Folia Linguistica XXXI/1-2, 57-75, 1997

Data Mining as a Method for Linguistic Analysis: Dutch Diminutives*

Walter Daelemans
Computational Linguistics, Tilburg University
PO Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: Walter.Daelemans@kub.nl

Peter Berck and Steven Gillis
Center for Dutch Language and Speech
University of Antwerp
Universiteitsplein 1
2610 Wilrijk, Belgium
e-mail: Peter.Berck@uia.ua.ac.be
Steven.Gillis@uia.ua.ac.be

Abstract

We propose to use data mining techniques (inductive techniques for the automatic acquisition of comprehensible knowledge from data) as a method in linguistic analysis. In the past, such techniques have mainly been used in linguistic engineering applications to solve knowledge acquisition bottlenecks. In this paper we show that they can also assist in linguistic theory formation by providing a new tool for the evaluation of linguistic hypotheses, for the extraction of rules from corpora, and for the discovery of useful linguistic categories. By applying a rule induction method to a particular linguistic task (diminutive formation in Dutch) we show that data mining techniques can be used to test linguistic hypotheses about this morphological proces, and to discover interesting morphological and phonological rules and categories.

* Preparation of this paper was supported by a Research Grant of the Fund for Joint Basic Research (FKFO 2.0101.94) of the National Fund for Scientific Research (NFWO) and by a VNC project of NFWO - NWO (contract number G.2201.96), and a grant from the Research Council of the University of Antwerp.

1. Introduction

The dominant view about the role of computers in linguistics has been that computer modeling is a useful tool for enforcing internal consistency, completeness, and empirical validity of the linguistic theory being modeled. In this paper, we argue that by using *data mining* techniques, the role of computation in linguistics can be significantly broadened.

Data Mining is a branch of computer science concerned with the automatic extraction from data of implicit and previously unknown information which is nontrivial, understandable, and useful, using techniques from Machine Learning and Statistical Pattern Recognition (clustering, rule induction, classification, etc.) (Piatetsky-Shapiro & Frawley 1991).

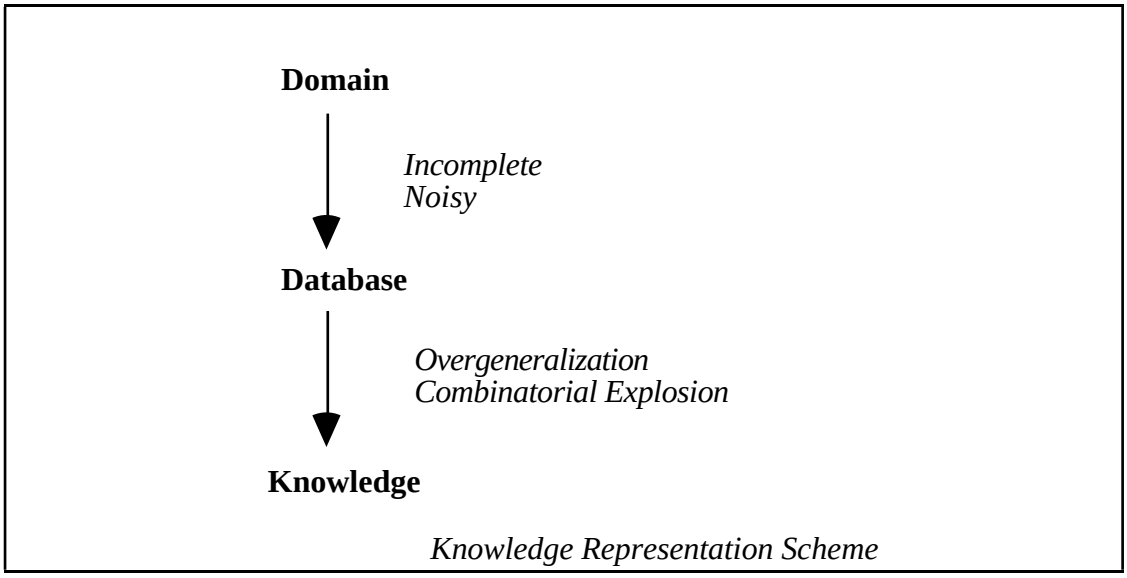

Figure 1: The process of data mining

Data Mining (Figure 1) presupposes a database (or several databases) containing data from a particular domain. The collected data may constitute a noisy and incomplete description of the domain. The Machine Learning or pattern Recognition technique is used to extract structured knowledge from these data, in terms of a knowledge representation scheme such as if-then rules or decision trees. During this process problems of *overgeneralization* (a learning system should go beyond the data, making inductive leaps which may be incorrect) and *combinatorial explosion* (in general a computationally intractable number of hypotheses can explain the same data) should be dealt with. Moreover this process of knowledge extraction occurs on the basis of incomplete and noisy data.

In linguistic applications, the domain comprises the (hypothesized) regularities (in terms of rules and concepts) governing linguistic behavior that the linguist wants to discover. The data are corpora of actual language use, and the resulting knowledge is the proposed theory about the regularities accounting for the corpora.

In this paper we will show that there are basically two ways in which data mining techniques can be used in linguistic theory formation:

- Evaluation of hypotheses: In order to evaluate competing theories about a linguistic phenomenon, the following procedure can be applied:
  1. Collect a representative corpus of data about the phenomenon.
  2. Annotate the corpus according to the requirements of each competing theory or hypothesis, i.e. implement an annotation using the concepts deemed necessary by each different theory or hypothesis for the description of the phenomenon.
  3. Compute the learnability of the phenomenon using a learning algorithm or by analysing the different annotations of the corpus using statistical techniques.

  By comparing the performance of the system when trained on these differently annotated corpora, claims about the necessity of particular information for the explanation of the phenomenon can be tested.
- Discovery of theories: In order to discover new generalizations or concepts, the following procedure can be used:
  1. Collect a representative corpus of data.
  2. Annotate the corpus with any linguistic information which may be relevant.
  3. Extract generalizations and categories using learning algorithms.

For instance, if one wants to evaluate different competing theories about sentence accent computation, a corpus can be collected with appropriate sentence accent marking. Different hypotheses about which type of linguistic knowledge plays a role in sentence accent assignment can be evaluated by encoding the corpus using these different knowledge sources (phonological, morphological information, syntactic tags, discourse structure, etc.) By comparing the learnability of sentence accent using these contrastively annotated corpora as training material, the hypotheses can be evaluated. Alternatively, a maximally annotated corpus, i.e. an annotation incorporating all possibly relevant knowledge, can be used to infer new generalizations about the role of different levels of annotation and their interaction.

In this paper we will describe a case study in the application of data mining techniques as a method in linguistic analysis. A generally available inductive rule learning algorithm, viz. C4.5 (Quinlan 1993), will be used to test linguistic hypotheses and to discover regularities and categories. The data mining technique will be used in the domain of allomorphy in Dutch diminutive formation, which is, according to Trommelen (1983), "one of the more vexed problems of Dutch phonology" and "one of the most spectacular phenomena of modern Dutch morphophonemics". We will use C4.5 to illustrate both the hypothesis testing and the linguistic discovery aspects of data mining.

2. Dutch Diminutive Formation

2.1 Allomorphy in Dutch Diminutive Formation

In standard Dutch, the diminutive is formed by attaching a form of the Germanic suffix *-tje* to the base form of a noun (diminutives of other word classes appear far less frequently and their meaning is often lexicalized, hence we will only deal with nouns). The suffix shows allomorphic variation. The five variants are exemplified in Table 1.

Table 1: Diminutive allomorphy in Dutch.

Allomorph	Example	IPA Transcription	Gloss
-tje (/ʧə/)	kikker-tje	/kɪkərʧə/	'frog-DIM'
-etje (/əʧə/)	roman-etje	/ro:mɑnəʧə/	'novel-DIM'
-pje (/pjə/)	lichaam-pje	/li:χa:mpjə/	'body-DIM'
-kje (/kjə/)	koning-kje	/ko:nɪŋkjə/	'king-DIM'
-je (/jə/)	wereld-je	/we:rəlʧə/	'world-DIM'

The frequency distribution of the variants is given in the Table 2. The CELEX lexical database (Burnage 1990) was consulted and all items marked as diminutivized nouns were collected. The frequency of each diminutive allomorph is represented in Table 2 in terms of its raw frequency and the percentage of each allomorph on the total number of diminutivized nouns. For the purpose of comparison, the corpus frequency, i.e. the relative frequency of each allomorph in the text corpus on which the word list was based, is added.

Table 2: Frequency distribution of diminutive allomorphs in Dutch

Suffix	Frequency	Database %	Corpus %
-tje	1896	48.0	50.9
-je	1478	37.4	30.4
-etje	395	10.0	10.9
-pje	104	2.6	4.0
-kje	77	1.9	3.8

The allomorph *-tje* is the most frequent one, followed by *-je*. The three other allomorphs are far less frequent: in the database 10% of the nouns take the suffix *-etje*, and approximately 2% take the allomorphs *-pje* and *-kje*.

2.2 Linguistic Analysis of Diminutive Formation

Diminutive formation in Dutch has a long history. Te Winkel (1862) was presumably the first to propose an analysis. Since then, analyses have taken different rules for the choice of the diminutive suffix and the linguistic concepts that play a role (e.g., Kruizinga 1915, Cohen 1958, Haverkamp-Lubbers & Kooij 1971, Trommelen 1983, Booij & Van Santen 1995, De Haas & Trommelen 1993).

The desciptive generalizations as they appear from these studies can be summarized as follows:

-je is used after an obstruent.

E.g. pop-je /pɔpjə/ 'doll-DIM'

-pje is used after a long vowel, diphthong or schwa followed by /m/ (*lichaam, pluim, bezem*) and after a short vowel followed by a liquid (/r/ or /l/) plus /m/ *(olm).*

E.g.	lichaam-pje	/li:χa:mpjə/	'body-DIM'
	pluim-pje	/plœympjə/	'feather-DIM'
	bezem-pje	/be:zəmpjə/	'broom-DIM'
	olm-pje	/ɔlmpjə/	'elm-DIM'

-etje is used after a nasal (/m/, /n/ or /ŋ/) or the liquid /l/ preceded by a short vowel (*roman*, *bal*). This allomorph is also used with monosyllabic words ending in /r/ preceded by a short vowel (*bar*). The latter restriction distinguishes monosyllabic words from polysyllabic words such as *radar* (/ra:dɑr/, 'radar') which take the allomorph *-tje.*

E.g.,	roman-etje	/ro:mɑnətʃə/	'novel-DIM'
	bal-etje	/bɑlətʃə/	'ball-DIM'
	bar-etje	/bɑrətʃə/	'bar-DIM'

ring-etje	/rɪŋətʃə/	'ring-DIM'

-kje is used in multisyllabic words ending in /ɪŋ/) (*koning*), which is to be distinguished from monosyllabic words (such as *ring*) which take the allomorph *-etje*. The allomorph *-etje* is also used when the penultimate syllable is unstressed, as in *zoldering* (diminutive: *zoldering-etje).* If the penult is stressed, the suffix *-kje* is used as in *soldering* (diminutive: *soldering-kje*). But not all words with penultimate stress take *-kje* as a suffix: *koning*, *teerling* do, but *leerling* and *tweeling* take the suffix *-etje*.

E.g.			
	koning-kje	/ko:nɪŋkjə/	'king-DIM'
	ring-etje	/rɪŋ/	'ring-DIM'
	zoldering-etje	/zɔldərɪŋ/	'ceiling-DIM'
	soldering-kje	/sɔlde:rɪŋ/	'soldering-DIM'
	teerling-kje	/te:rlɪŋ/	'die-DIM'
	leerling-etje	/le:rlɪŋ/	'pupil-DIM'
	tweeling-etje	/twe:lɪŋ/	'twin-DIM'

-tje is the default which applies in those conditions not stipulated above.

In addition to these general rules which assign a particular allomorph, certain wordforms allow an additional diminutive suffix (see De Haas & Trommelen 1993 for a complete description and a formal specification). We will provide only a few examples of these 'double diminutives':

Some words that take *-etje*, also allow *-tje*, *-pje* or *-kje* depending on the final consonant. More specifically this alternation occurs with disyllabic trochaic words that have a short vowel followed by a nasal or the liquid /l/ in the second syllable. If the word ends in /n/ or /l/ *-tje* is possible in addition to *-etje*, if the word ends in /m/ *-pje* is also allowed, and if the word ends in /ŋ/, *-kje* is allowed in addition to *-etje*.

E.g.,			
	sultan-etje	*sultan*-tje	'sultan-DIM'
	consul-etje	*consul*-tje	'consul-DIM'
	pelgrim-etje	*pelgrim*-pje	'pilgrim-DIM'

Monosyllabic words ending in the obstruent /p/, /b/ or /ɣ/ allow *-etje*, next to the regular *-je*:

E.g.,			
	pop-je	*pop*-etje	'doll-DIM'
	kip-je	*kip*-etje	'chicken-DIM'
	rug-je	*rug*-etje	'back-DIM'

Words ending in a long vowel followed by a sonorant, take either *-pje* (after /m/) or *-tje* according to the rule mentioned above. In many cases also the allomorph *-etje* is

possible. Note however that this alternation may involve a shift in meaning.

E.g.,	*bloem*-pje	'flower-DIM'	('a little flower')
	bloem-etje	'flower-DIM'	('a little flower' or 'a bunch of flowers')

Trommelen (1983) argues that diminutive formation is a local phenomenon in which concepts such as word stress and morphological structure (proposed in earlier analyses) do not play a role.[1] The description of diminutive formation presented above can be captured in terms of a metrical model of the syllable (inspired by Selkirk 1982). Only the rhyme of the last syllable is necessary and sufficient for predicting the correct diminutive allomorph.

The natural classes (or concepts) which are used in the description of diminutives in Dutch and hypothesized in the rules for diminutive formation include *obstruents*, *sonorants* and *bimoraic vowels* (the class consisting of long vowels, diphthongs and schwa). According to Trommelen the concept of word stress and morphologically relevant concepts are unnecessary to account for the allomorphic variation.

In summary, diminutive formation in Dutch is a relatively transparent linguistic domain for which different competing theories have been proposed, and for which different generalizations (in terms of rules and linguistic categories) have been proposed. In the following sections we will show how data mining techniques may be used to (i) test competing hypotheses, and (ii) discover generalizations in the data which can be compared to the generalizations formulated by linguists. We will first introduce the data mining technique used in this study.

3. Machine Learning Method

In the data mining experiments that will be focused on in the the next sections, we used C4.5 (Quinlan 1993). C4.5 incorporates an inductive learning algorithm that is geared to constructing a decision tree on the basis of a set of examples. The algorithm can be formulated in a simplified way as follows:

Procedure **make-decision-tree**

---

[1] Trommelen proposes a morphological motivation for distinguishing the final /ɪŋ/ in *koning* from that in *leerling*, a distinction which is not uncontroversial, see e.g. Booij (1984). Consequently although her analysis does not use morphological information, the underlying segmental representation is morphologically 'informed' in a number of cases.

Given a training set *T* (a collection of examples) and a finite number of classes C_1 ... C_n:

- If T contains one or more cases all belonging to the same class Cj, then the decision tree for T is a leaf node with category C_j.
- If T is empty, a category has to be found on the basis of other information (e.g. domain knowledge). The heuristic used here is that the most frequent class in the initial training set is used.
- Otherwise,
  - Choose a test (feature) with a finite number of outcomes (values), and partition T into subsets of examples that have the same outcome for the test chosen. The decision tree consists of a root node containing the test, and a branch for each outcome, each branch leading to a subset of the original set.
    - Apply **make-decision-tree** recursively to all subsets created this way.

Thus the algorithm creates a decision tree in which tests (feature names) are the nodes and in which the outcome of the tests (feature values) are the branches. The leaf nodes are labeled with a category name and constitute the output of the system. The description of the algorithm does not specify how a test is chosen to split a node into subtrees at a particular point. Taking a test at random usually results in a large decision tree that hardly captures solid generalizations, as uninformative tests may be chosen. Considering all possible trees consistent with the data is computationally intractable so that a reliable test selection procedure is called for, a procedure that takes into account the 'relevance' of a test in the task domain. The method used in C4.5 is based on the concept of *information gain.* Selection of a test is based on the information gain of the features: the feature that maximizes the information gain is chosen. For the actual computation of information gain and the additional normalization of the measure adjusting for the bias in favor of tests with many outcomes, we refer to Quinlan (1993: 20-24).

Decision trees can be automatically transformed into sets of if-then rules (or production rules), which are easier to understand by domain experts; linguists in our case. C4.5 also contains a value grouping method which collapses different values of a feature into the same category on the basis of statistical information. This results in more concise decision trees and rules because a rule condition or branch of the tree can make reference to a class of values instead of requiring a rule condition of branch for each value seperately.

In order to assess whether C4.5 has learned the problem, i.e. to assess the extent to

which the tree incorporates a good representation of the regularities of the problem domain, the *generalization* accuracy is determined. Generalization accuracy is measured by using the induced decision tree (or the induced rules) on a set of words that were not part of the training set. The specific method is discussed in the following section.

4. Experiment 1: Learnability

4.1 Method

The training set used in the experiment consists of a set of 3950 diminutivized nouns from the CELEX database. For each noun the following information was included in the training set:

1. The phonemic transcription of the word and its syllabic structure: the onset, nucleus and coda of the three last syllables of each word is used as training material.
2. For each syllable it is indicated whether it carries main stress or not.
3. The correct diminutive allmorph(s) is (are) indicated.

The format of the training items is exemplified in the following table.

Table 3: Format of training items as used in the induction experiment

Feature	Example1	Example2	Example3	Example4
Stress Antepenultimate	-	=	=	=
Onset Antepenultimate	b	=	=	=
Nucleus Antepenultimate	i	=	=	=
Coda Antepenultimate	=	=	=	=
Stress Penultimate	-	=	+	+
Onset Penultimate	z	=	b	b
Nucleus Penultimate	ə	=	ɛɪ	ɛɪ
Coda Penultimate	=	=	=	=
Stress Ultimate	+	+	-	-
Onset Ultimate	m	b	b	b
Nucleus Ultimate	ɑ	ɪ	a	ə
Coda Ultimate	nt	χ	n	l
Diminutive Suffix	JE	ETJE	TJE	TJE

Table 3 shows four examples of training items: *biezenmandje* (example 1, 'basket'),

big (example 2, 'piglet'), *bijbaan* (example 3, 'job on the side'), and *bijbel* (example 4, 'bible'). A syllable's stress is indicate by '+' for main stress, and '-' for secundary stress or unstressed. An equality sign ('=') means that there are no values for that feature in that word.

The generalization accuracy of C4.5 was assessed in a ten-fold cross-validation experiment (Weiss & Kulikowski 1991). In this experimental procedure the database is partitioned ten times, each time a different 10% of the database is used as test set and the remaining 90% as training set. The cross-validation was stratified, i.e. the 10 folds are stratified so that they contain approximately the same proportions of diminutive allomorphs as the complete database (see Table 1). The success rate of the algorithm is obtained by calculating the average accuracy over the ten test sets in the stratified ten-fold cross-validation experiment. This means that success rate of the algorithm is taken to be the average number of test pattern categories (the diminutive allomorph) correctly predicted over the ten experiments.

In this experiment the learning material ranged from 100 words to the full database of 3950 words. Up to 1000 items, 100 words were added in each experiment. From 1000 words onwards 250 words were added to the learning material.

4.2 Results

In Figure 3 the overall learning curve is displayed: this learning curve expresses the the system's generalization accuracy for increasing sizes of training sets as the mean percentage over the ten folds in the stratified ten-fold cross-validation experiment. With a training set of 100 items, a success score of 91% is reached. This score increases to 97% in the final experiment with 3950 words.

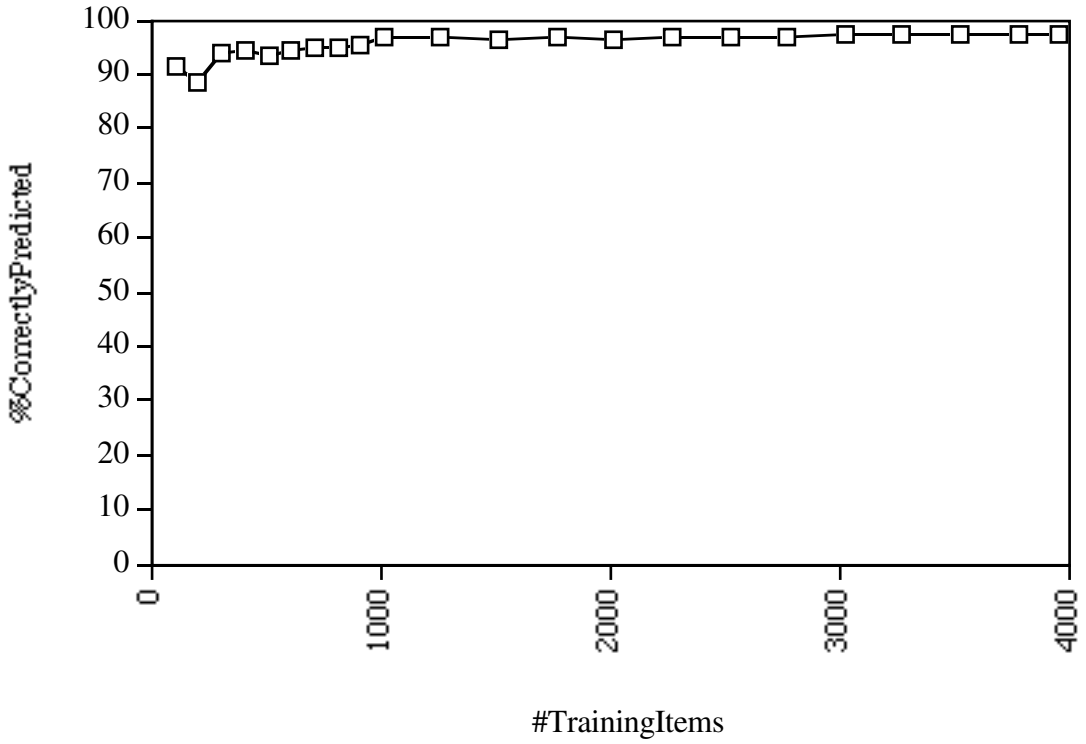

Figure 2: Global learning curve expressed as the average percentage correctly predicted diminutives of C4.5 with increasing number of examples in a stratified ten-fold cross-validation experiment.

This result clearly indicates that the diminutive formation problem is learnable in a data-oriented way, i.e. by extraction of regularities from examples. By observing the generalization performance of the system with different amounts of examples, we obtain an insight into the relative learnability of the different allomorphs. In Figure 3 the learning curves for the individual allomorphs are shown.

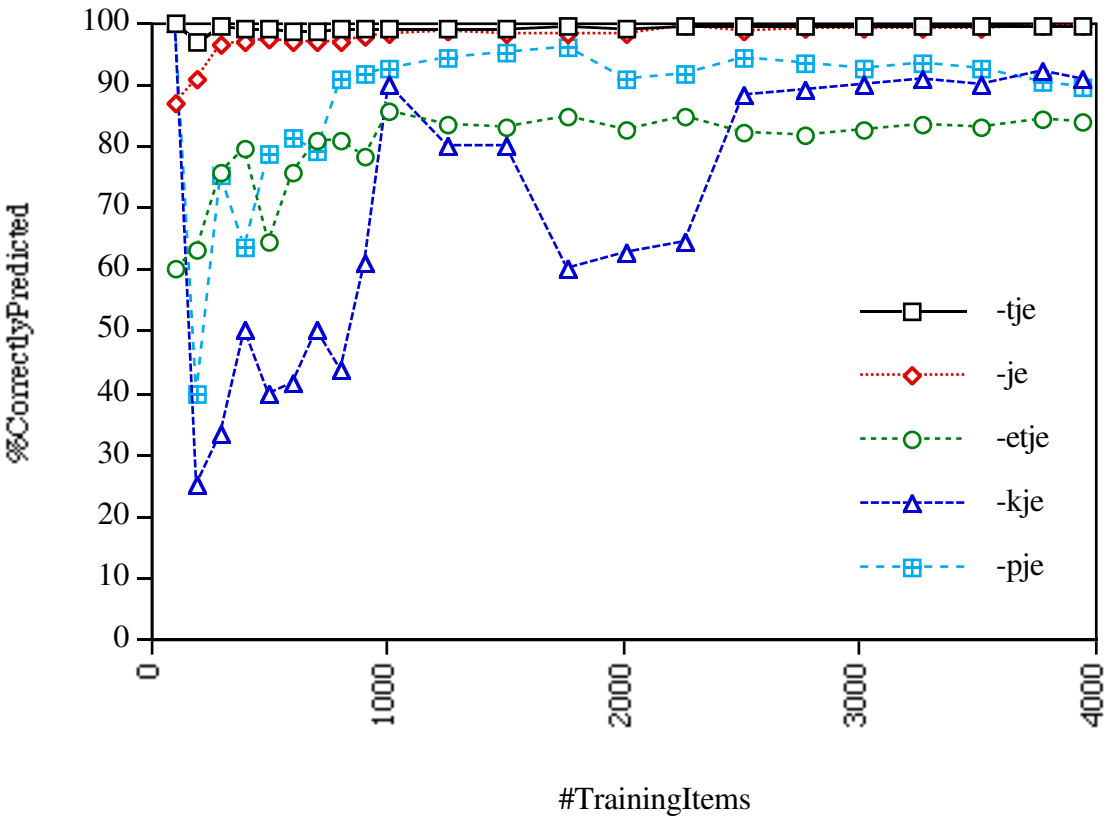

Figure 3: Learning curves for individual diminutive allomorphs.

From this graph it appears that *-tje* and *-je* are almost perfectly learned, even when the system is provided only a relatively limited set of examples. The generalization accuracy for these two allomorphs reaches 99.3% for *-tje* and 99.2% for *-je* (mean percentages over 10 experiments in the ten-fold cross-validation experiment). The other allomorphs are not so easily learned: *-kje* and *-pje* eventually reach a success score of approximately 90% and *-etje* 84%. These scores should be evaluated in the light of the relatively inferior frequency of these allomorphs in the training sets. Given the abundant presence of *-tje* and *-je* (together 85% of the entire database) the other allomorphs will not be represented by a large amount of examples in a ten-fold cross-validation experiment, especially when the training set is fairly restricted. Especially *-kje* needs a lot of examples, judging from the low generalization accuracy up to a training set of 2500 items.

So far we have provided a rationale for using data mining techniques in linguistic research. We have selected a domain, viz. diminutive formation in Dutch, and a data

mining technique, viz. the C4.5 machine learning system. In a first experiment we have shown that the problem is indeed learnable by the learning method selected. In the following sections we will concentrate on qualitative aspects of the induction of linguistic knowledge using the data mining technique: in the next section we describe an experiment that was set up to focus on the evaluation of conflicting hypotheses. Next we will analyze in qualitative terms the domain knowledge (rules and categories for diminutive formation) that the system discovers.

5. Experiment 2: Linguistic Hypothesis Testing and Linguistic Discovery

In the previous experiment we concentrated on the success rates of the algorithm (Experiment 1). In the second set of experiments we will examine the linguistic knowledge induced by C4.5 from the perspective of linguistic hypothesis testing and linguistic discovery. As to the former, we will investigate more specifically the hypothesis formulated by Trommelen (1983) that only the rhyme of the final syllable of the noun is relevant for determining the diminutive allomorph. As to the latter, we will analyze the rules induced by the system in order to find out if linguistically relevant generalizations appear. In other words, we will analyze the linguistic knowledge C4.5 abstracted from the learning material.

5.1 Method

In section 2 the regularities of diminutive formation in Dutch were briefly introduced. According to Trommelen's (1983) analysis, the correct diminutive allomorph can be determined solely on the basis of the rhyme of a noun's final syllable. This entails a number of interesting testable hypotheses:

1. Only information about the last syllable is relevant for determining the correct allomorph, and, hence, adding information about other syllables should not lead to superior performance on the task.
2. Only information about the rhyme of the last syllable is relevant, hence information about the onset of the final syllable is irrelevant in predicting the correct allomorph.
3. Stress is irrelevant for predicting the correct allomorph. Hence, adding information about a word's main stress should not lead to a better prediction of the correct diminutive allomorph.

In order to test these hypotheses learning material (training and test sets) were constructed that differed in a number of crucial respects relevant to the hypotheses. C4.5 was trained and tested with different learning and test material that incorporated the following information:

1. Only the rhyme (segments in the nucleus and the coda) of the last syllable of the word was represented (NC corpus).
   This corpus thus incorporates the claim of Trommelen that only the rhyme of the last syllable is necessary.
2. The segments of the last syllable (onset, nucleus and coda) are represented (ONC corpus).
3. The segmental information of the last syllable (as in the ONC corpus) supplemented with the stress level (main stress or not). This corpus will be referred to as the SONC corpus.
4. The segmental information of the three last syllables and for each syllable its stress level (main stress or not). Examples are provided in Table 3. This corpus will be referred to as the 3SYLL corpus.

The 3950 words used in the previous experiment were also used in this experiment. The decision trees and rules were obtained in a single run of C4.5 over the entire set of 3950 words. This means that since we are mainly interested in the output representations of C4.5, the complete corpus is used as input for the system. The decision trees are used to determine the number of items in the training set that they correctly handle.

5.2 Linguistic Hypothesis Testing

In Table 4 the results are displayed of the application of C4.5 to the four corpora. The number of misclassified items are indicated in raw figures as well as in the percentage of the total number of items in the training set. Recall that C4.5 is performing here the task of a linguist who collects a number of observations, formulates rules to account for the observations and in the process of rule discovery or formulation checks how well the rules account for the data.

Table 4: Errors in diminutive formation: comparison of four encodings

		NC		ONC		SONC		3SYLL	
Suffix	Total # Items	#	%	#	%	#	%	#	%
-tje	1896	12	0.60	6	0.32	6	0.32	7	0.37
-je	1478	6	0.41	6	0.41	6	0.41	6	0.41
-etje	395	81	20.51	80	20.25	78	19.75	59	14.94
-kje	77	0	0.0	0	0.0	0	0.0	1	1.30
-pje	104	5	4.81	1	0.96	1	0.96	1	0.96
	3950	104	2.63	93	2.35	91	2.30	74	1.87

Table 4 shows the results for the different corpora. First of all, the number of errors is very small: 2.63% for the NC corpus down to 1.87% for the 3SYLL corpus. Thus, the overall best result is obtained with the corpus that contains all information about the three last syllables. This suggests that, contra Trommelen, important information is lost by restricting attention exclusively to the last syllable. The difference between the success score of the NC corpus and the 3SYLL corpus is statistically significant (χ^2 = 5.173, p < .0226). The difference between the encodings of the last syllable (NC, ONC, SONC) is very small, which corroborates Trommelen's claim that stress and onset are not necessary to predict the correct diminutive allomorph.

An analysis of the results for individual allomorphs shows that *-etje* shows an interesting pattern. Adding more information to the last syllable does not significantly change the error rate for the other allomorphs, even adding information about the three last syllables does not yield a significant improvement. But the error rate for *-etje* drops significantly (χ^2 = 4.2018, p < .0405) when information about other syllables is added in addition to information about the final one. In the second section, we pointed out that words ending in the velar nasal show an intricate pattern of alternation between *-kje* and *-etje* and that additional information such as the word's stress pattern or morphological information would be needed in order to deal in a satisfactory way with that alternation. It appears that if only information about the last syllable is available, the system formulates an overgeneral rule for *-kje*, which correctly applies to words requiring that allomorph but in addition incorrectly applies to words requiring *-etje*. In the next section the specific rules induced from the NC corpus and the 3SYLL corpus will be analyzed and the exact source of this overgeneralization will be revealed.

The learnability results discussed in this section corroborate the broad lines of Trommelen's (1983) analysis: only the segmental information of a word's final rhyme is necessary for predicting the correct diminutive allomorph. This conclusion holds for all allomorphs, except for *-etje*: adding more information (as in the 3SYLL corpus, improves the success score significantly. We will now turn to the use of C4.5 as a generator of generalizations about the domain and investigate the induced rules for diminutive formation and the linguistically relevant categories established during the induction process.

5.3 Linguistic Discovery

In this section we analyze the linguistic knowledge C4.5 has induced from the corpus of examples. First of all, a number of interesting generalizations appear from the decision tree induced from the NC corpus (segmental information of the final syllable). The decision tree is represented in Figure 4.

```
Decision Tree:

coda in {rk, k, s, t, lt, p, st ,χt, mt, f, ts, nt, χ, ns, rt, lf, ŋk, nst, ls, ft, rs, lk, rχ,
         mp, rst, lp, ks, rp, kst, b, lχ, kt}: JE
coda in { n, l, r, =, ŋ, rn, m, j, w, rm, lm}:
|  nucleus in {  ɪ, ɑ, œ, ɛ, ɔ}:
|  |  coda in { n, l, r, m}: ETJE
|  |  coda in { rn}: TJE
|  |  coda in { rm, lm}: PJE
|  |  coda = ŋ:
|  |  |  nucleus = ɪ : KJE
|  |  |  nucleus in {  ɑ, ɛ, ɔ}: ETJE
|  nucleus in {  a, e, ə, u, œy, ɛɪ, i, ø, o, y, ɑu, ɛ:, ɒ:}:
|  |  coda in { n, l, r, =, rn, j, w}: TJE
|  |  coda = m : PJE
```

Figure 4: Induced decision tree from the NC corpus.

The decision tree should be read as follows. First of all, C4.5's value grouping mechanism has created a number of phonological classes by collapsing different segments into sets indicated by curly brackets. The tree starts with a test on the coda of the last syllable: if it ends in an obstruent, the correct allomorph is *-je.* If not, check the nucleus of the last syllable. Two possibilities are indicated: either the nucleus

belongs to the set {ɪ, ɑ, œ, ɛ, ɔ}, which is the set of monomoraic (short or lax) vowels in Dutch or the nucleus belongs to the set {a, e, ə, u, œy, ɛɪ, i, ø, o, y, ɑu, ɛ:, ɒ:}, which is the set of bimoraic vowels (long or tense vowels, diphthongs, and schwa). If the vowel is monomoraic, check the coda. If the coda is /ŋ/ then *-kje* is selected if the nucleus is /ɪ/, else *-etje* is the predicted allomorph (this is where the overgeneralization to *-kje* for words in /ɪŋ/ occurs). If the coda is a liquid or a nasal (except /ŋ/) *-etje* is the correct allomorph. The clusters /rm, lm/ require *-pje*, and in the remaining cases (a short vowel followed by the cluster /rn/) *-tje* is selected. If the nucleus is bimoraic, the only relevant test is whether the coda is /m/. If so, *-pje* is selected, else *-tje* is the correct allomorph.

Interestingly, while constructing the decision tree, several phonological relevant categories are 'discovered'. As already indicated, C4.5 comes up with a division of the vowels into a bimoraic category and a monomoraic one. The proposed division completely coincides with what is generally accepted in the literature, but especially the inclusion of schwa in the set was new when Trommelen proposed it in 1983. Other relevant categories include the obstruents (in the obstruent final clusters), the category of nasals and liquids, where the velar nasal is left out, reflecting its exceptional distributional properties: /ŋ/ can only occur preceded by a short vowel.

In the previous section we noted that for words ending in /ɪŋ/ the choice between *-etje* and *-kje* could not be accurately made by only referring to the last syllable (C4.5, and any statistical induction algroithm for that matter, overgeneralizes *-kje*). When we analyze the knowledge derived by C4.5 from the full 3SYLL corpus, which contains all segmental information of the last three syllables as well as information about the stress pattern, this problem is solved. We first present the knowledge induced by the system in its rule format (which is completely equivalent to and is automatically inferred from decision trees such as the one in Figure 4).

Default class is *-tje*

Rule 1:
IF coda last in {rk, k, s, t, lt, p, st ,χt, mt, f, ts, nt, χ, ns, rt, lf, ŋk, nst, ls, ft, rs, lk, rχ, mp, rst, lp, ks, rp, kst, b, lχ, kt}
THEN *-je*

Rule 2:
IF coda last in {rm, lm}
THEN *-pje*

Rule 3:

IF nucleus last in {a, e, ə, u, œy, ɛɪ, i, ø, o, y, ɑu, ɛ:, ɒ:}
AND coda last = m
THEN -*pje*

Rule 4:
IF nucleus penultimate {=, ə}
AND nucleus last in {ɪ, ɑ, œ, ɛ, ɔ}
AND coda last in {n, l, r, ŋ, rn, m}
THEN -*etje*

Rule 5:
IF nucleus last in {I, A, }, E, O}
AND coda last in {n, l, r, m}
THEN -*etje*

Rule 6:
IF nucleus penultimate in {ɪ, ɛ, i, a, ɔ, ø, ɑ, e, o, ɛɪ, ʉ, u, ɑu, œy}
AND coda last = ŋ
THEN -*kje*

The default class -*tje* is the allomorph chosen when no other rule is triggered.

The first rule refers to the class of obstruents in the coda of the last syllable (equivalent to the first test in the decision tree in Figure 4): if the final consonant is an obstruent (or a consonant cluster ending in an obstruent) then -*je* is the correct choice. The second and the third rule deal with the allomorph -*pje* which is the correct choice if the word ends in /m/ preceded by a liquid (Rule 2) and if the nucleus of the last syllable is a bimoraic vowel (Rule 3).

Rule 4 accounts for a number of interesting phenomena in the choice for the allomorph -*etje*. The first condition refers to the nucleus of the penultimate syllable, viz. those cases where the nucleus is empty ('='), and hence the word is monosyllabic, or the nucleus is a schwa. The second condition further restricts the relevant cases to the words with a monomoraic vowel in the last syllable and the third condition adds that only words ending in a liquid or a nasal (including the cluster /rn/) are eligible. This rule accurately predicts that monosyllabic words in /r/ get -*etje* instead of the default -*tje*: *bar-etje* (/bɑrətʃə/, 'bar-DIM'). Also the opposition between monomoraic and bimoraic vowels in monosyllabic words is accurately dealt with: for instance, *bon-etje* (/bɔnətʃə/, 'ticket-DIM') versus *boon-tje* (/bo:ntʃə/, 'bean-DIM'), and *bom-etje* (/bɔmətʃə/, 'bomb-DIM') versus *boom-pje* (/bo:mpjə/, 'tree-DIM'). [2]

[2] For the sake of completeness it should be added that there is a single small category of words that

As to the overgeneralization of the allomorph *-kje*, Rule 4 opposes monosyllabic and multisyllabic words in /ɪŋ/: monosyllabic words correctly receive *-etje*, for instance *ring-etje* (/rIŋətʃə/, 'ring-DIM'). Rule 4 also filters out multisyllabic words with a schwa in the penultimate syllable, which correctly opposes word pairs such as *zoldering-etje* (/zɔldərIŋətʃə/, 'ceiling-DIM') versus *soldering-kje* (/sɔldeːrIŋkjə/, 'soldering-DIM'). Thus, Rule 4 correctly applies for monosyllabic words ending in /ɪŋ/ and longer words with a schwa in the penultimate syllable. Other words in /ɪŋ/ are correclty excluded. Note that by referring to schwa in the penultimate syllable, the rule does not use stress as the distinguishing factor between these two types of words.

The remaining words in /ŋ/ are dealt with by Rule 6. Briefly stated it stipulates that if the nucleus of the penultimate syllable is not empty and is not a shwa, or in other words if it is a full vowel or a diphthong, *-kje* is the correct choice. *Koning-kje* (/koːnɪŋkjə/, 'king-DIM') and *teerling-kje* /teːrlɪŋkjə/, 'die-DIM') receive the correct allomorph. But note that *leerling* (/leːrlɪŋ/, 'pupil-DIM') and *tweeling* (/tweːlɪŋ/, 'twin-DIM') are not satisfactorily dealt with: they should get *-etje* instead of the predicted *-kje*. These words require additional (morphological) information in order to distinguish the monomorphemic *koning* and the plurimorphemic *tweeling*.

Finally, Rule 5 deals with words ending in a short vowel followed by a nasal (except /ŋ/) or a liquid. It correctly predicts the allomorphic variation in pairs such as *man-etje* (/mɑnətʃə/, 'man-DIM') versus *maan-tje* (/maːntʃə/, 'moon-DIM'). Note that C4.5 misses the generalization that polysyllabic words in /r/ preceded by a short vowel (*radar,* /raːdɑr/, *sonar* /soːnɑr/, etc.) should be assigned *-tje* instead of *-etje*. These words, which are very infrequent, are successfully blocked from Rule 4, but are not blocked from Rule 5, which in a high majority of cases makes the correct prediction.

An interesting fact about this rule set is that it makes use of only three features represented in the training material: the coda and the nucleus of the last syllable and the nucleus of the penultimate syllable. This means that contrary to our description in section 2 where some regularities were formulated in terms of the stress patterns of words, all regularities can be formulated in terms of segmental content without

---

this rule will treat incorrectly, viz. words monosyllabic words ending in a short vowel followed by /rn/. There are only three monosyllabic words in our corpus of 3950 words that meet this condition. For these words *-etje* is incorrectly predicted instead of *-tje*. These words receive the correct allomorph if the decision tree in Figure 4 is used.

reference to suprasegmental features. Secondly, contrary to the hypothesis formulated by Trommelen (1983), apart from the rhyme of the last syllable, the nucleus of the penultimate syllable is relevant as well.

When we compare the decision tree based on the NC corpus with the rule set induced from the 3SYLL corpus, it appears that they largely overlap. However, the NC corpus only allowed a global differentiation between *-kje* and *-etje* in terms of words ending in a velar nasal: /ɪŋ/ words were all categorized as receiving *-kje* and all others received *-etje*. This overgeneralization is taken care of by the rule set from the 3SYLL corpus: Rule 4 differentiates monosyllabic words from polysyllabic ones. The former get *-etje*. In addition polysyllabic words with schwa as the nucleus of the penultimate syllable are also correctly assigned *-etje* by Rule 4. Rule 6 adds the correct condition for *-kje*: a long vowel or diphthong as the nucleus of the penultimate syllable is decisive in this respect. C4.5 induces a quite elegant set of rules: Rule 4 treats the sequence of a short vowel followed by a nasal or a liquid as one single case, specifying a unified condition for determining the correct allomorphy. Rule 6 handles the exceptional status of the velar nasal: *-kje* is the correct allomorph in specific circumstances involving only the velar nasal. As far as we know a similar analysis of the *-kje/-etje* alternation has not been proposed in the literature on diminutive formation.

6. Conclusion

We have shown by example that data mining techniques can profitably be used in linguistics as a tool for comparing linguistic theories and hypotheses or for the discovery of new linguistic theories in the form of linguistic rules or categories.

The case we presented concerns diminutive formation in Dutch, a relatively simple and well documented domain, which nevertheless shows enough complexity to reveal the power and usefulness of the proposed techniques. We showed that data mining techniques can be used to corroborate and falsify some of the existing theories about the phenomenon. In particular, by using corpora annotated using the concepts deemed necessary by the theoretical accounts, and submitting the corpora to a machine learning procedure, the necessity of the concepts for solving the linguistic task was assessed. We showed that although Trommelen's (1983) claim that only the rhyme of a word's final syllable is necessary for predicting the correct allomorph, holds for four of the five diminutive allomorphs. No substantial changes of the error rate was noted using more information about the final syllable of words. However, we also

found that when information about other syllables was added to the corpus, the error rate decreased significantly, which indicates that Trommelen's thesis was only partially corroborated. More specifically, the allomorph *-etje* can only be accurately predicted when the nucleus of the penultimate syllable is known.

Secondly, we showed that using data mining techniques, linguistically interesting generalizations can be discovered. In learning the rules of diminutive formation, categories were extensionally defined which constitute interesting generalizations from a linguistic point of view. In particular, the set of bimoraic vowels, obstruents, etc. surfaced. The rules induced by the algorithm were shown to be highly similar to those proposed in linguistic analyses, and C4.5 even induced rules for the *-kje / -etje* alternation that appear to throw an innovative light on that issue.

In conclusion: this case study dealt with a relatively small and well documented linguistic domain, which provided the opportunity to show the strength of data mining techniques as a method in linguistic analysis: (i) linguistic analyses can be corroborated (and falsified) in a straightforward way; (ii) the categories established by linguists were also induced by the machine learner, holding promises for future work on less well-studied and more complex domains; and (iii) the machine learner proposed new generalizations (rules) within the domain.

References

Booij, G. & Van Santen, A. 1995. *Morfologie: De woordstructuur van het Nederlands.* Amsterdam: Amsterdam University Press.

Booij, G. 1984. Syllabestructuur en verkleinwoordvorming in het Nederlands. *Glot* 7: 207-226.

Burnage, G. 1990. *CELEX: A guide for users.* Nijmegen: Centre for Lexical Information.

Cohen. A. 1958. Het Nederlands diminutiefsuffix, een morfonologische proeve. *De Nieuwe Taalgids* 51: 40-45.

De Haas, W. & Trommelen, M. 1993. *Morfologisch handboek van het Nederlands. Een overzicht van de woordvorming.* 's-Gravenhage, SDU Uitgeverij.

Haverkamp-Lubbers, R. & Kooij, J. eds. 1971. *Het verkleinwoord in het Nederlands.* Amsterdam: University of Amsterdam. Publikaties van het Instituut voor Algemene Taalwetenschap, 1.

Kruisinga, E. 1915. De vorm van de verkleinwoorden. *De Nieuwe Taalgids* 9: 96-97.

Piatetsky-Shapiro, G. & Frawley, W. 1991. *Knowledge discovery in databases.* AAAI Press.

Quinlan, J. 1993. *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann.

Te Winkel, L. 1862. Over de verkleinwoorden. *De Taalgids* 4: 81-116.

Trommelen, M. 1983. *The syllable in Dutch*. Dordrecht: Foris.

Weiss, S. & Kulikowski, C. 1991. *Computer systems that learn.* San Mateo: Morgan Kaufmann.