Extraction of Dutch definitory contexts for eLearning purposes

Eline Westerhout and Paola Monachesi

Utrecht University

Abstract

The aim of the Language Technology for eLearning project is to facilitate the retrieval, management and distribution of learning material within a Learning Management System by exploiting Natural Language Processing techniques as well as semantic knowledge. One of the functionalities provided by the project is the possibility to create glossaries semi-automatically. Glossaries are derived from the learning objects in order to capture the exact definition which the author of these documents uses. A rule-based approach is employed to identify the relevant lexical and linguistic patterns which underlie the definition. In this paper, we discuss the grammar developed to identify the definitory contexts in the Dutch learning objects and we present the results of the quantitative evaluation.

1 Introduction

The aim of the European project Language Technology for eLearning (LT4eL)[1] is to show that the integration of Language Technology based functionalities and Semantic Web techniques will enhance the management, distribution and retrieval of the learning material within Learning Management Systems (LMS). The functionalities are being developed for eight languages represented in our consortium that is Bulgarian, Czech, Dutch, English, German, Polish, Portuguese and Romanian (Monachesi, Lemnitzer and Simov 2006b), (Monachesi, Cristea, Evans, Killing, Lemnitzer, Simov and Vertan 2006a).

Language Technology resources and tools, such as corpora and taggers which have been produced in the context of other projects are employed in the development of new functionalities that will allow the semi-automatic generation of metadata for the description of learning objects in an LMS: to this end, a keyword extractor is being developed (Lemnitzer, Vertan, Killing, Simov, Evans, Cristea and Monachesi 2007).

Furthermore, the project will integrate the use of ontologies, a key element in the Semantic Web architecture, to structure and retrieve the learning material within the LMS. An ontology of 1000 concepts for the domain of computer science for non-experts and eLearning has been developed as well as an English vocabulary and English annotated learning objects. The ontology should facilitate the multilingual retrieval of learning objects.

Another objective of the project is the semi-automatic construction of glossaries which will be built on the basis of the definitory contexts which are presented in the learning objects themselves in order to capture the exact definition which the author of these documents uses. This definition in many cases overrides a more general definition of the term.

[1] http://www.lt4el.eu

In the project, definitory contexts are identified in a bottom-up manner. First, a substantial amount of definitions are selected and annotated manually in the learning objects which are the asset of this project. From these examples, grammars with the complexity of regular languages are abstracted (cf. Muresan and Klavans (2002) for a similar approach). These language-specific grammars are applied to a test set from the same language in order to estimate their coverage.

In this paper, we focus on the definitory contexts attested in the Dutch learning objects and the grammar necessary to identify them. As a basis for the extraction and annotation, we use linguistically annotated learning material which has been converted into XML. This process is discussed in section 2. Our approach to the detection and extraction of definitory contexts is rule-based. The patterns covered by our grammar are discussed in section 3 and 4 while the grammar is presented in section 5. Section 6 deals with the results we have obtained with the current version of the grammar. In section 7, we compare our methodology with other approaches while section 8 contains our conclusions and suggestions for future work.

2 The corpus

The learning material which constitutes our corpus from which definitions are extracted, can have different formats, such as HTML, PDF or DOC. Figure 1 illustrates the conversion process from the original file into the final XML output which conforms to the LT4ELAna DTD. This DTD has been derived from the XCES DTD for linguistically annotated corpora (Ide and Suderman 2002). For our purposes, the XCES DTD has been enriched with elements which are relevant for our project and contains – besides the content of the original files (that is, information about layout and the text itself) – the possibility to encode information about part-of-speech, morphosyntactic features and lemmas. This information is used for the extraction of keywords and the detection of definitory contexts.

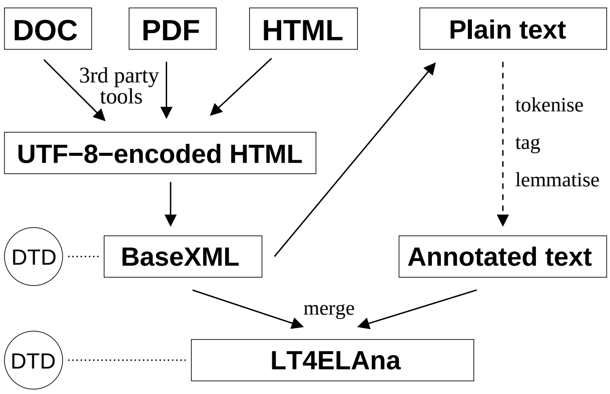

Figure 1: Data flow for the processing of learning objects.

The Wotan tagger (Daelemans, Zavrel, Berck and Gillis 1996) has been used for annotating the Dutch documents with part-of-speech information and morphosyntactic features whereas the CGN lemmatizer (Bosch and Daelemans 1999) was used for the lemmatization.

Figure 2 presents an example sentence in the LT4ELAna format. The *id* attribute is a unique identifier for each word, the *base* attribute contains the lemma of the word, the *ctag* attribute is related to the part-of-speech tag and the *msd* attribute gives the morpho syntactic information. The layout information is stored in the *rend* attribute. The rules of the grammar for the extraction of the definitory context are based on the information encoded in the LT4ELAna format.

```
<s id="s150">
  ...
  <tok id="t2254" class="word" base="het" ctag="Art"
    msd="bep,onzijd,neut">het</tok>
  <tok id="t2255" class="word" rend="b"
    base="eLearning-actieplan" ctag="N" msd="soort,ev,neut">
    eLearning-actieplan</tok>
  <tok id="t2256" class="punc" rend="b" base="." ctag="Punc"
    msd="punt">.</tok>
</s>
```

Figure 2: Part of a sentence in LT4ELAna format

3 The use of definitory contexts

Research on the detection and identification of definitory contexts has been pursued mainly in the context of question answering systems, where finding answers to definitory questions is a particularly difficult problem (cf. among others Miliaraki and Androutsopoulos (2004), Blair-Goldensohn, McKeown and Hazen Schlaikjer (2004) and Fahmi and Bouma (2006)). Very often pattern matching techniques are used to detect definitions such as the one exemplified below:

(1) Een vette letter is een letter die zwarter wordt afgedrukt dan de andere letters.
a bold character is a character that blacker is printed than the other characters

‘A bold character is a character which is printed darker than the other characters’.

Definitory contexts are expected to contain at least three elements: (1) the definiendum, that is the element that is defined (i.e. *een vette letter*), (2) the connector, which indicates the relation with the third element (i.e. *is*) and (3) the definiens, that is the definition of the definiendum (*een letter die zwarter wordt afgedrukt dan de andere letters*) (Walter and Pinkal 2006), (Fahmi and Bouma 2006). The number of patterns distinguished by the various systems differs largely. The documents used to extract definitory contexts are usually dictionaries or encyclopedias, which contain well structured definitions.

The LT4eL project is quite innovative with respect to the research in this area because it has adopted well known techniques to extract definitions and provided a totally new application: in the field of eLearning, identifying definitory contexts is relevant for the construction and maintenance of glossaries (Monachesi et al. 2006b). Furthermore, within our project the extracted definitions are employed in the construction of a domain ontology.

Glossaries are an important kind of secondary index to a text. They can be seen as small lexical resources which support the reader in decoding the text and understanding the central concepts which are conveyed. Since a glossary can be built on the definitory contexts which are present in the learning objects themselves, the advantages for the learning process are obvious: the learner accesses the appropriate definition which is the one used by the author of the learning object, which can in certain cases be different from the general definition of the term that could be found in a dictionary. For example, when we encounter the word ‘enter’ in a tutorial about Word, it will not have the meaning given by the Merriam Webster dictionary: ’to go or come in’. Instead, it will most times stand for the enter key and therefore have a completely different definition, that is: ’Also known as a return key, the enter key is used to return a cursor to the next line or execute a command or operation. It is common for most standard keyboards to have two enter or return keys, one on the keyboard and another on the numeric keypad’.

4 Types of definitory contexts

In order to identify the typology of definitions attested in our corpus, we have manually extracted 303 definitory contexts from our learning objects and grouped them according to the connector used. It should be mentioned that the collection of Dutch learning objects comprises 77 files within three different domains: computer science for non-experts (e.g. manuals on software programs), eLearning and the Pulman documents which deal with digitization. The average number of tokens per file is 6568 and the average number of types is 463.

The creation of the grammar has been done on the basis of the patterns found in 21 files. These 21 files contain 303 definitory contexts. We call this the training corpus. It should be noted that we are not using machine learning techniques yet, the files have not been used for training in the sense of training a classifier but only to identify the most common patterns. The test corpus consists of 14 files and has only been used for evaluating the grammar. It contains 159 definitory contexts

We distinguish three elements in definitory contexts (i.e. the definiendum, the connector and the definiens) in our approach. According to the patterns, the definitory contexts were classified into five groups:

1. Definitory contexts in which a form of the verb *zijn* (‘to be’) is used as connector verb;

```
Gnuplot is een programma om grafieken te maken .
'Gnuplot is a program for drawing graphs'
```

2. definitory contexts in which other verbs are used as connector (e.g. *betekenen* ('to mean'), *wordt ... genoemd* ('is called'), *wordt gebruikt om* ('is used to'));

```
E-learning omvat hulpmiddelen en toepassingen die via het
internet beschikbaar zijn en creatieve mogelijkheden
bieden om de leerervaring te verbeteren .
'eLearning comprises resources and application that are
available via the internet and provide creative possibilities
to improve the learning experience'
```

3. definitory contexts having specific punctuation features (e.g. *:*, *(..)*);

```
Passen: plastic kaarten voorzien van een magnetische strip,
die door een gleuf gehaald worden, waardoor de gebruiker
zich kan identificeren en toegang krijgt tot bepaalde
faciliteiten.
'Passes: plastic cards equipped with a magnetic strip, that
can be swiped through a card reader, by means of which the
identity of the user can be verified and the user gets
access to certain facilities'
```

4. definitory contexts in which the layout plays an important role (e.g. in tables, defined term in margin, defined term in heading);

```
RABE
Een samenwerkingsverband van een aantal Duitse bibliotheken,
die gezamenlijk een Internet inlichtingen dienst bieden,
gevestigd bij de gemeenschappelijke catalogus, HBZ, in
Keulen.
'RABE,
Cooperation of a number of German libraries, that
together provide an Internet information service, residing
at the common catalogue, HBZ, in Cologne'
```

5. definitory contexts in which relative and demonstrative pronouns (e.g. *dit* ('this'), *dat* ('that'), *deze* ('these')) and words like *hiermee* ('with this'), *hierdoor* ('because of this') are used to point back to an earlier used defined term. The definition of the term then follows after the pronoun, so these are often multisentence definitory contexts.

```
Dedicated readers.
Dit zijn speciale apparaten, ontwikkeld met het exclusieve
doel e-boeken te kunnen lezen.
'Dedicated readers.
These are special devices, developed with the exclusive
goal to make it possible to read e-books.'
```

Some definitions can be classified in more than one category. For these cases, we have chosen the category that was most important for the identification of the pattern. For example, in the last example, both the layout and the pronoun 'Dit' can be used as clues. We classified it as a pronoun definition, because 'dit' gives a stronger clue than the layout does. Table 1 shows how the definitory contexts are divided over the 5 types. From this table we can see that the definitions with a form of the verb *zijn* ('to be') as connector verb account for respectively 27.7 % and 38.4 % of the definitions and that in both the test and the training corpus around 40 % of the definitory contexts does not have a verb as main indicator.

	Training corpus	Test corpus
Type 1	84	61
Type 2	99	41
Type 3	46	13
Type 4	7	1
Type 5	46	27
Other patterns	48	31
# sentences	330	174
# definitory contexts	303	159

Table 1: Division of the definitory contexts into types

Although there are 303 definitions in the training corpus, we have more sentences, because definitory contexts have been identified which consist of more than one sentence (i.e. often two sentences are present). In most multisentence definitory contexts, one of the sentences contains only the defined term and no explanation of its meaning. These sentences in which only the defined term is mentioned do not meet our definition of a definitory context, and are therefore not identified by our grammar and also not mentioned in table 1. This is for example the case in the multisentence definitory context below:

```
Een gebruiker kan meer dan een programma tegelijkertijd
draaien. Dit wordt multi-tasking genoemd.
'A user can run more than one program at a time.
This is called multi-tasking.'
```

We leave out the sentences which contain only the defined term when we evaluate the performance of the grammar. As a consequence, we have only 27 multisentence definitory contexts left in the training corpus and 15 in the test corpus. The second part of the multisentence definitory contexts fits either in the fifth definition category or does not have a definitory context pattern. For these cases, both sentences give information on the meaning of the term defined.

```
TEX is een computerprogramma van Donald E. Knuth.
Het is speciaal ontworpen voor het zetten en drukken
van mathematische teksten en formules.
'TEX is a computer program developed by Donald E. Knuth.
It has been designed for setting and printing mathematical
texts and formulas.'
```

As already mentioned, most approaches to definition extraction use dictionaries or encyclopedias as corpus. This is not the case of our project in which the learning objects which constitute our corpus are mainly manuals and articles. As a consequence we have identified a variety of definitory context patterns which have not been taken into consideration in previous studies. This is the case for type 3, 4 and 5 patterns which are less common in dictionaries and encyclopedias. For some of these definitions, it is even not immediately clear that they are definitory contexts. The context of the patterns then determines whether or not we have to do with a definition. The type 3, 4 and 5 patterns make our work challenging and innovative.

5 The grammar

As already mentioned, in the LT4eL project, we have adopted a rule-based approach to the extraction of definitory contexts. Because of the variety of patterns present in our learning objects, we believe this is the best approach to use. Previous research has shown that grammars which match the syntactic structures of the definitory contexts are the most successful approaches if deep syntactic and semantic analysis of texts is not available (Muresan and Klavans 2002), (Liu, Chin and Ng 2003).

Therefore, we have developed a Dutch grammar in order to extract the definitory context patterns. The XML transducer *lxtransduce* developed by Tobin (2005) is used to match the grammar against files in the LT4eLAna format. Lxtransduce is an XML transducer, especially intended for use in NLP applications. It supplies a format for the development of grammars which are matched against either pure text or XML documents. The grammars must be XML documents which conform to a DTD (lxtransduce.dtd, which is part of the software). In each grammar, there is one “main” rule which calls other rules by referring to them. The XPath-based rules are matched against elements in the input document. When a match is found, a corresponding rewrite is done.

The grammar contains rules that match the grammatical patterns described in the previous section. The rules have been written on the basis of the 303 manually selected definitory contexts. At the moment, we focus on the extraction of patterns in which verbs are used as connector (type 1 and type 2). For type 3, we can extract the patterns with the colon as connector and the patterns between brackets. For type 5, we can extract patterns in which words like ‘hiermee’ (‘with this’) are used and definitions starting with ‘dit’ (‘this’). Type 4 has not been implemented yet.

The grammar consists of four parts. In the first part, the part-of-speech information is used to make rules for matching separate words (e.g. verbs, nouns, adverbs). The second part consists of rules to match chunks (e.g. noun phrases, prepositional phrases). We did not use a chunker, because we want to be able to put restrictions on the chunks. The third part contains rules for matching and marking the defined terms and in the last part the pieces are put together and the complete definitory contexts are matched. The rules were made as general as possible to prevent overfitting to our training corpus.

Figure 3 shows one of the rules described in the fourth part, namely the rule for the extraction of definitory contexts in which a form of *to be* (‘zijn’) is used as connector. The *name* attribute in the element *ref* refers to a previously described rule with this name, so the first element of the rule refers to a rule defined in the third part of the grammar with the name *markedTerm* and matches ‘een vette letter’. Thereafter, the verb is matched (‘is’). After the verb, a noun phrase follows (‘een letter’). The rest of the sentence is matched with the rule ‘tok_or_chunk’, which identifies the relevant material until the end of the sentence.

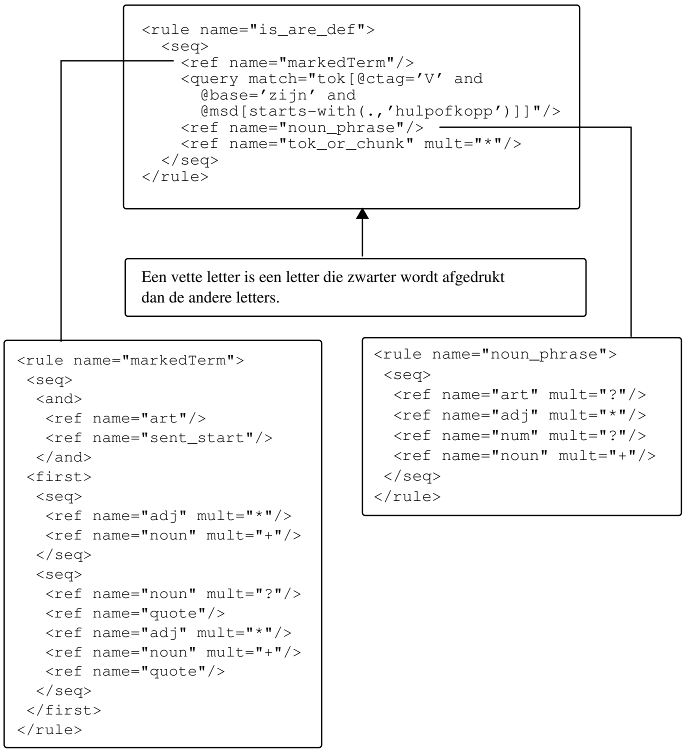

Figure 3: Grammar rule for extracting *is*-patterns

6 Results

The current grammar is able to detect type 1, type 2, type 3 and type 5 patterns. We have left type 4 (layout patterns) aside, for the moment, due to the low frequency of this pattern which makes the identification of the appropriate rules for its detection a complex task.

We calculated the performance of the grammar for each of the types in terms of precision, recall and F-score. In the evaluation, precision and recall were calculated at two levels: at the token level and at the sentence level, as both ways of the evaluation of definition extraction may be found in the literature. At the token level, precision is understood as the number of tokens simultaneously belonging to a manual definition and an automatically found definition, divided by the number of tokens in automatically found definitions. Correspondingly, recall is the ratio of the number of tokens simultaneously in both definition types to the number of tokens in manual definitions. At the sentence level, a sentence is taken as a manual or automatic definition sentence if and only if it contains a (part of a), respectively, manual or automatic definition. Given that, precision and recall are calculated in a way analogous to token level precision and recall. It is important to select the appropriate units when measuring precision and recall. We think for our task the sentence is the most appropriate unit and therefore we report the results obtained when using the sentence as a unit (Przepiórkowski, Degórski, Spousta, Simov, Osenova, Lemnitzer, Kubon and Wójtowicz 2007).

We did not only calculate the usual F-score, but also the F_2-score. In this score, recall is weighted twice as much as precision [2]. For the task at hand, where recall is more important than precision, the latter measure in which recall is measured seem appropriate (Przepiórkowski et al. 2007). The performance of the grammar has been evaluated for both the training set and the test set.

		Precision	Recall	F_1-score	F_2-score
Type 1	training	22.63	73.81	34.64	42.08
	test	20.97	91.80	34.15	43.19
Type 2	training	44.64	75.76	56.18	61.48
	test	25.76	41.46	31.78	34.46
Type 3	training	5.71	54.35	10.33	14.15
	test	2.58	76.92	4.99	7.25
Type 5	training	9.18	41.30	15.02	19.06
	test	6.15	40.74	10.68	14.16

Table 2: Performance of the grammar

For type 1 (the *is*-patterns), we had a recall of 73.81, a precision of 22.63 and an F_2-score of 42.08 on the training corpus and a recall of 91.80, a precision of 20.97 and an F_2-score of 43.18 on the test corpus (Table 2).

Within the test set, the grammar was able to detect 56 out of 61 definitory contexts. For three of the non-detected sentences, the verb 'is' was followed by an adverb or an adverbial used adjective. The other two sentence were not found due to an error of the part-of-speech tagger (e.g. the word 'uitwerken' (elaborating) was tagged as a verb, whereas it is used as a noun in this context). The recall is slightly better for the training set.

The type 2 patterns are those in which a verb different from *zijn* ('is') is used as connector. For the training corpus, recall was 74.76, precision was 44.46 and the F_2-score was 61.48. For the test corpus, both recall and precision were remarkably

[2] $F_\alpha = (1+\alpha) \cdot (\text{precision} \cdot \text{recall})/(\alpha \cdot \text{precision} + \text{recall})$. For F_2, $\alpha = 2$

lower, namely 41.46 and 25.76. The F_2-score on the test corpus was 34.46.

It should be noticed that a number of verbs can be used as connector, such as *betekenen* ('to mean'), *omvatten* ('to comprise'), *bestaan uit* ('to consist of'), *wordt gedefinieerd als* ('can be defined as'). However, there are also verbs that are used within definitory contexts that are normally not used as connector, such as the verb *voorkomen* ('prevent').

(2) Een vaste spatie voorkomt dat een regel tussen twee woorden
A non-breaking space prevents that a line between two words
wordt afgebroken.
is splitted
'A non-breaking space prevents a line from being splitted between two words'.

Whereas not everybody will consider this as a definition, they probably will consider the next sentence, which contains the same information, as a definition:

(3) Een vaste spatie is een spatie die voorkomt dat een regel
A non-breaking space is a space that prevents that a line
tussen twee woorden wordt afgebroken.
between two words is splitted
'A non-breaking space is a space that prevents a line from being splitted between two words'.

Because of the diversity of possible type 2 patterns, the recall score for type 2 is lower than the recall score for type 1. The precision is higher for type 2, because the patterns in which connector verbs different from a form of 'to be' are used, are less common in non-definitory contexts.

The third type of patterns comprises the patterns in which there is a punctuation character indicating that the sentence is a definition (e.g. a colon or brackets). The main problem for the identification of this type of definition is that it also occurs very often in non-definitory contexts. The precision is therefore very low (5.71 on training corpus and 2.58 on the test corpus). Recall is higher for the test corpus than it is for the training corpus (76.92 and 54.35 respectively), but the F-score is higher for the training corpus.

Within the type 5 patterns, two groups can be distinguished. The first group contains definitions starting with *dit* and the second group contains definitions starting with words like *hiermee*. The first type of definitions has roughly the same pattern as the type 2 definitions, whereas within the second type other patterns are used. All scores are higher for the training corpus: precision is 41.30 on the training corpus and 40.74 on the test corpus. Recall is respectively 9.18 and 6.15, and the F-scores are also higher for the training corpus.

In our project, we have a broad definition of what a definitory context is. Our learning objects present us with patterns that are often not attested in encyclopedias and dictionaries. Around 60 % of our patterns are standard definition patterns (i.e. definitions in which a verb is used as connector). However, this implies that we also have around 40 % non-standard patterns (that is, patterns of type 3, 4 or 5). Because of the variety of patterns attested in our corpus, we believe that a rule-based approach is the most appropriate for our task.

In the analysis of our results, we should take into account that there are several definition patterns that can also occur in non-definitory contexts. This is often the case for *to be* patterns and punctuation patterns and this has obviously a negative influence on the precision scores, as shown by the example below, which has the structure of a definition but it is obviously not a definition.

```
De stad is een belangrijke havenstad aan de Middellandse Zee.
'The city is an important port in the Mediterranean.'
```

Even though we used a state-of-the-art tagger (Bosch and Daelemans 1999), some of the definitory contexts were not found due to a tagger error. Most times, errors are nouns tagged as verbs (e.g. ‘leren’ in ‘Levenslang leren’ (‘learning’ in ‘lifelong learning’) or English words or commands (e.g. ‘sort’ referring to the Unix command ‘sort’ is tagged as verb). For the *zijn*-pattern, 27.3 % of the definitory contexts (6 definitory contexts) that were not found by the grammar, were not detected due to errors of the part-of-speech tagger.

6.1 Interannotator agreement

Because it is not always clear whether a sentence is a definitory context or not, it would be relevant to have more annotators expressing their judgments. We should let them analyze both the manually annotated definitory contexts to see whether they really are definitory contexts and the definitory contexts extracted by the grammar which were not marked by the annotator to check whether some of these can also be accepted as definitory contexts. These judgments could lead to the deletion and addition of some definitory contexts, which would result in an improvement of both precision and recall.

More generally, it would be relevant to identify the interannotator agreement in the annotation of definitions within our corpus and therefore we have carried out a small experiment to this end (Muresan and Klavans 2002). One of our texts was provided to three other persons which were asked to annotate the definitions and their headwords in this text. In total, 87 different sentences were marked as definitory context by the 4 annotators, 52 of which were unique.

We measured the interannotator agreement using Cohens kappa (κ) and several adapted versions of it (Table 3). Cohens kappa is the standard version of kappa. It assumes that the scores are equally divided over the categories. However, we have a large difference between the number of definitions and non-definitions in a text. Therefore, we also used another statistical measure in which this is taken into account. This score, the PABAK score (prevalence-adjusted bias-adjusted kappa), accounts for prevalence and bias of the data (Byrt, Bishop and Carlin 1993). The True Skill Statistic (TSS) can be used when one of the annotators is considered to be an expert (Allouche, Tsoar and Kadmon 2006). The annotation of the expert is then taken as model and the definitions marked by the other annotators are compared to this. In this case, we used our own annotation as expert annotation (annotator 4) and compared the results of the other annotators to these definitions.

Annotators	Cohens κ	PABAK	TSS
1 + 2	0.26	0.4	
1 + 3	0.27	0.43	
1 + 4	0.24	0.45	0.58
2 + 3	0.37	0.6	
2 + 4	0.42	0.69	0.77
3 + 4	0.42	0.74	0.62

Table 3: Interannotator agreement

The experiment with more annotators shows that the agreement between different annotators is not very high when definitions have to be annotated. From the fact that 87 different sentences were marked as definitory context by 4 annotators from which only 35 were marked by more then one person, we can already see that it is not easy to distinguish definitory contexts. The statistics in table 3 support this intuitive thought: both the Cohens κ score and the PABAK score show that the agreement between the different annotators is not very high. Although the agreement is better when we consider our own annotation as expert annotation and compare the others to this (TSS-scores) the agreement is higher, it is still not very high.

For measuring the interannotator agreement, it should be investigated which is the best statistical method to evaluate interannotator agreement for our purposes. Besides, the experiment should be repeated with a larger set of documents to make it possible to draw stronger conclusions.

7 Related work

Research on the detection of definitory contexts has been pursued mainly in the context of question-answering tasks. The answers to ‘What is’-questions are usually definitions of concepts. A common approach in this field is to search the corpus for sentences consisting of a subject, a copular verb and a predicative phrase. If the concept matches the subject, the predicative phrase is returned as answer. However, although the recall is high for this approach, the precision is often low, because there are many sentences which have the relevant syntactic form but are not definitions (Tjong Kim Sang, Bouma and de Rijke 2005). We encountered this problem within our approach for the patterns with a form of *zijn* (‘to be’) as connector. Fahmi and Bouma (2006) tried to solve this problem by applying machine learning techniques on the potential definitory contexts they extracted. They succeeded to improve the precision with 16.3 %. For this reason, we plan to adopt machine learning techniques to improve our results.

Within the German HyTex project (Storrer and Wellinghof 2006), 19 definitor verbs were distinguished on the basis of 174 manually extracted definitory contexts. Sentences in which one of these verbs was used were extracted. The results were calculated for each of the different definitors. They differed highly for the 19 verbs and depended also on the number of times the pattern was used. For the precision, the most problematic verb was the verb *sein* ('to be'), for which a precision of only 31 % was achieved. This is comparable to our precision score for this type of patterns. The recall was worst for the verb *nennen* ('to call') (20 %).

The DEFINDER system (Muresan and Klavans 2002) combines shallow natural language processing with deep grammatical analysis to identify and extract definitions and the terms they define from on-line consumer health literature. Four persons were provided with a set of nine articles, and were asked to annotate the definitions and their headwords in text. The gold-standard against which the system was compared, was determined by the set of definitions marked up by at least 3 out of the 4 subjects and consisted of 53 definitions. Nearly 60% of the definitions are introduced by a limited set of text markers '–', '()'), the other 40% being identified by more complex linguistic phenomena (anaphora, apposition, conjoined definitions). DEFINDER identified 40 out of the 53 definitions obtaining 86.95% precision and 75.47% recall. We used the same approach for one of our files to investigate whether this would lead to a different set of definitions. Because we used only one text, the differences for type 1 and type 2 were small compared to the results obtained by comparisaon to the set of definitions annotated by one person. It is difficult to compare our results to the DEFINDER results, because they use more structured texts.

8 Conclusions and future work

One of the functionalities developed within the LT4eL project is the possibility to derive glossaries semi-automatically on the basis of the definitory contexts identified within the learning objects.

A rule-based approach is employed to identify the definitory contexts. The current grammar is able to identify most types of definitory contexts and we obtain an acceptable recall while precision should be improved. However, due to the embedding of this functionality within an eLearning context in which human intervention is foreseen, the results are quite good.

At the moment, we are working on the improvement of the results at several levels.

First, we will investigate to which extent machine learning techniques can be used to improve the results and adopt an approach similar to Fahmi and Bouma (2006) to filter out unwanted results. More generally, we will have an identification step in which definitions will be detected on the basis of NLP techniques which will be followed by a filtering step based on machine learning techniques. We believe that we would always need to identify the definitions by means of a grammar, because this is the best approach to identify the relevant patterns and will enable us to to generalize the approach across all the languages involved in our project. Furthermore, we are not aware of machine learning approaches that account for the extraction of definitory contexts of type 3, 4 and 5.

As for the grammar, we will extend it with additional rules to cover also the less frequent patterns. In addition, we will investigate to which extent the grammar can be made more language independent. To this purpose, we are closely cooperating with the German and English grammar developers within the project to see whether the patterns of definitions are similar in closely related languages.

More generally, we wonder whether a quantitative evaluation is the best way to evaluate our results. Due to the variety of patterns attested and the lack of agreement among users about what should be considered a definition, it might be more appropriate to evaluate our grammar also from a qualitative point of view. Given the eLearning context in which we operate, the definitory contexts will be used to develop glossaries that are linked to the various learning objects, it might be thus more relevant to evaluate the degree of satisfaction of the users. These are both the content providers who will exploit this functionality in order to develop glossaries semi-automatically and they can thus select among the proposed definitions those that they consider the most appropriate as well as the learners who thanks to this functionality will have glossaries at their disposal that should facilitate their learning process.

References

Allouche, O., Tsoar, A. and Kadmon, R.(2006), Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss), *Journal of applied ecology* **43**(6), 1223–1232.

Blair-Goldensohn, S., McKeown, K. and Hazen Schlaikjer, A.(2004), *New Directions In Question Answering*, AAAI Press, chapter Answering Definitional Questions: A Hybrid Approach.

Bosch, A. v. d. and Daelemans, W.(1999), Memory-based morphological analysis, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL '99*, pp. 285–292.

Byrt, T., Bishop, J. and Carlin, J.(1993), Bias, prevalence and kappa, *J Clin Epidemiol* **46**(5), 423–429.

Daelemans, W., Zavrel, J., Berck, P. and Gillis, S.(1996), Mbt: A memory-based part of speech tagger generator, *in* E. Ejerhed and I. Dagan (eds), *Proceedings of the Fourth Workshop on Very Large Corpora*, pp. 14–27.

Fahmi, I. and Bouma, G.(2006), Learning to identify definitions using syntactic features, *in* R. Basili and A. Moschitti (eds), *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*.

Ide, N. and Suderman, K.(2002), XML Corpus Encoding Standard, document XCES 0.2, *Technical report*, Department of Computer Science, Vassar College, and Equipe Langue et Dialogue, New York, USA and LORIA/CNRS, Vandoeuvre-lés-Nancy, France,. `http://www.cs.vassar.edu/XCES/`.

Lemnitzer, L., Vertan, C., Killing, A., Simov, K., Evans, D., Cristea, D. and

Monachesi, P.(2007), Improving the search for learning objects with keywords and ontologies, *Proceedings of ECTEL 2007*.

Liu, B., Chin, C. and Ng, H.(2003), Mining topic-specific concepts and definitions on the web, *Proceedings of WWW-2003*.

Miliaraki, S. and Androutsopoulos, I.(2004), Learning to identify single-snippet answers to definition questions, *Proceedings of COLING 2004*, pp. 1360–1366.

Monachesi, P., Cristea, D., Evans, D., Killing, A., Lemnitzer, L., Simov, K. and Vertan, C.(2006a), Integrating language technology and semantic web techniques in elearning, *Proceedings of ICL 2006*.

Monachesi, P., Lemnitzer, L. and Simov, K.(2006b), Language technology for elearning, *in* W. Nejdl and K. Tochtermann (eds), *Proceedings of EC-TEL 2006*, Springer LNCS, pp. 667–672.

Muresan, S. and Klavans, J.(2002), A method for automatically building and evaluating dictionary resources, *Proceedings of the Language Resources and Evaluation Conference (LREC 2002)*.

Przepiórkowski, A., Degórski, L., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kubon, V. and Wójtowicz, B.(2007), Towards the automatic extraction of denitions in slavic, *Proceedings of BSNLP workshop at ACL*.

Storrer, A. and Wellinghof, S.(2006), Automated detection and annotation of term definitions in german text corpora, *Proceedings of LREC 2006*.

Tjong Kim Sang, E., Bouma, G. and de Rijke, M.(2005), Developing offline strategies for answering medical questions, *in* D. Mollá and J. Vicedo (eds), *Proceedings AAAI 2005 Workshop on Question Answering in Restricted Domains*.

Tobin, R.(2005), Lxtransduce, a replacement for fsgmatch. `http://www.ltg.ed.ac.uk/~richard/ltxml2/lxtransduce-manual.html`.

Walter, S. and Pinkal, M.(2006), Automatic extraction of definitions from german court decisions, *Proceedings of the workshop on information extraction beyond the document*, pp. 20–28.