying e2 (a major part of its installations and equipment [patient e2]) [result e1])"

Figure 2: Examples of LIRICS semantic role definitions in the form of ISO data categories (from Schiffrin & Bunt, 2007)

4 Conclusion

In this paper we have described a number of fundamental decisions in the process of defining an international ISO standard for the annotation of semantic roles. Starting from the conceptual view of predication in natural language as referring to (actual or hypothetical) eventualities and their participants, and of semantic roles as ways in which a participant may be involved in an eventuality, we outlined a metamodel which specifies the categories of basic concepts involved in semantic role annotation, and which shows how these concepts are interrelated. We subsequently defined an annotation language, SemRoleML, which has an XML-based pivot representation format for semantic role annotations, and a semantics that is defined for an abstract syntax that underlies these representations. We showed how the formalization of semantic role definitions can in principle be the basis of a semantics of semantic role annotations.

Two advantages of defining the semantic role annotation language SemRoleML in this way, following the CASCADES methodology of defining semantic annotations, are

(1) that different representation formats, used to encode the same underlying abstract structures,

share the same semantics, and are thus semantically interoperable;

(2) that integration of the annotation of semantic roles with the annotation of other types of semantic information, such as information about time and events according to ISO 24617-1, or about spatial information (ISO 24617-7, under development) or about discourse relations (ISO 24617-8, under developent) is facilitated, since these all follow the same design methodology;

(3) that annotations of other linguistic phenomena, especially when following the ISO Linguistic Annotation Framework (ISO 24613:2012), such as annotations of syntactic, pragmatic and contextual information, can be combined with semantic role annotations; many of these are helpful and sometimes even necessary to determine word senses and resolve references for the automatic recognition of semantic roles.

All this helps to make these annotation schemes mutually interoperable and combinable.

Important work that remains to be done is the formalization of all the semantic role definitions which are included in ISO CD 24617-4, including the specification of meaning postulates for the predicates used in their interpretation, in order to fully specify the inferences that may be drawn from the semantic roles used in an annotated corpus.

References

Bonial, Claire, William Corvey, Volha Petukhova, Martha Palmer, and Harry Bunt (2011) A Hierarchical Unification of LIRICS and VerbNet Thematic Roles, in *Proceedings ICSC Workshop on Semantic Annotation for Computational Linguistic Resources (SACL-ISCS 2011)*, September 21, 2011, Stanford, CA.

Bunt, Harry (2010) A Methodology for designing semantic annotation languages exploiting syntactic-semantic iso-morphisms. In: Alex Fang, Nancy Ide, and Jonathan Webber (eds.) *Proceedings ICGL 2010, the 2nd International Conference on Global Interoperability for Language Resources*, Hong Kong, pp. 29–45.

Bunt, Harry (2012) Annotations that effectively contribute to semantic interpretation. Forthcoming in Harry Bunt, Johan Bos and Stephen Pulman (eds) *Computing Meaning, Vol. 4*. Springer, Berlin.

Bunt, Harry (2013) A methodology for designing semantic annotations. Forthcoming in *Language Resources and Evaluation*.

Bunt, Harry and Laurent Romary (2004) Standardization in Multimodal Content Representation: Some Methodological Issues. In *Proceedings LREC 2004*, Lisbon, pp. 2219-2222.

Dowty. David (1991) Thematic Proto-Roles and Argument Selection. *Language*, 67:547-619.

EngVallex 2011 Charles University in Prague. Available at (Accessed 9/10/2012): http://ufal.mff.cuni.cz/lindat/EngVallex.html

Fellbaum, Christiane (1998) *WordNet: An Electronic Lexical Data-base. Language, Speech and Communications.* MIT Press, Cambridge, MA.

Fillmore, Charles; Collin Baker, and Hiroaki Sato (2004). FrameNet as a "Net". In *Proceedings 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1091-1094

ISO 24612:2012 Language Resource Management - Linguistic Annotation Framework. Spring 2012. ISO, Geneva.

ISO 24617-1:2012 Language Resource Management - Semantic Annotation Framework, Part 1: Time and events. ISO International Standard, Spring 2012. ISO, Geneva.

ISO CD 24617-4:2013 Language Resource Management - Semantic Annotation Framework, Part 4: Semantic roles. ISO Committee Draft, March 2013. ISO, Geneva.

Johnson, Christopher R., Charles J. Fillmore, Esther J.Wood, Josef Ruppenhofer, Margaret Urban, Miriam R. L. Petruck, and Collin F. Baker, 2001. The FrameNet Project: Tools for Lexicon Building. Unpublished Report, University of Berkeley.

Kamp, Hans and Uwe Reyle, (1993) From discourse to logic. Kluwer, Dordrecht.

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

Miller, George A., 1990, WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4) pp. 235312

Miltsakaki, Eleni, Livio Robaldi, Alan Lee and Aravind Joshi (2008) Sense Annotation in the Penn Discourse Treebank. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science Vol. 4919. Springer, Berlin, pp. 275-286.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank:An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71106.

Petukhova, Volha, Harry Bunt and Amanda Schiffrin (2007) LIRICS semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings 6^{th} International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Schiffrin, Amanda and Harry Bunt. 2007. LIRICS Deliverable D4.3. Documented compilation of semantic data categories. `http://lirics.loria.fr`.