

# Definition Extraction using Linguistic and Structural Features

Eline Westerhout  
Utrecht University  
*E.N.Westerhout@uu.nl*

## Abstract

In this paper a combination of linguistic and structural information is used for the extraction of Dutch definitions. The corpus used is a collection of Dutch texts on computing and elearning containing 603 definitions. The extraction process consists of two steps. In the first step a parser using a grammar defined on the basis of the patterns observed in the definitions is applied on the complete corpus. Machine learning is thereafter applied to improve the results obtained with the grammar. The experiments show that using a combination of linguistic (n-grams, type of article, type of noun) and structural information (layout, position) is a promising approach to the definition extraction task.

## Keywords

definition extraction, machine learning, grammar, linguistic features, text structure

## 1 Introduction

Definition extraction is a relevant task in different areas. Most times it is used in the domain of question answering to answer ‘What-is’-questions, but it is also used for dictionary building, ontology development and glossary creation. The context in which we apply definition extraction is the automatic creation of glossaries within elearning. Glossaries can play an important role within this domain since they support the learner in decoding the learning object he is confronted with and in understanding the central concepts which are being conveyed in the learning material.

The glossary creation context provides its own requirements to the task. The most relevant one is constituted by the corpus of learning objects which includes a variety of text genres (such as manuals, scientific texts, descriptive documents) and also a variety of writing styles that pose a real challenge to computational techniques for automatic identification and extraction of definitions together with the headwords. Our texts are not as structured as those employed for the extraction of definitions in question-answering tasks which most times include encyclopedias and Wikipedia. Furthermore, some of our learning objects are relatively small in size, thus our approach has not only to favor precision but also recall. That is, we want to make sure that as many as possible definitions present in a text are proposed to the user for the creation of the relevant glossary. Therefore, the

extraction of definitions cannot be limited to sentences consisting of a subject, a copular verb and a predicative phrase, as is often the case in question-answering tasks, but a much richer typology of patterns needs to be identified than in current research on definition extraction.

Different approaches for the extraction of definitions can be distinguished. We use a sequential combination of a rule-based approach and machine learning to extract them. As a first step a grammar is used to match sentences with a definition pattern and thereafter, machine learning techniques are applied to filter out those sentences that – although they have a definition pattern – do not qualify as definitions.

Our work has several innovative aspects compared to other work in this area. First, we address less common definition types in addition to ‘to be’ definitions. Second, we apply a machine learning algorithm designed specifically to deal with imbalanced datasets, which seems to be more appropriate for us because we have data sets in which the proportion of ‘yes’-cases is extremely low. The third innovative aspect on which this paper focuses has to do with the combination of different types of information for the extraction of definitions. Not only linguistic information (n-grams, type of article, type of noun) has been used, but also experiments with structural and textual information have been carried out (position, layout).

The paper is organized as follows. Section 2 introduces relevant work in definition extraction, focusing on the work done within the glossary creation context. Section 3 describes the data used in the experiments and the definition categories we distinguish. In section 4 the way in which grammars have been applied to extract definitions and the results obtained with them are discussed. Section 5 talks about the machine learning approach, covering issues such as the classifier, the features and the experiments. Section 6 reports and discusses the results obtained in the experiments. Section 7 provides conclusions and presents some future work.

## 2 Related research

Research on definition extraction has been pursued mainly in the context of automatic dictionary building from text, question-answering and ontology development. Initially, mainly pattern-based methods were used to extract definitions (cf. [12, 15, 16, 19]) but recently, several researchers have started to apply also machine learning techniques and combinations of

pattern-based methods and machine learning in this area (cf. [2, 9, 11]). [20] provides an overview of the work done in the different areas and compares it to the task within the glossary creation context.

Definition detection approaches developed in the context of question-answering tasks are often definiendum-centered, that is, they search for definitions containing a given term. Our approach, in contrast, is connector-centered, which means that we search for verbs or phrases that typically appear in definitions with the aim of finding the complete list of all definitions in a corpus independently of the defined terms. Despite the challenges that the eLearning application involves, we believe that the techniques for the extraction of definitions developed within the Natural Language Processing and the Information Extraction communities can be adapted and extended for our purposes.

Our work on definition extraction started within the European LT4eL project. Within the scope of this project experiments for different languages have been carried out. [13] describe experiments on definition extraction in Slavic languages and present the results obtained with Bulgarian, Czech and Polish grammars. The three grammars show varying degrees of sophistication. The more sophisticated the grammar, the more patterns are covered. Although the recall improves when more rules are added, the precision does not drop and is comparable for the three languages (22.3-22.5%).

For Polish, [10, 14, 7] put efforts in outperforming the pattern-based approach using machine learning techniques. To this end, [10] describe an approach in which the Balanced Random Forest classifier is used to extract definitions from Polish texts. They compare the results obtained with this approach to results obtained with experiments on the same data in which grammars were used [14] and to results of experiments with standard classifiers [7]. The best results are obtained with the approach designed for dealing with imbalanced datasets. The differences with my approach are that (1) they used either only machine learning or only a grammar and not a combination of the two, (2) they did not distinguish different definition types and (3) they only used relatively simple features, such as  $n$ -grams.

[3] applies Genetic Algorithms to the extraction of English ‘to be’ definitions. Her experiments focus on assigning weights to a set of features for the identification of such definitions. These weights act as a ranking mechanism for the classification of sentences, providing a level of certainty as to whether a sentence is actually a definition or a non-definition. They obtain a precision of 62% and a recall of 52 % on the extraction of is definitions by using a set of features such as ‘has keyword’ and ‘contains ‘is a’.

[8] focus on the extraction of Portuguese ‘to be’ definitions. First, a simple grammar is used to extract all sentences in which the verb ‘to be’ is used as main verb. Because their corpus is heavily imbalanced and only 10 percent of the sentences are definitions, they investigate which sampling technique gives the best results and present results from experiments that seek to obtain optimal solutions for this problem.

Previous experiments for Dutch focused on using a

grammar [22], and using several combinations of machine learning and a grammar to extract definitions [21, 23, 20]. A comparison of a standard classifier (naive Bayes) and the Balanced Random Forest (BRF) classifier showed that, especially for the more imbalanced data sets, the BRF classifier outperforms the naive Bayes classifier [20]. In all these previous experiments the features used were either only  $n$ -grams or a combination of  $n$ -grams and linguistic features.

### 3 Data

Definitions are expected to contain at least three parts. The definiendum is the element that is defined (Latin: that which is to be defined). The definiens provides the meaning of the definiendum (Latin: that which is doing the defining). Definiendum and definiens are connected by a verb or punctuation mark, the connector, which indicates the relation between definiendum and definiens [19].

Based on the connectors used in the 603 manually annotated patterns, four common definition types were distinguished. The first type are the definitions in which a form of the verb ‘to be’ is used as connector (called ‘is definitions’). The second group consists of definitions in which a verb (or verbal phrase) other than ‘to be’ is used as connector (e.g. to mean, to comprise). It also happens that a punctuation character is used as connector (most times the colon), such patterns are contained in the third type. The fourth category contains the definitory contexts in which relative or demonstrative pronouns are used to point back to a defined term that is mentioned in a preceding sentence. The definition of the term then follows after the pronoun. Table 1 shows an example for each of the four types.

### 4 Grammar

The first part of the extraction process is rule-based in our approach. Based on the part-of-speech tag patterns observed in the development part of the corpus a grammar was written to detect the four types of definitions. For a proper extraction of both sentences of multi-sentence pronoun definitions, anaphora resolution would have to be included in the system. As this is a completely different topic, we decided to restrict ourselves to only looking at the part of the definition containing the pronoun and connector verb (phrase). When the tool is integrated into the Learning Management System, it shows for each definition candidate one sentence to the left and one sentence to the right to see the context in which it is used. For the multi-sentence pronoun definitions this makes it possible to see which term is defined in the previous sentence and to select it manually.

The XML transducer LXTransduce developed by [18] has been used to match the grammars against files in XML format. LXTransduce is an XML transducer that supplies a format for the development of grammars which are matched against either pure text or XML documents. The grammars are represented in XML using the `lxtransduce.dtd` DTD, which is part

Type	Example sentence
is	Gnuplot is een programma om grafieken te maken <i>'Gnuplot is a program for drawing graphs'</i>
verb	E-learning omvat hulpmiddelen en toepassingen die via het internet beschikbaar zijn en creatieve mogelijkheden bieden om de leerervaring te verbeteren . <i>'eLearning comprises resources and application that are available via the Internet and provide creative possibilities to improve the learning experience'</i>
punctuation	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. <i>'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.'</i>
pronoun	Dedicated readers. Dit zijn speciale apparaten, ontwikkeld met het exclusieve doel e-boeken te kunnen lezen. <i>'Dedicated readers. These are special devices, developed with the exclusive goal to make it possible to read e-books.'</i>

**Table 1:** Examples for each of the definition types

of the software. A sentence is classified as a definition sentence if the parsing algorithm finds a match in this sentence of at least one token (not necessarily spanning the whole sentence).

type	R	P	F	F <sub>2</sub>
is	0.83	0.36	0.50	0.58
verb	0.75	0.45	0.56	0.61
punctuation	0.93	0.07	0.13	0.18
pronoun	0.64	0.09	0.16	0.21
all	0.79	0.16	0.27	0.34

**Table 2:** Results with the grammar

Table 4 shows the results obtained with the grammar. As can be seen from this table, the precision is quite low for all types, especially for the punctuation and pronoun types. The grammar rules were thus not specific enough to filter the incorrect sentences. To improve these low precision scores, machine learning has been applied on the grammar results.

## 5 Machine learning

The datasets obtained with the grammar are imbalanced, especially for the punctuation and pronoun definitions. Our interest leans towards correct classification of the smaller class (the ‘positive’ class), that is, the class containing the definitions. Therefore, a classifier specifically designed to deal with imbalanced datasets has been used, namely the Balanced Random Forest classifier. After describing how this classifier works, the features and feature settings are set out.

### 5.1 Balanced Random Forest Classifier

The Random Forest classifier is a decision tree algorithm, which aims at finding a tree that best fits the training data. Whereas normally the underlying tree is a CART tree, in the Weka package it is a modified variant of REPTree. The Weka algorithm follows the same methods of introducing randomness and voting of models. At the root node of the tree the feature that best divides the training data is used. In the Random Forest classifier [5] the *Gini index* is used as splitting measure.

In the Random Forest classifier there is not just one tree used for classification but an ensemble of trees [4]. The ‘forest’ is created by using bootstrap samples of the training data and random feature selection in tree induction. Prediction is made by aggregating the predictions of the ensemble. This idea behind Random Forest can be used in other classifiers as well and is called *bagging* (bootstrap **agg**regating).

A disadvantage of the Random Forest approach is that when data are extremely imbalanced, there is a significant probability that a bootstrap sample contains few or even none of the minority class. As a consequence, the resulting tree will perform poor when predicting the minority class. To solve this problem, [6] proposed the Balanced Random Forest classifier. This is a modification of the Random Forest method specifically designed to deal with imbalanced data sets using down-sampling. In this method a adapted version of the bagging procedure is used, the difference being that trees are induced from *balanced* down-sampled data. The procedure of the Balanced Random Forest (BRF) algorithm is described by [6]:

1. For each iteration in random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacement, from the majority class.
2. Induce a classification tree from the data to maximum size, without pruning. The tree is induced with the CART (Classification and Regression Trees) algorithm [5], with the following modification: at each node, instead of searching through all variables for the optimal split, only search through a set of  $m$  randomly selected variables<sup>1</sup>.
3. Repeat the two steps above for the number of times desired. Aggregate the predictions of the ensemble and make the final prediction.

### 5.2 Features

The features that have been used can be divided into five categories. Several combinations of these features resulted in 16 settings.

<sup>1</sup> [4] experimentend with  $m = 1$  and a higher value of  $m$  and concluded that the procedure is not very sensitive to the value of  $m$ . The average absolute difference between the error rate using  $F=1$  and the higher value of  $F$  is less than 1%

1. **Text properties:** these include various types of  $n$ -grams with different values for  $n$ .
2. **Syntactic properties:** features of this category give information on syntactic properties of the sentences, in these experiments the type of article used in definiens and definiendum are considered.
3. **Word properties:** in this category information on specific words is included, in these experiments, whether the noun in the definiens is a proper or a common noun.
4. **Position properties:** these include several features which give information on the place in the document where the definition is used.
5. **Lay-out properties:** this category contains features on layout information used in definitions.

## N-grams

In many text classification tasks  $n$ -grams are used for predicting the correct class (cf. [1] and [17]). For the classification of definitions two types of  $n$ -grams have been used, with  $n$  being 1, 2 or 3. We used Part-of-Speech tag (PoS-tag)  $n$ -grams. The tagger used distinguished 9 parts of speech: adjective, adverb, article, conjunction, interjection, noun, numeral, preposition, pronoun, verb. In addition it used the tag ‘Misc’ for unknown words and ‘Punc’ for punctuation marks.

## Articles

[9] investigated whether there is a connection between the type of article used in the definiendum (definite, indefinite, other) and the class of sentences (definition or non-definition). Although our definition corpus contains less structured texts than the data used by [9] (Wikipedia texts), part of the figures are quite similar for our data (table 3). In the Wikipedia sentences, the majority of subjects in definition sentences did not have an article (63%), which is the same in our corpus (62%). A difference with their data is the proportion of indefinite articles, which is 25% in our data and 13% in the data from [9].

	definition	non-definition
definite	12.8%	44.4%
indefinite	25.0%	8.3%
no article	62.2%	43.7%
other	0%	3.6%
	100%	100%

**Table 3:** Proportions of article types used in definiendum of *is*-definitions

The differences in distribution observed for the *is*-definitions is not seen to the same extent for the verb and punctuation definitions. In the verb definition candidates, for instance, both in definitions and non-definitions, definite articles tend to be used. However, also for these types there is a difference between definitions and non-definitions with respect to this feature.

The article used in the predicate complement has also been included. Again, we observe similarities and

differences between our data and the data from [9]. In both data sets the vast majority of articles tends to be indefinite at the start of the definiens (72% and 64%), which is quite different from the proportions for the non-definitions (30% and 29%). Differences between the two data sets are the proportion of definite articles in the definitions group (15% and 23%) and the proportion of no articles in the non-definitions (18% and 1%), which is much higher in the LT4eL data set.

	definitions	non-definitions
definite	14.7%	30.0%
indefinite	71.8%	30.0%
no article	9.0%	18.7%
other	4.5%	21.3%
	100%	100%

**Table 4:** Proportions of article types used at start of definiens in *is*-definitions

## Nouns

Nouns can be divided into two types, namely proper nouns and common nouns. Unfortunately, with our linguistic annotation tools it was not possible to get more detailed information about the type of proper noun (e.g. person, location), so we can only distinguish between proper and common nouns. The distribution of these types is different for definitions and non-definitions, especially for *is*-definitions. In the *is*-definitions the proportion of proper nouns in the definiendum is considerably higher for the definitions than for the non-definitions (53% versus 31%). For the other definition types the difference observed is much smaller.

## Layout

Because definitions contain important information you might expect special layout features (e.g. bold, italics, underlined) to occur more often in definitions than in non-definitions. Because in our data information on the original layout of the documents has been stored per word it was possible to check whether this was the case. No other research on definitions included this property in their research as far as we know. For each of the sentences it was indicated whether a specific layout feature was used in the definiendum. Because of the small numbers for some of the properties we decided to combine all layout features into one group. A comparison shows that *is*, *verb* and *punctuation* definition sentences contain significantly more layout information in the definiendum than non-definition sentences.<sup>2</sup> For each of the definition types the proportion of layout information is about twice as high in definitions than in non-definitions.

## Position

[9] in their research on definition extraction from Wikipedia texts reduced the set of definition candi-

<sup>2</sup> The pronoun definitions were not included in this investigation, because the definiendum of these sentences is often not in the same sentence as the definiens

dates extracted with the grammar by selecting only the first sentences of each document as possible candidates. It seems that Google’s define query feature also relies heavily on this feature to answer definition queries. However, as [9] also state, the first position sentence is likely to be a weaker predictor of definition versus non-definition sentences for documents from other sources, which are not as structured as Wikipedia. The texts from the LT4eL corpus are such less structured texts and therefore using this restriction would not be a good decision when dealing with these documents. In addition to being less structured, they are also often longer and contain on average 10.6 definitions, so applying the first sentence restriction would cause a dramatic decrease of recall and make it impossible to fulfil our aim of extracting as much definitions as possible because at most one sentence per document would be extracted using this method.

Although we thus cannot use the same restriction, it is nevertheless possible to include information on the position of the definition candidate in a document as feature in the machine learning experiments to see whether it helps the classifier in predicting the correct class. To this end, three types of positional information were included in the features, namely information on the position of the sentence within the paragraph, information on the position of the definition within the sentence and information on the (relative and absolute) position of the definiendum compared to other occurrences of the term in the document.

**Position in paragraph** Each document is divided into paragraphs which are again divided into sentences. It is thus possible to see where in the paragraph a definition is used. When we consider each paragraph as a separate block of information, we would expect definitions to appear at the beginning of such a block. The fact that sentence position is such a strong predictor in Wikipedia articles supports this idea.

The first property related to position in paragraph is the absolute position of the definition sentence within the paragraph. When we compare definitions and non-definitions with respect to this feature we see that for three of the four definition types the absolute position is lower for the definitions. Only of the pronoun definitions there is no significant difference. The pronoun definitions tend to be used later on in the paragraph compared to the non-definitions for this type. This might be caused by the fact that they are used more often at the second position of the paragraph where the term is mentioned in the first sentence.

In addition to the absolute position of a sentence, we also included a score on the relative position taking into account the number of sentences in a paragraph, because the beginning of a paragraph is a relative property. When we compare the scores on this property for definitions and non-definitions, for three of the four types there is a significant difference, only the result for the *punctuation*-definitions is not significant.

**Position in sentence** When we look at the four definition types, one of the differences observed is the place in the sentence where it can start and end.

Whereas *is* and *verb* definitions tend to span a complete sentence, the rules for punctuation definition are less strict for this feature. On the basis of this observation I investigated whether information on this could be used to distinguish definitions from non-definitions.

In addition to this, a second reason has to do with the conversion from original document to XML document. During this process sentences were split automatically and marked as <s>. However, not all sentences were splitted correctly, because the sentence splitter tool made errors sometimes which were not corrected manually. Therefore, an extra rule had to be used to detect the beginning of a sentences saying that each word starting with a capital could indicate the start of a sentence.

The position is given by indicating the number of tokens in the <s> before the definition starts. For all definition types, the absolute position of the definition candidate within the sentence is significantly lower for definitions than for non-definitions.

**Position of definiendum** When a term is defined, one would expect that it has not been used a lot of times before it is explained in the definition. Although it is possible that it has been used two or three times before already (e.g. in title of document, table of contents or heading), intuitively you would expect it to be used more after it has been explained. Based on this intuition three measures have been included.

The first two are the absolute number of occurrences of the term before and after it is used in the definition candidate. For all types the average number of occurrences before is lower for definitions. This difference is significant for all types except for the *is*-definitions. The number of occurrences of the term after it has been defined seems to be a less good predictor and is only significantly lower for the *is*-definitions. When we look at the relative position of the definiendum the score is significantly lower for the definition sentences for all types except the *is*-definitions for which there is no difference observed.

### 5.3 Feature settings

The first setting are the n-grams of part-of-speech tags. This setting is the baseline to which all other settings are compared. The four types of features – articles, nouns, position and layout – have been combined in all possible ways resulting in 16 settings in total. In the second group the four types of feature settings were tried separately (setting 2 to 5). Settings 6 to 11 are all possible combinations of two of the four settings. Then there are four settings (12 to 15) in each of which three types were combined and in the last setting all four types are integrated. Table 5 shows the settings.

## 6 Results

The final results after applying both the grammar and machine learning are shown in table 6. The sentences not detected with the grammar rules could of course not be retrieved anymore, and as a consequence the recall after applying machine learning is always lower

setting	IS				VERB				PUNCTUATION				PRONOUN			
	R	P	F	A	R	P	F	A	R	P	F	A	R	P	F	A
1.	0.57	0.49	0.53	0.60	0.58	0.54	0.56	0.56	0.51	0.16	0.24	0.74	0.40	0.15	0.22	0.64
2.	0.74	0.56	0.64	0.66	0.49	0.53	0.51	0.54	0.50	0.13	0.21	0.70	0.55	0.17	0.26	0.61
3.	0.49	0.47	0.48	0.58	0.49	0.43	0.46	0.43	0.43	0.11	0.18	0.68	0.49	0.21	0.29	0.70
4.	0.57	0.50	0.54	0.61	0.52	0.53	0.53	0.54	0.47	0.13	0.20	0.70	0.47	0.19	0.27	0.67
5.	0.17	0.52	0.26	0.61	0.15	0.56	0.24	0.53	0.39	0.14	0.21	0.76	0.57	0.09	0.15	0.21
6.	0.70	0.56	0.62	0.66	0.49	0.61	0.55	0.60	0.60	0.11	0.19	0.58	0.56	0.18	0.27	0.62
7.	0.64	0.63	0.64	0.71	0.56	0.56	0.56	0.58	0.53	0.15	0.24	0.73	0.47	0.22	0.30	0.72
8.	0.74	0.57	0.64	0.68	0.44	0.54	0.49	0.54	0.52	0.13	0.21	0.68	0.53	0.17	0.26	0.62
9.	0.54	0.52	0.53	0.62	0.56	0.56	0.56	0.57	0.50	0.15	0.23	0.73	0.45	0.18	0.26	0.68
10.	0.53	0.47	0.50	0.58	0.22	0.46	0.30	0.49	0.42	0.16	0.24	0.78	0.53	0.20	0.29	0.67
11.	0.57	0.52	0.54	0.62	0.52	0.51	0.51	0.52	0.48	0.14	0.21	0.72	0.44	0.19	0.26	0.68
12.	0.63	0.62	0.62	0.70	0.56	0.58	0.57	0.59	0.53	0.17	0.26	0.75	0.47	0.23	0.31	0.73
13.	0.69	0.57	0.62	0.67	0.51	0.64	0.57	0.62	0.53	0.14	0.22	0.70	0.57	0.19	0.29	0.64
14.	0.66	0.64	0.65	0.72	0.59	0.57	0.58	0.58	0.54	0.16	0.24	0.73	0.46	0.22	0.30	0.72
15.	0.57	0.53	0.54	0.63	0.53	0.52	0.52	0.53	0.45	0.14	0.21	0.73	0.47	0.20	0.28	0.70
16.	0.63	0.63	0.63	0.71	0.56	0.57	0.56	0.58	0.47	0.15	0.23	0.75	0.42	0.22	0.29	0.73

**Table 6:** Final results after applying grammar and machine learning

#	setting
1.	n-grams
2.	article
3.	noun
4.	position
5.	layout
6.	article + noun
7.	article + position
8.	article + layout
9.	noun + position
10.	noun + layout
11.	position + layout
12.	article + noun + position
13.	article + noun + layout
14.	article + position + layout
15.	noun + position + layout
16.	article + noun + position + layout

**Table 5:** The sixteen feature settings

than the recall obtained in the first step. For each experiment four measures are reported. The first three are the recall, precision, and f-score of the definition class. The fourth score is the overall classification accuracy. The separate results for the non-definition class are not shown. As the aim of the experiments is to improve the precision obtained with the grammar, this is the most important measure. However, recall and accuracy may not become too low and therefore also recall, f-score and accuracy are reported.

For each of the types it is described in this section how the results should be interpreted and to which extent the settings can compete with setting 1 (n-grams).

## 6.1 Results per type

**Is definitions** The first block of information in table 6 shows the results for the is definitions. We see that for this type the article is the best feature for classification. Using only this feature gives better results than the results obtained with the n-grams. The second best individual feature is the information on position, although for this type the results with the n-grams are almost the same. A combination of article, noun and position (setting 14) gives the best result, which is equally good as the result obtained with a combination of article and position (setting 7) and a

combination of all feature settings (setting 16).

For the layout setting the recall is very low, which is not strange given the fact that only in a small subset of the definitions there was special layout used. Although there is a slight improvement when it is used in combination with other features, the added value is not big. Adding the noun to other settings generally leads to either lower or similar classification results.

The maximum improvement of precision compared to the precision obtained with the grammar is 77.8% (setting 14).

**Verb definitions** The second group of definitions in table 6 are the verb definitions. For this type none of the individual settings outperforms the baseline set by the n-grams. The best feature here is position. Using a combination of features makes it possible to perform better than the n-grams. The highest precision is obtained with setting 13, which is a combination of article, noun and layout. The results with the layout setting are comparable to the results for the is definitions. The grammar precision for this type was 0.45 so the maximum improvement is 42.2% (setting 13).

**Punctuation definitions** For the punctuation definitions the accuracy is highly determined by the non-definitions, as these constitute over 90% of the data set. For the individual feature settings the best precision and accuracy are obtained with the layout setting, however, the recall is quite low for this type. Only one of the settings gives better results than the n-grams, namely setting 12 (article, noun and position). The maximum improvement of precision compared to the precision obtained with the grammar is 142.9% with this setting.

**Pronoun definitions** Just as for the punctuation definitions, the pronoun definitions data set is highly imbalanced. The noun is the most important individual feature setting, which is surprising as many of these definitions do not have a definiendum. In most settings the recall improves compared to the result on this score of the n-grams, but it often goes with a drop of precision. An overall improvement compared to the base line is observed in most of the settings, especially

in setting 7 (article and position) and 10 (noun and layout) and the best result is obtained with setting 12 (article, noun and position), which is considerably better than the result of setting 1. With this setting the increase of the precision score compared to the precision obtained with the grammar is 155.6% (setting 12).

## 6.2 General observations

When looking from the perspective of the settings, we see that the article and position in general are the best features. The problem with the layout feature setting mainly is that the recall obtained with it is quite low. Also, adding it as an extra feature to other settings does not lead to much improvement of these results.

A second general observation is that for none of the types the best results are obtained when a combination of all features is used. It is thus not the case that the more information is included the better results will be obtained. For all types one or more feature settings outperform the n-grams results.

## 7 Conclusions and future work

The influence of the inclusion of linguistic and structural features on classification accuracy differs per type and per combination of settings. Except for the layout setting all individual settings perform well on at least one of the definition types. Combining the different feature settings generally improves the results.

The precision improved in all cases. The two types on which the grammar performed best (is and verb) showed a substantial improvement of 77.8% and 44.2%. And even though precision was still low for punctuation and pronoun patterns after applying machine learning, the percentual improvement was huge for these types (142.9% and 155.6% respectively).

The fact that it is possible to obtain better results with linguistic and structural features than with part-of-speech n-grams is encouraging for several reasons. First, because it shows that it makes sense to use other information in addition to linguistic information (position and layout settings) and to structure the linguistic information (article and noun settings). A second issue is that those features provide us more insight on how definitions are used, which is relevant for research on definitions.

As the results are promising, future work will proceed in this direction. We plan to conduct experiments in which other feature settings that go beyond use of linguistic information are used in addition to the settings discussed in this paper. An example of such a setting is the importance of words in a text ('keywordiness'). Another future experiment will investigate whether the number of included n-grams (in these experiments we included all n-grams) can be decreased to lower the computational load while keeping the same results. Initial experiments with 100 n-grams for the is definitions did not show much decrease in performance.

## References

- [1] R. Bekkerman and J. Allan. Using bigrams in text categorization. Technical report, Technical Report IR, 2003.
- [2] S. Blair-Goldensohn, K. R. McKeown, and A. Hazen Schlaikjer. *New Directions In Question Answering*, chapter Answering Definitional Questions: A Hybrid Approach. AAAI Press, 2004.
- [3] C. Borg. *Automatic definition extraction using evolutionary algorithms*. PhD thesis, University of Malta, 2009.
- [4] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
- [5] L. Breiman, J. Friedman, R. Olshen, C. Stone, L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth, 1984.
- [6] C. Chen, A. Liaw, and L. Breiman. Using Random Forest to learn imbalanced data. Technical Report 666, University of California, Berkeley, 2004.
- [7] L. Degórski, M. Marcińczuk, and A. Przepiórkowski. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of LREC 2008*, 2008.
- [8] R. Del Gaudio and A. Branco. Extraction of definitions in portuguese: An imbalanced data set problem. In *Proceedings of Text Mining and Applications at EPIA 2009*, 2009.
- [9] I. Fahmi and G. Bouma. Learning to identify definitions using syntactic features. In R. Basili and A. Moschitti, editors, *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*, 2006.
- [10] L. Kobyliński and A. Przepiórkowski. Definition extraction with balanced random forests. In B. Nordström and A. Ranta, editors, *Advances in Natural Language Processing: Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008*, pages 237–247. Springer Verlag, LNAI series 5221, 2008.
- [11] S. Miliaraki and I. Androutsopoulos. Learning to identify single-snippet answers to definition questions. In *Proceedings of COLING 2004*, pages 1360–1366, 2004.
- [12] S. Muresan and J. Klavans. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference*, 2002.
- [13] A. Przepiórkowski, L. Degórski, M. Spousta, K. Simov, P. Osenova, L. Lemnitzer, V. Kubon, and B. Wójtowicz. Towards the automatic extraction of denitions in Slavic. In *Proceedings of BSNLP workshop at ACL*, 2007.
- [14] A. Przepiórkowski, M. Marcińczuk, and L. Degórski. Dealing with small, noisy and imbalanced data: Machine learning or manual grammars? In *Proceedings of TSD 2008*, 2008.
- [15] H. Saggion. Identifying definitions in text collections for question answering. In *Proceedings of the Language Resources and Evaluation Conference*, 2004.
- [16] A. Storrer and S. Wellinghof. Automated detection and annotation of term definitions in German text corpora. In *Proceedings of LREC 2006*, 2006.
- [17] C. Tan, Y. Wang, and C. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management*, 38(4):529–546, 2002.
- [18] R. Tobin. Lxtransduce, a replacement for fsgmatch, 2005. <http://www.ltg.ed.ac.uk/~richard/ltxml2/lxtransduce-manual.html>.
- [19] S. Walter and M. Pinkal. Automatic extraction of definitions from German court decisions. In *Proceedings of the workshop on information extraction beyond the document*, pages 20–28, 2006.
- [20] E. Westerhout. Extraction of definitions using grammar-enhanced machine learning. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 88–96, Athens, Greece, 2009. Association for Computational Linguistics.
- [21] E. Westerhout and P. Monachesi. Combining pattern-based and machine learning methods to detect denitions for elearning purposes. In *Proceedings of RANLP 2007 Workshop "Natural Language Processing and Knowledge Representation for eLearning Environments"*, 2007.
- [22] E. Westerhout and P. Monachesi. Extraction of Dutch definitory contexts for elearning purposes. In *Proceedings of CLIN 2006*, 2007.
- [23] E. Westerhout and P. Monachesi. Creating glossaries using pattern-based and machine learning techniques. In *Proceedings of LREC 2008*, 2008.