An Information Extraction Approach to the Semantic Annotation of Folktales

Thierry Declerck
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
declerck@dfki.de

Antonia Scheidel
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
66123 Saarbrücken, Germany
Antonia.Scheidel@dfki.de

ABSTRACT
We propose an Information Extraction (IE) approach to the automated semantic annotation of folktales. We introduce and motivate the type of templates that we consider for encoding the (possibly underspecified) information extracted from textual and linguistic units of the tales. Each template is (possibly partially) instantiated on the base of a combined use of linguistic annotation and semantic resources. We opt for an incremental strategy: already instantiated templates can be further specialized on the base of subsequently instantiated templates. Once the full text has been processed, a round of specialization and of merging of the instantiated templates can take place.

1. INTRODUCTION
The work we describe here is part of the projects CLARIN[1] and D-SPIN[2]. While CLARIN is focusing on the establishment of an integrated and interoperable research infrastructure of language resources and technologies that aims at enabling eHumanities research in cooperation with Human Language Technology (HLT), the D-SPIN project, which is the German contribution to CLARIN, is additionally providing for integrated language processing Web services that generate linguistic annotation, which can be concretely used in eHumanities research.

A use case in CLARIN/D-SPIN, conducted in cooperation with the AMICUS Network[3], is investigating the possibilities of an automated processing of folktales that generates annotation that can be exploited by specialists in this specific field of narratives. We propose for this an Information Extraction (IE) strategy, which is applied on linguistically

---

[1]http://www.clarin.eu/

[2]http://weblicht.sfs.uni-tuebingen.de/

[3]AMICUS – Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts – is a research network on the topic of computational models of motifs in cultural heritage text and in scientific communication. See http://amicus.uvt.nl/

annotated folktales. On the base of such annotation, relevant IE units of a tale are detected. With *IE relevant units* we mean textual and linguistic units out of which basic information related to an event taking place in a particular temporal interval can be extracted and encoded in a corresponding IE template. Great importance is thus given to the recognition of temporal expressions in tales, which provides for a semantic means for text segmentation[4].

The filling – or instantiation – of the templates is done on the base of a combination of linguistic annotation and semantic resources, which will be described in more details in the paper. We adopt an incremental approach: while a template is being generated for each IE unit of the folktale text, this template can remain underspecified and be more specifically instantiated on the base of information detected and extracted in the context of subsequent units. Once the complete text has been processed and the corresponding amount of templates generated and (possibly partly) instantiated, an additional round of template comparisons, filling and merging can start, so that each template is fully specified[5].

The IE task is limited in a first step to providing for an automatic extraction of the characters of a tale, the particular relations existing between them, and the events they are involved in. Beyond this, we aim at establishing profiles of the characters, including their emotional states (if any), and we plan to collect information about all kind of objects mentioned in the tale. The IE templates we implement at this stage can in a sense be considered as giving the basic and generic information about the content of the tale.

First on the top of the instantiated (generic) templates, a classification of the characters in terms of character types and of the events in term of actions that correspond to a specific theory (like the typology of narrative structures suggested by Vladimir Propp in [11]) can be envisaged. We

---

[4]The detection of spatial expressions is also quite important, but we do not use (yet) this information for segmenting a tale.

[5]This IE approach is guided by our actual work in the Monnet project. Monnet (Multilingual ONtologies for NETworked knowledge) is a FP7 R&D project co-funded by the European Commission with Grant No. 248458. See also http://www.monnet-project.eu/. While Monnet deals with e-Government and Business Information use cases, we are here testing the Monnet approach to Ontology-based information extraction when applied to a new domain.

assume that elements of arbitrary theories of narratives[6] can be added to the templates within a specific slot or field. We also assume that our approach can support the recognition of the motifs of a tale, since following Uther "a motif can be a combination of statements about an actor, an object, or an incident - [or of] all three of these elements"[7]. The IE templates contain this information, but we still need to investigate how to compute the relevant "combination".

The paper is organized as follows: we give first a brief description of the kind of linguistic annotation we use. This is followed by an introduction to IE. We present then some of the semantic resources we have consulted for supporting the IE process and the resulting semantic annotation. And finally we expand on the motivation of the IE templates used in our actual work and the incremental approach we follow.

2. LINGUISTIC ANNOTATION

We follow two annotation strategies[8]:

1. Stand off annotation, meaning that the annotation is not added to the text, which we call here *primary data*, but resides in an external data structure that is containing a referential system for pointing back to segments of the text.

2. A multi-layer approach to (linguistic) annotation. From the linguistic point of view, we annotate the tales with the following information:

   - Segmentation in tokens: in EN, GE, FR etc., sequences of characters separated by punctuation or blanks, and the punctuation signs.

   - Morpho-Syntactic properties of tokens. If a token is a verb, specify its person and tense, etc.; if a token is a noun, specify its gender, number, case, etc. Tokens are upgraded to word forms.

   - Constituency: grouping word forms in phrases (nominal phrase, verbal phrase, prepositional phrase, etc.) clauses and sentences.

   - Dependency: grammatical relations between elements of constituents (head elements vs modifiers, etc.) and between constituents (subject, direct object, etc.)

   - Semantic relations at the linguistic level (for example time, space, co-reference etc.)

This annotation strategy (stand-off and multi-layered) is not restricted to the linguistic data, but is valid for all information we want to use for annotating the tales. We give a short and simplified example of the possible linguistic annotation of the tale *The Magic Swan Geese*, which we take from [10]. The annotation is displayed here in an in-line fashion in order to ease readability, and we show only the morpho-syntactic and constituency annotation levels, as they are applied to five tokens of the tale:

---

[6]We think here at the analysis of narratives proposed by Greimas in [4] or by Bremond in [3].

[7]This quotation from http://oaks.nvg.org/uther.html

[8]In compliance with ISO recommendations on the annotation of linguistic data, see [6] for details.

```
<wordForms>
<W ID="w11" POS="ART" LEMMA="the"
   MORPH="Sg" tokenID="t11">the</W>
<W ID="w12" POS="NN" LEMMA="daughter"
   MORPH="Sg" tokenID="t12">daughter</W>
<W ID="w13" POS="ADV" LEMMA="soon"
   tokenID="t13">soon</W>
<W ID="w14" POS="ADV" LEMMA="enough"
   tokenID="t14">enough</W>
<W ID="w15" POS="VVFIN" LEMMA="forget"
   MORPH="Past" tokenID="t15">forgot</W>
...
</wordForms>
```

In the morpho-syntactic annotation above, the value of the TokenID of the 12th word is pointing to the original data (*daughter* is the 12th token in the text).

In the constituency annotation level displayed below, words are grouped into syntactic constituents (e.g. the nominal phrase *the daughter*). The span of constituents is marked by the value of the features `from` and `to`, which are pointing to the previous morpho-syntactic annotation layer.

```
<phrases>
<phrase id="p4" from="w11" to="w12" type="NP">
   the daughter</phrase>
<phrase id="p5" from="w13" to="w14" type="ADVP">
   soon enough</phrase>
<phrase id="p6" from="w15" to="w15" type="VG"/>
   forgot</phrase>
<phrase id="p7" from="w16" to="w20" type="REL_COMP">
   what they had told her</phrase>
...
</phrases>
```

On the top of this linguistic annotation, which is described in more details in [9], one can add additional annotation layers, like the various results of IE.

3. INFORMATION EXTRACTION

In this section we give a brief and selective introduction to Information Extraction (IE), presenting the elements that are playing a role for our current research related to the semantic annotation of folktales. We base our introduction on the slides of the lectures *Intelligent Information Extraction* given by Günter Neumann and Feiyu Xu at the ESSLLI Summer School 2004 in Nancy[9]. We just "verbalize" the relevant slides for our purpose, modifying slightly the original text and adding some more extensive explications. The slides contain also references to many classical papers on IE.

As Neumann & Xu state, the goal of IE is to build systems that find and link relevant information from text and to fill up predefined data records/templates with this information. The input to IE is thus twofold: templates that encode the type of information that is of interest for an application, and the textual documents out of which the concrete information can be extracted for filling up (or instantiate) the templates. Core problems of IE are the identification of a general mapping strategy between text fragments and template descriptions and the specification of all possible textual paraphrases

---

[9]See http://www.dfki.de/~neumann/ie-esslli04.html

for a relevant IE natural language expression. As a concrete example, we can consider a template representing a company profile, with respective fields for the name, the legal status, the branch of activity, its address, the number of employee, the name of the members of boards, etc. of a particular company. Natural Language Processing (NLP) of textual documents will be used for helping in finding names of companies and the associated information, and encode this as instances of the template representing a normalized company profile. IE can thus be considered as an interface between natural language processing and domain knowledge[10].

IE is traditionally subdivided in 5 sub-tasks:

- Named Entity (NE) task. Mark into the text each string that represents a person, organization, or location name, or a date or time, or a currency or percentage figure.

- Template Element (TE) task. Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text (TE consists in generic objects and slots for a given application)

- Template Relation (TR) task. Extract relational information on employee_of, manufacture_of, location_of relations etc. (TR expresses domain independent relationships between entities identified by TE)

- Scenario Template (ST) task. Extract prespecified event information and relate the event information to particular organization, person, or artifact entities (ST identifies domain and task specific entities and relations)

- Co-reference (CO) task. Capture information on co-referring expressions, i.e. all mentions of a given entity, including those marked in NE and TE (Nouns, Noun phrases, Pronouns)

Interesting *shared tasks* in the field of IE were in the past the series of Message Understanding Conferences (MUC)[11]:

- MUC-1 (1987) and MUC-2 (1989) dealing with messages on naval operations

- MUC-3 (1991) and MUC-4 (1992) dealing with news articles on terrorist activity

- MUC-5 (1993) dealing with news articles about joint venture on microelectronics

- MUC-6 (1995) dealing with news articles on management changes

- MUC-7 (1997) dealing with news articles on space vehicle and missile launches

A commonality between all those editions of MUC was that they concentrated on the detection of events and their related arguments. So for example for the detection of events of succession in corporate executive personal, the IE systems had to detect not only the event, but the related position, the name of the persons involved, the reason for the change, the organizations involved (where is the new person coming from, where is the leaving person going to, etc.)

The ACE (Automated Content Extraction) program (1999-2008)[12] was another shared tasks initiative in the broader field of IE. A goal of this program was to develop core information extraction technology by focusing on the detection of specific semantic entities and relations over a very wide range of texts, and so discouraging highly domain- and genre-dependent solutions. ACE stressed the importance of detecting *unique* entities, relations, events and to find *all of their mentions* in documents. The relevance of ACE for our research can be summarized by the following points:

- Syntactic analysis of the text is a vehicle for organizing the information

- Toward the detection of each entity, relation, and event of a specific type

- Recognize all mentions of entities, relations and events, including the resolution of all mentions of the proper entity, relation, or event

- Convert information in human language into structured data, since structured data supports knowledge modeling & analysis

- Extract semantics of communication (this point was particularly missing from MUC)

ACE proposes a way towards the specification of the components of a broader semantic model for the content of different types of text.

- Entities – Individuals in the world
  - Simple entities: singular objects
  - Collective entities: sets of objects of the same type

- Attributes – Timeless unary properties of entities

- Temporal points and intervals

- Relations – Properties that hold of one or more entities over a time interval

- Events – A particular kind of relation among entities implying a change in relation state at the end of the time interval.

This type of semantic model is guiding our approach to IE applied to folktales, especially for the first processing step consisting in identifying generic characters, relations, and events. We need therefore to identify what kind of linguistic mentions refer to different components of this model, taking into account here for example the different types of phrases (nominal phrases for entities, prepositional phrases for relations, pronouns for co-reference, etc., naturally dependent on the language in use).

We close our summary of the lectures by Neumann & Xu by stressing that we have to deal with text types that have not been considered till now by the large IE shared tasks campaigns mentioned above. We note for example that in the 10 tales we have been looking at, only very few Named Entities are used. Especially persons are not named very often. And a main type of relation between persons is the one of family relation. Also names of locations are very seldom. And the temporal expressions used are widely underspecified ("one day", "later", "when she came back", "evening" or "winter"). So that compared to the standard IE tasks, we can not normalize extracted temporal expressions to calendar dates and times, but have to confine ourselves to a topological representation of time. One of the consequences for our IE approach is that we take stronger advices from a temporal ontology, as described in [8], and from a family ontology, which is currently under development[13].

As a general strategy for the semantic annotation of folktales, we will first remain at the level of the extraction of entities, relations and events, corresponding roughly to the semantic model of ACE, before trying to further specify the entities, relations and events in terms of a specific theory of folktales or narratives. We identified various semantic resources for guiding our semantic annotation of folktales, and those are briefly described in the next section.

4. SEMANTIC RESOURCES
Besides the temporal and family ontologies mentioned in the former section, we consider the use of FrameNet[14] at the level of the extraction of generic information. For the theory specific annotation we are for sure considering resources in the field of folktales, like the ATU (Aarne-Thompson-Uther) classification system[15] and Vladimir Propp's seminal work *Morphology of the Folktale* (see [11]). And we plan to extend our work for populating the ProppOnto Ontology[16], which we can not present here.

We note that while the ProppOnto ontology or the PftML annotation scheme[17] represent a formalized account of certain aspects of the theory of Propp, we are not aware of any formalization of (parts of) the ATU classification system.

[13] Similar to http://www.owldl.com/ontologies/family.owl but with more complex relations and extended to topics of fairy tales.
[14] http://framenet.icsi.berkeley.edu
[15] http://en.wikipedia.org/wiki/Aarne-Thompson_classification_system
[16] http://www.fdi.ucm.es/profesor/fpeinado/projects/kiids/apps/protopropp/
[17] PftML – Proppian fairy tale Markup language, see http://clover.slavic.pitt.edu/sam/propp/theory/propp.html. See also section 4.4 below for a short discussion of PftML.

4.1 FrameNet
We started to investigate the use of FrameNet (FN)[18] as a semantic resource. FN is dealing with the creation of lexical resources based on frame semantics. FrameNet is available for four languages (English, German, Spanish and Korean), whereas we are aware of developments for Italian as well. The FrameNet consortium developed corpora, annotated with syntactic and grammatical roles information associated to the semantic frames[19].

The motivation behind the use of FN is the ability mark up natural language expressions with relational frame semantics. For example we can annotate the verb "rewarded" (in one version of *Red Little Riding Hood*, the hunter who saved the hero is rewarded with wine), with the semantic Frame Element (FE) `Rewards and punishments`. This Frame Element specifies following core arguments to the `reward.v` lexical unit (the letter v staying for the Part-of-Speech `verb`): *Agent*, *Evaluee* and *Reason*. On the base of this frame semantics, we can map natural language expressions to those frame arguments (filling a corresponding template). FN also allows for further non-core arguments that can be associated with the lexical unit: *degree*, *manner*, *instrument*, ..., *place* and *time*. For all of those arguments, the IE engine is trying to find corresponding text segments. FN also provides a list of associated lexical items, with their corresponding Part-of-Speech, which are associated with the same Frame Element `Rewards and punishments`:
`discipline.v`, `punish.v`, `recompense.v`, `reward.v`.

FrameNet proposes a hierarchy of FEs, and for example `Rewards and punishment` is inherited from `Intentionally affect`, which is inherited from `Intentionally act`, which is itself a sub-type of `Event`. Due to this inheritance structure we are able to detect and annotate relevant events in the tales, and also to classify those along the lines of the subclasses of the `Event` Frame Element.

However, we have to note that the examples in the corpus of FN are mostly taken from newspapers, whereas we are dealing with texts belonging to the folktale genre. The question arises on how to enrich – in an automated fashion – FN with new types of annotated examples. Following this direction, our work would support not only intelligent access to folktales but also would also give feedback in a way that enables the enrichment of the lexical semantic resources of FN. We note that a specialized FN for the soccer domain is already available[20], and we will investigate if a similar specialization in the field of folktales can be proposed.

4.2 ATU
The Aarne-Thompson-Uther (ATU) classification system[21] analyzes folktales by motif, such as *Supernatural or enchanted relatives*, *Persecuted heroine* or *Wild and domestic animals*,

[18] See http://framenet.icsi.berkeley.edu and [2]
[19] This corpus resource is a reason why we prefer in our context FN to WordNet (WN, see http://wordnet.princeton.edu/), since in FN syncategorematic information is associated with lexical units and through this with the corresponding semantic frames.
[20] See http://www.kicktionary.de
[21] http://en.wikipedia.org/wiki/Aarne-Thompson_classification_system

but is also a source of vocabulary, since the names of the tales that are categorized under the types reveal some of the typical characters and events that one can encounter in tales, so for example the motif type `Supernatural Opponents`[22]:

```
The Dragon-Slayer, 300
The Three Kidnapped Princesses, 301
The Giant Without A Heart, 302
The Twin Brothers, 303
Seven Sisters, Seven Brothers, 303A
The Trained Hunter, 304
The Twelve Dancing Princesses, 306
The Princess in the Coffin, 307
Rapunzel, 310
Killed by a Giant, 311
The Bluebeard, 312
The Magic Flight, 313
The Golden-Haired, 314
The Treacherous Sister, 315
The Mermaid in the Pond, 316
```

All the nouns, and other lexical units, listed in those titles of tales can be stored in a kind of gazetteer that can guide the IE process, like this is usually done for the recognition of Named Entities. But not only the lexical units are relevant, also the syntactic information is very valuable: so for example the information encoded in prepositional phrases: A princess can be *in* a coffin, someone can be killed *by* a Giant etc. We can relate this syntactic valency information to FEs of FrameNet and so get semantic roles associated with characters mentioned in the titles of the tales. And this again can help in semi-automatically define the templates for the folktale specific Information Extraction task.

A closely related semantic resource is the Thompson Motif Index[23]. Of particular interest for us is the fact that this index offers a kind of specialization of motifs, which can be considered as quite close to a taxonomy, as the example from the *Ogre* Index shows:

```
G500--G599. Ogre defeated
G500. Ogre defeated
G510. Ogre killed, maimed, or captured
G520. Ogre deceived into self-injury
G530. Ogre's relative aids hero
G550. Rescue from ogre
G560. Ogre deceived into releasing prisoner
G570. Ogre overawed
G580. Ogre otherwise subdued
```

In this example one can see again how we could transform this listing – and the vocabulary included in it – into related semantic frames, or even onto a real taxonomy (*being killed* as a sub-class of *being defeated*, formalizing thus Thompson Motif Index and maybe also parts of the ATU system in a semi-automatic manner.

---

[22]The digits in the listing are the so-called AT number entries
[23]We consulted here an online source: http://www.ruthenia.ru/folklore/thompson/

4.3   Propp's Morphology of the Folktale

This section is widely borrowed from [12], which is describing complementary material to the research presented in this paper.

From the analysis of Alexander Nikolayevich Afanasyev's collection of Russian tales (cf. [1]), Propp identified a number of common components, which we list below:

7 Character types. Propp puts forward the notion that the folktale know no more than seven *dramatis personae*: The villain, the donor, the helper, the princess and her father (sometimes treated as two dramatis personae, resulting in a total of 8), the dispatcher, the hero and the false hero.

31 Functions. At the heart of *Morphology of the Folktale* is the introduction and detailed description of 31 "functions", i.e. (mostly) actions which can be attributed to the dramatis personae of a folktale. According to Propp, every folktale consists of a subset of these 31 functions, arranged in one or more "move". The order of the functions is fixed, with a number of scrupulously defined variations. Functions are frequently divided into sub-functions: In the case of function A: *Villainy*, they range from A^1: *The villain abducts a person* to A^{19}: *The villain declares war*.

150 Elements. In Appendix I of *Morphology of the Folktale*, Propp provides what he calls a "list of all the elements of the fairy tale". The list contains 150 elements, distributed over six tables:

1. The Initial Situation

2. The Preparatory Section

3. The Complication

4. The Donors

5. From the Entry of the Helper to the End of the First Move

6. Beginning of the Second Move

Some of the 150 elements appear alone, others are grouped under a descriptive heading. If these "element clusters" are counted as one, as shown below in Fig. 1, the appendix contains 56 - as they shall tentatively be called in the following - narratemes.

About a third of the narratemes can be mapped directly to functions, such as the aforementioned *30-32. Violation of an interdiction*. Other narratemes can be combined to form an equivalent to a function (together, narratemes *71-77: Donors* and *78: Preparation for the transmission of a magical agent* can presumably be considered as a superset to the information expressed by function D: *First Function of the donor*.

```
30-32.  Violation of an interdiction
     30.  person performing
     31.  form of violation
     32.  motivation
```

Figure 1: Example for a narrateme

For the time being, our approach aims at extending the Proppian classification with a set of semantic relations, on the basis of the FrameNet approach.

4.4 APftML

APftML (Augmented Proppian fairy tale Markup Language)[24] is a markup scheme that combines linguistic, generic and domain-specific (folktales) semantic information. The scheme builds on and extends the mark-up language PftML (Proppian fairy tale Markup Language). PftML has been designed for transforming the grammar-like functions, subfunctions and the rules concerning their combination from *Morphology of the Folktale* into a DTD, allowing for an XML annotation of fairy tales. APftML extends and revises PftML in various ways, two of those being that the augmented scheme does not limit itself to the Proppian functions and the Proppian "information" is integrated in textual and linguistic annotation standards as proposed by TEI (Text Encoding Initiative) and ISO TC37/SC4 on language resources management. APftML, developed in parallel to our IE work, is the annotation scheme that is used for encoding the results of the IE applied to folktales.

5. THE IE TEMPLATES

We present now in an informal way the kind of templates we are using, in our two-level approach to IE applied to folktales.

5.1 The generic Semantic Roles

We designed the templates so that they contain the information about the *WHs* of the tale, namely *Who, WhatObject, When, Where, WhatAction, ToWhom, Why, How*, etc. We are following in this an approach, which is similar to the scheme defined by the MPEG-7 standard for the *structured textual annotation* of multimedia data[25]. In MPEG-7 this annotation has the function to add semantic metadata to the content analysis of images or videos, which very normally remains at the level of physical descriptors (also called *Low-Level Features*). We provide the information about the characters (active or passive), the relations between them, the time and place in which they are mentioned, the actions (or events) in which they are involved etc.

In a first phase, the values that can be given to those descriptors (or slots in the templates) are extracted directly from text, allowing in certain cases for normalization or for establishing equality of information on the basis of basic inferences that can be derived from our family ontology. For illustration we take the first sentence of the tale *The Magic*

Swan Geese[26]: "Once upon a time a man and a woman lived with their daughter and small son." The (simplified) corresponding template looks like:

```
When: (T1, past)
Where: Somewhere (P1, inferred:
     someone has to live somewhere)
Who: M1, W1,
     D1, S1
 age(S1) < age(D1), inferred)
WhatAction: Exist((M1,W1, D1, S1)
Updates: Introduction
     characters and relations
     hasChildren(M1,D1 & S1)
     hasChildren(W1, D1 & S1))
Speaker: Narrator
```

In this pseudo-logical representation, we just mark the fact that we have four characters introduced in the tale, and the relations existing between the man (M1) and the children and between the woman (W1) and the children. The family ontology and its associated rules allow us to group the daughter (D1) and the son (S1) under the class `Children`. But nothing allows us to state the M1 and W1 are married.

Within the *WH* features we include temporal information, which is also indicated by the tense of the verb *lived*, local information (giving the global context of the described situation) extracted from text or inferred. The values of *Who* are extracted from text on the basis of the heuristic that indefinite nominal phrases (NPs) are introducing referents, following here broadly theories like [5] or [7], and we use variables for naming those referents. Clearly this approach has to be adapted for languages not using indefinite NPs (or determiners). In the course of the tales, we then consider most of the occurrences of definite NPs as co-referent expressions (for example *the girl*, in "When the girl came back", will be co-referent to *daughter*, mentioned in the first sentence of the tale. Our concrete co-reference algorithm is making here also use of our family ontology, which is stating that both classes *daughter* and *girl* have female gender, and we thus do not rely solely on textual and linguistic clues. This ontology-based resolution of co-reference is already a big step toward a better semantic annotation of folktales, since the user searching for all actions involving *daughters* in tales, will not be forced to formulate her/his query in dependency of the strings that are present in the tale.

We can not consider all definite NPs as co-referring to formerly introduced referents. Examples are like "In the cabin was the old witch Baba Yaga...". In this case we have to deal with a Named Entity, and we consider that such expressions introduce a referent per se. But there are also cases where a character is first introduced by means of a definite NP, so for example the *swan geese* in "In swooped the swan-geese, snatched up the little boy, and flew away with him". Clearly

we have to adapt our approach here, and a way for dealing with this case, is to have the *Swan Geese* within a gazetteer for folktales, as we already mentioned in the section 4.2, and so to consider it as kind of Named Entity.

Another issue we have to deal with: In the *Magic Swan Geese* tale the girl sees in the field an oven (introduced in the tale by an indefinite NP, as this is expected by us) and gets involved in a discussion with it. But while she seats in the hut of Baba Yaga, a mouse told her: "She [Baba Yaga] is going to steam you, put you in the oven, ...". Here we can not avoid the co-referencing mechanism to start, since we have in the tale both an indefinite and a definite NP referencing to an *oven*. But due to the fact that in our list of referents, the first oven is included in a template having as information on the location a *field*, we can assume here that we have two different ovens, the second one being located in the hut and not co-referencing to the first one, and so also not to be considered as a character of the tale (for which we assume that there are either introduced by an indefinite NP or by a Named Entity).

Additionally to the *WH* information, we add the features *Speaker* and *Updates*. With updates we mean something very similar as in dynamic predicate logic (see [5]): every utterance is describing a change of information of the interpreter (or reader).[27] At this IE level we could already apply the Proppian theory, and add to the template the information that we have to deal with an `Initial State`, since this would be quite straightforward. We can also postpone this step till we have analyzed the full text and generated all generic templates related to IE relevant units.

For the second sentence: "Dearest daughter," said the mother, "we are going to work. Look after your brother! Don't go out of the yard, be a good girl, and we'll buy you a handkerchief." Here the template looks like:

```
When: (T2, <= T1,  per inference (pi))
Where: H1 = P1 (H(ouse), pi)
Who: Mother = W1 (pi)
ToWhom: D1
WhatAction: talking
About What:
    NextActionOf(W1)
    AdvicesAbout(Brother=S1, pi),
    InterdictionFor(D1, GoOut)), ...
Updates:
    HaveWork (E1)
    WillLeaveHouse(W1)
    Sibling(D1,S1)
    Has(H1, CY1)
    Speaks(M1, D1)
Speaker: Narrator and W(1)
```

The updates are already interesting here. We know that the mother is talking to the daughter, and we can recognize that

some commands/interdiction are formulated. At this level we could already apply the Proppian theory, and add to the template that an interdiction has been uttered to a central character of the tale. But we can also postpone this step and first see how often the *girl* is mentioned and in which kind of situation she is involved in the whole tale before marking the *girl* as the *hero* of the story.

Since in this sentence we have to do with a dialog, in which the persons are using the present form of the verbs, we know that we have to deal with another kind of events as if the narrator would be the sole "teller" (or speaker). We see in this particular example that the IE has to take into account two different "worlds": the factual one (the description of events by the narrator) and the description of possible factual worlds, as they are uttered by the participants of the dialog. We have to be careful in this case on extracting from a dialog only the relevant factual information (for example the formulated interdiction).

Such an approach to IE and domain-specific semantic annotation is particularly relevant when one considers all the possible co-reference linkings and more specially the anaphora used in the tale. Let us look again at the second sentence of the *Magic Swan Geese*, which we mentioned in the former section.

The interesting (and complex) point here is the fact that we deal with a dialog, introduced by the narrator. The mother speaks to her daughter, and says "We are going to leave for work". on the basis of the sole string sequence of this sentence, one can not resolve the anaphora *We*. A strategy would be to either consider the whole set of referents, or the two persons involved in this dialog, or just the mother. In the latter case, the co-reference algorithm can subsequently add other entities: if one looks at the next sentence of the tale *The father and mother went off to work*, we can then add the father (M1) to the set of entities denoted by the pronoun *We*. Since it seems to be easier to add referents to a pronoun in the course of the further analysis of the tale, then to remove members out of the set, we go for the minimal solutions, and after the analysis of this segment of the tale only the mother (i.e. the speaker in the first person) is added to the set of referents meant with *We*.

A case in which the attribution of a specific type to a character is definitely better postponed, and not directly attributed is the *snatching up of the boy*. We have in this textual context no full evidence that this is a kidnapping event, and we also can not categorize the Swan Geese as the villain. And in fact first when the girl reaches the hut of Baba Yaga, we can infer that the boy has been kidnapped and the Swan Geese is not the villain, but rather the witch (the Swan Geese can be considered as the Villain's helper, and we can add this information to the first template in which the Swan Geese is introduced as a neutral character.).

5.2 The assignment of character roles and functions

Just to more precisely motivate our two-level annotation strategy: We do not want to follow only the logic of Propp and assume (or infer) that the person receiving an interdiction, and violating this one, is automatically the hero of the

tale or of the story. We want to have this role assignment also supported by linguistic and semantic evidence. At the lowest level, this can be due to the frequency of the mentioning of a character (supported by a co-reference algorithm in order to make sure that really all the mentions are collected, see again the requirements of the ACE program, described in the section 3). We are currently in the process of writing some rules to allow to map the generic character information to the Proppian descriptors.

6. CONCLUSIONS

We have been presenting the possible components of an Information Extraction approach to the semantic annotation of folktales. We suggest to follow a two-steps procedure, and to adopt the kind of semantic model described by the ACE initiative for defining the templates of IE in a first processing stage. We described for this the type of linguistic annotation and of semantic resources we are using. The second IE processing stage is dealing with the theory specific annotation of folktales, for which we have been consulting the Aarne-Thompson-Uther classification system and Propp's Morphology of the Folktales.

Parallel to our IE approach, an annotation scheme, APftML, has been developed and will be used for annotating the folktales with the results of the IE process. Future work will be dedicated to extending our approach to Ontology-based Information Extraction, allowing to populate existing or future ontologies in the fields of folktale or narratives in general. We will also propose a multilingual extension of our work.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] A. Afanas'ev. *Russian fairy tales*. Pantheon Books, New York, 1945.

[2] H. Boas. From theory to practice: Frame semantics and the design of framenet. In S. Langer and D. Schnorbusch, editors, *Semantisches Wissen im Lexikon*, pages 129–160. Narr, Tübingen, 2005.

[3] C. Bremond. *La Logique du Récit*. Editions du Seuil, Paris, 1973.

[4] A. J. Greimas. *Sémantique structurale*. Larousse, Paris, 1966.

[5] J. Groenendijk and M. Stokhof. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–101, 1991.

[6] N. Ide and L. Romary. Representing linguistic corpora and their annotations. In *LREC 2006- The fifth international conference on Language Resources and Evaluation*. ELRA, 2006.

[7] H. Kamp. Discourse representation theory. In J. Verschueren, J.-O. Östman, and J. Blommaert, editors, *Handbook of Pragmatics*, pages 253–257. Benjamins, 1995.

[8] H.-U. Krieger, B. Kiefer, and T. Declerck. A framework for temporal representation and reasoning in business intelligence applications. In K. Hinkelmann, editor, *AI Meets Business Rules and Process Management. Papers from AAAI 2008 Spring Symposium. AAAI 2008 Spring Symposium: AI Meets Business Rules and Process Management, March 26-28, Stanford, CA, United States*, volume SS-08-01 of *Technical Report*, pages 59–70. AAAI Press, 2008.

[9] P. Lendvai, T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado. Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[10] P. Lendvai, T. Declerck, S. Darányi, and S. Malec. Propp revisited: Integration of linguistic markup into structured content descriptors of tales. In *Digital Humanities 2010*. Oxford University Press, 7 2010.

[11] V. Propp. *Morphology of the folktale*. University of Texas Press:, Austin, 1968.

[12] A. Scheidel and T. Declerck. Apftml – augmented proppian fairy tale markup language. In *Proceedings of the First AMICUS Workshop*, 2010.