6

Discovery of association rules between syntactic variables

Data mining the Syntactic Atlas of the Dutch Dialects

Marco René Spruit
Meertens Instituut

Abstract

This research applies an association rule mining technique to purely syntactic dialect data. The paper answers the research question of how relevant associations between syntactic variables can be discovered. The method calculates the proportional overlap between geographical distributions of syntactic microvariables and incorporates rule quality factors such as accuracy, coverage and completeness to measure the interestingness of the variable associations. The exploratory review of the results discusses several highly ranked association rules and also examines an implicational chain of syntactic variables.

6.1 Introduction

This work[1] investigates a data mining technique to discover associations between syntactic variables in Dutch dialects using a rule induction system based on proportional overlap. The research aims to contribute to the understanding of the as-

[1]The research for this paper is being carried out in the context of the NWO project The Determinants of Dialectal Variation, number 360-70-120, P.I. J. Nerbonne. Please visit http://dialectometry.net for more information and relevant software.

Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands
Edited by: Peter Dirix, Ineke Schuurman, Vincent Vandeghinste, and Frank Van Eynde.

sociations between syntactic variables by examining geographical distributions of syntactic microvariation. The current paper addresses the following two research questions:

1. How can relevant associations between syntactic variables be discovered?
2. What are interesting associations between syntactic variables?

This research integrates expertise from the research fields of data mining and ecology to answer these questions quantitatively. In essence this investigation exhaustively evaluates levels of association between combinations of syntactic variables based on the proportional overlap between their geographical distributions.

This work proceeds from the observation that linguistic research frameworks such as generative syntax and functional typology share a primary interest in understanding the structural similarities and differences between language varieties. The frameworks aim to identify which universal syntactic properties can vary across language varieties and which remain constant. The ultimate goal is to characterise the superficial structural diversity of all language varieties as particular settings of relatively few parametric patterns. Unfortunately, the search for syntactic universals is still very much a topic of ongoing research. Gianollo et al. (to appear) most notably define an extensive parametric framework to model language variation in the internal structure of Determiner Phrases based on a relatively wide sample of languages and language families.

Haspelmath (to appear) compiles a list of seven universal syntactic parameters for which there is a wide consensus in the field. One well-known example of a syntactic universal is the pro-drop/null-subject parameter, which states that the subject position in a clause may be empty or must be filled by a subject pronoun. It was originally thought to universally correlate with syntactic phenomena such as null thematic subjects and null expletives (Rizzi 1986). However, the generalisation quickly became untenable once more language varieties were analysed (Newmeyer 2005). This example adequately illustrates that a large data set of comparable language varieties is required to investigate syntactic variable relationships more reliably. Such an examination needs to be automated using verifiable methods because of the exhaustive and repetitive nature of the comparison procedure.

The current research aims to contribute to the global research effort of parametrisation of the structural diversity of language varieties by proposing a computational method to discover syntactic variable associations automatically. The technique facilitates exploration of previously unknown variable relationships and validation of existing parametric generalisations. The second research question is addressed through an exploratory review of the method's application to a large syntactic microvariation database.

The paper is structured as follows. Section 2 describes the unique syntactic variation database under investigation. Section 3 introduces the sample data subset used in Section 4 to illustrate the association rule mining procedure based on proportional overlap. Section 5 reviews the evaluation factors to accurately measure the quality of the association rules. Section 6 explores the most interesting rules

discovered in the sample data. Section 7 highlights results of the association rule mining application to the entire syntactic variation database under investigation. Section 8 recapitulates the main findings. The paper concludes with a discussion and directions for future research in Section 9.

6.2 Syntactic variation database

This research examines the first volume of the *Syntactische Atlas van de Nederlandse Dialecten* (SAND1; 'Syntactic Atlas of the Dutch Dialects'; Barbiers et al. (2005)) from a quantitative perspective. SAND1 contains 145 geographical distribution maps of individual syntactic variables in 267 Dutch dialects in the Netherlands, the Northern part of Belgium and a small northwestern part of France.[2] It covers syntactic variation related to the left periphery of the clause and pronominal reference. This includes variation with respect to complementisers, subject pronouns and expletives, subject doubling and subject clitisation following yes/no, reflexive and reciprocal pronouns, and fronting phenomena. The second and final volume of the SAND is due to appear near the end of 2007 and will describe syntactic variation in Dutch dialects with respect to verbal clusters, negation and quantification. Cornips and Jongenburger (2001) review the methodological aspects of the written and oral syntactic elicitation techniques which were employed to reliably collect the SAND data.

From a quantitative research perspective SAND1 also represents a syntactic microvariation database containing 106 syntactic contexts and 485 syntactic variables among varieties of a single language. This work defines a syntactic variable as a form or word order in a syntactic context in which two dialects can differ (Spruit 2006). The number of available syntactic contexts is somewhat lower than the number of geographical maps because SAND1 also contains numerous correlation maps which show syntactic variables from different perspectives. Also, some syntactic contexts are presented using multiple maps.

Tables 1 to 4 provide examples of syntactic variation in the complementisers, subject doubling, reflexives and fronting domains, respectively. For example, Table 1 shows the attested variation throughout the Dutch language area in the realisation of the complementiser position in comparative if-clauses as presented in SAND1 map B on page 14. In standard Dutch people say *'t lijkt wel of er iemand in de tuin staat* 'it looks [affirmative] if there someone in the garden stands', but in colloquial Dutch the following form also frequently occurs in the southern provinces: *'t lijkt wel of dat er iemand in de tuin staat*. There are even a few northern and southern regions within the Dutch language area where the verb occurs in the second position of the if-clause: *'t lijkt wel of er staat iemand in de tuin*. The last example also illustrates that both word form and word order may vary within a syntactic context.

[2]The online version of this paper at http://marco.info/pro/pub/mrs2007clin.pdf includes geographical distribution maps of the SAND dialect locations with relevant province names and also contains additional data mining results and references.

Table 1. Map 14b in SAND1 shows seven syntactic variables in the complementisers domain.

Context	Complementiser of comparative if-clause											
Variables	{ of, *of dat, dat, as/of + V2, at, as, et }											
Example	't	lijkt	wel	of	dat	er	∅	iemand	in	de	tuin	staat.
Gloss	it	looks	[affirmative]	if	that	there	[v2]	someone	in	the	garden	stands
Translation	"It looks as if there is someone in the garden."											

Table 2. Map 54a in SAND1 shows four syntactic variables in the subject doubling domain.

Context	Subject doubling 2 singular											
Variables	{ V_{FINITE} __ , *__ V_{FINITE} __ , C __ , *$C_{COMPARATIVE}$ __ }											
Example	Ge	gelooft	gij	zeker	niet	dat	hij	sterker	is	as	-ge	gij.
Gloss	you_{weak}	believe	you_{strong}	certainly	not	that	he	stronger	is	than	you_{weak}	you_{strong}
Translation	"You do not seem to believe that he is stronger than you."											

Table 3. Map 68a in SAND1 shows five syntactic variables in the reflexives domain.

Context	Weak reflexive pronoun as object of inherent reflexive verb						
Variables	{ zich, hem, *zijn eigen, zichzelf, hemzelf }						
Example	Jan	herinnert	zijn	eigen	dat	verhaal	wel.
Gloss	John	remembers	his	own	that	story	[affirmative]
Translation	"John certainly remembers that story."						

Table 4. Map 84a in SAND1 shows four syntactic variables in the fronting domain.

Context	Short subject relative, complementiser following relative pronoun									
Variables	{ *1:die 2:as/at/da(t), 1:die 2:-t, 1:dien 2:at/da(t), 1:die/dat 2:wat }									
Example	Dat	is	de	man	die	dat	het	verhaal	verteld	heeft.
Gloss	that	is	the	man	who	that	the	story	told	has
Translation	"That is the man who told the story."									

6.3 Sample data illustration and diagram

Figures 1 and 2 illustrate the data mining procedure presented in the next section by defining a small subset of the actual SAND1 data. Figure 1 marks the geographical occurrences in seven Dutch dialects (1-7) of the four example variables (A-D) shown in Tables 1 to 4. For example, Figure 1 shows that in the dialects of Ouddorp (1), Merckeghem (2), Brussel (3) and Gemert (4), people can say *'t Lijkt wel of dat er iemand in de tuin staat* (A). This variable does not occur in the dialects of Nieuwmoer (5), Boskoop (6) and Nijkerk (7). Likewise, only in the village of Nieuwmoer have all of the following three variables been attested: *Als gij gezond leeft, leef-de gij langer* (B), *Jan herinnert z'n eigen dat verhaal wel* (C), and *Dat is de man die dat het verhaal verteld heeft* (D). Figure 2 shows a symbolic representation of the sample data in Figure 1. The remainder of the current article uses the symbolic variable characters (A-D) and dialect numbers (1-7) to refer to the sample data components to enhance readability.

6.4 Association rule mining based on proportional overlap

The SAND1 sample data described above are used to illustrate how relationships between variables in a database can be discovered using a technique best known as data mining but arguably more accurately described with its synonym Knowledge Discovery in Databases (KDD). Data mining is an umbrella term for various knowledge representation techniques such as association *rules*, decision *trees* and neural *networks*. Frawley et al. (1992) define data mining as the nontrivial ex-

Figure 1. This SAND1 sample marks the occurrences in seven dialects (1-7) of the syntactic variables (A-D) in Tables 1 to 4.

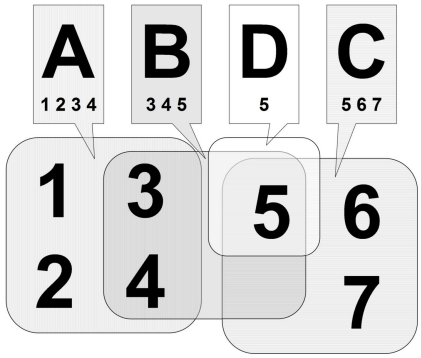

Figure 2. Symbolic representation of the SAND1 sample shown in Figure 1.

traction of implicit, previously unknown, and potentially useful information from data. Hand et al. (2001) formulate data mining more generally as the science of extracting useful information from large data sets or databases.

This work explores associations between syntactic variables in Dutch dialects using a rule induction system based on proportional overlap. Generally speaking, association rules show attribute-value conditions that occur frequently together in a given dataset. The left side of an association rule is called the antecedent and may consist of multiple predicting attributes. The right side of a rule is called the consequent and defines the predicted class(es). Association rules are typically written as 'A → C' and should be read as 'IF variable A THEN variable C'. A widely-used example of association rule mining is Market Basket Analysis, a method which examines a long list of supermarket transactions to determine which items are most frequently purchased together. It applies the Apriori algorithm to generate candidate association rules which relate the items within each transaction or basket (Agrawal et al. 1993).

$$C_n^k = C_4^3 = \binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4\times3\times2\times1}{3\times2\times1\times(1)} = \frac{24}{6} = 4$$

Figure 3. Calculation of the number of combinations with k=3 elements from the sample data set with n=4 variables.

The application of association rule mining between syntactic variables in the current paper examines all k-combinations (or k-subsets) of syntactic variables to determine which variable subsets most frequently co-occur geographically.[3] A k-combination is an unordered collection with k unique elements.[4] Figure 3 il-

[3]The Rule INduction Console (*rinc*) programme implements the association rule mining procedure. It has been developed with the wxWidgets C++ toolkit and the next_combination STL template. The console programme is available for all software platforms and can be downloaded from http://dialectometry.net/syntax.

[4]This is in contrast with a k-permutation, which is an *ordered* collection with k unique elements.

lustrates how to calculate the binomial coefficient of the number of combinations with three elements in the sample data set of the four variables {A,B,C,D}. In this example the binomial coefficient is four and represents the combinations {A,B,C}, {A,B,D}, {A,C,D} and {B,C,D}.

Table 5. Algorithm to non-recursively evaluate all association rules.

```
1.  FOR EACH k-combination of variable set v with n elements
2.      INITIALISE combination subset s from v
3.      REPEAT
4.          FOR EACH m-combination of s
5.              INITIALISE antecedent a from s with m elements
6.              REPEAT
7.                  INITIALISE consequent c as the complement of a with k-m elements
8.                  CALL evaluateAssociationRule with a and c
9.              UNTIL all antecedent combinations a have been processed
10.         ENDFOR
11.     UNTIL all combination subsets s have been processed
12. ENDFOR
```

Table 5 lists the association rule mining algorithm in pseudocode. The procedure is scalable to even larger data sets because it is non-recursive. Therefore, memory usage remains constant. Line 1 specifies that the procedure iterates through all combinations with k=2 to k=n variables. Line 2 selects the first combination subset s with k variables. Then, lines 3 to 11 repeatedly process subset s and select the next subset. Line 4 iterates through all combinations of subset s with m=1 to m=k-1 variables. Line 5 generates the first combination subset a as the antecedent variables subset from s with m variables. Then, lines 6 to 9 repeatedly process subset a and select the next subset. Line 7 determines the corresponding consequent variables by selecting the complementary set of a from s. Finally, line 8 evaluates the quality of the generated association rule using the unique antecedent-consequent tuple based on the proportional overlap between the geographical distributions of the rule variables. The candidate association rule is accepted when it satisfies previously specified criteria of interestingness.

The procedure remains modest in automatically discarding uninteresting candidate rules. The current version of the algorithm only prunes the combination space in two cases. In the first, self-explanatory situation the interestingness value is either equal to or below zero. The second condition applies when the coverage value has the maximum value. This indicates that the antecedent encompasses the entire data set, which implies that the rule does not have any explanatory power. Of course, manual factor threshold values may be applied as well in addition to these conditions to further minimise the amount of uninteresting rules.

The proportional overlap procedure in this work consists of the following three steps. First, the lists of geographical occurrences of all syntactic variables in the rule antecedent are disjunctively merged into the rule antecedent vector of geographical occurrences. Variable occurrences are not merged conjunctively because the procedure attempts to combine microvariables to discover more general patterns. Then, the procedure constructs the rule consequent vector of geographical

occurrences. Finally, the intersection and union sets of the two vectors of geographical co-occurrences are calculated as factor components to help determine the quality of the candidate rule using a combination of indicators as listed in Table 6. The intersection set |A&C| in Table 6 represents the geographical conjunction of antecedent and consequent variable occurrences. The concept of proportional overlap is predominantly applied in research areas such as ecology and biogeography and is notably explored in (Horn 1966).

6.5 Evaluating the quality of a rule

Table 6 lists several widely used factors to help determine the quality of an association rule: accuracy, coverage, completeness and interestingness. Many more factors have been proposed over the years to further enhance rule evaluation quality. McGarry (2005) reviews a range of objective and subjective measures such as actionability, surprisingness, unexpectedness, misclassification cost, class distribution and attribute ranking, among others. These factors are not taken into account in this work. However, the current paper does incorporate complexity as the total number of variable disjuncts in both the antecedent and consequent sets. Higher complexity results are interpreted as being less interesting.

Table 6. Evaluation factors to help determine the quality of association rule A→C.

Accuracy:	|A&C| / |A|	The number of dialects which have both variables A and C divided by the number of dialects which have variable A.
Coverage:	|A| / N	The number of dialects which have variable A divided by the total number of dialects in the data set.
Completeness:	|A&C| / |C|	The number of dialects which have both variables A and C divided by the number of dialects which have variable C.
Interestingness:	|A&C| - |A||C|/N	The number of dialects which have both variables A and C minus the product of the number of dialects which have variable A with the number of dialects which have variable C divided by the total number of dialects in the data set.

It is important to note that although a pattern is expressed as a rule, it does not mean that it is true all the time. An association rule does not imply causality. The antecedent of a rule does not necessarily cause the consequent of a rule to happen. Therefore, the uncertainty in a rule should be made explicit. This is what the accuracy of a rule indicates. It signifies how often a rule is correct and is also called the confidence of a rule. The coverage of a rule expresses how often a rule applies and is also called support. The factor completeness may be used to explore how much of the target class a rule covers. This work multiplies all accuracy, coverage and completeness values by one hundred to express the rule quality factors as percentages.

The three rudimentary interestingness factors described above are always integrated in proposed measures of rule interestingness. Intuitively, rules are interesting when they have high accuracy, high coverage and deviate from the norm. The effort, then, is to formulate the optimal trade off between coverage, accuracy and potentially other factors for a specific problem domain. The domain specificity of interestingness is one of the many reasons why the ability to interactively explore

the generated association rules is always desirable and maybe even inevitable. Although data mining algorithms may use objective factors to decide whether a rule is genuinely interesting or not, domain-specific, subjective notions of interestingness may be required as well to decide whether a potentially or technically interesting rule is also genuinely interesting in a specific domain. For example, a discovered association rule may be too well-known or too trivial.

Table 7. Piatetsky-Shapiro's principles for rule interestingness (RI) measures.

1.	RI = 0 if |A&C| = |A| |C| / N.
2.	RI monotonically increases with |A&C| when other parameters are fixed.
3.	RI monotonically decreases with |A| or |C| when other parameters are fixed.

This work applies the three principles for rule interestingness measures proposed in (Piatetsky-Shapiro 1991). They are reprinted in Table 7. The principles formulate the relations between the factors accuracy, coverage and completeness as objective evaluation criteria of interestingness measures. The first principle states that the rule interestingness is zero if the antecedent and consequent of the rule are statistically independent. The second principle defines that more co-occurring elements in the antecedent and consequent of the rule will result in higher accuracy and completeness values when all other parameters remain fixed, which increases the interestingness of the rule. The third principle's interpretation is two-fold. It formulates that rule interestingness monotonically decreases with completeness when all other parameters remain fixed. Similarly, rule interestingness also monotonically decreases with coverage when all other parameters remain fixed (Freitas 1999). Note that, in contrast with accuracy, coverage and completeness values, interestingness values do not necessarily range between zero and one.

Several enhancements and alternative measures of interestingness have been proposed since (Piatetsky-Shapiro 1991). Lenca et al. (to appear) most notably describes numerous measures of interestingness in detail. The current work restricts itself to Piatetsky-Shapiro's measure of interestingness because of its historical position and formulaic simplicity. Note, however, that its symmetric nature is a property where this measure seems lacking. This is not the case for the factors accuracy, coverage and completeness. To a certain extent the influence of symmetricity can be compensated by ranking the entire result set of association rules firstly on descending interestingness, secondly on ascending complexity, thirdly on descending accuracy and finally on descending coverage.

6.6 Discovery of association rules between syntactic variables

Table 8 lists the eight most interesting association rules based on occurrences in seven dialects of the four syntactic variables in the sample data as shown in Figures 1 and 2. The algorithm in Table 5 generates fifty variable combinations for the sample data. Fourteen candidate rules are potentially interesting based on the Piatetsky-Shapiro measure of interestingness and have at least some explanatory power. From a technical perspective this means that fourteen association rules

have an interestingness value greater than 0 and a coverage value smaller than 100 percent. The list in Table 8 is sorted on descending interestingness, ascending complexity and descending accuracy, respectively.[5]

Table 8. The eight most interesting association rules in the sample data set as shown in Figures 3 and 4 sorted on descending interestingness, ascending complexity and descending accuracy.

#	*Antecedent*	→	*Consequent*	*Interestingness*	*Complexity*	*Accuracy %*	*Coverage %*	*Completeness %*
1.	B	→	A ∨ D	0.86	1	100	42	60
2.	A ∨ D	→	B	0.86	1	60	71	100
3.	D	→	B	0.57	0	100	14	33
4.	D	→	C	0.57	0	100	14	33
5.	B	→	D	0.57	0	33	42	100
6.	C	→	D	0.57	0	33	42	100
7.	B	→	A	0.29	0	66	42	50
8.	A	→	B	0.29	0	50	57	66

The list of association rules is primarily sorted on descending interestingness since the main goal of this work is to discover the most interesting association rules between the variables. The list's secondary sort factor uses ascending values of complexity which can be interpreted as an extension of the measure of interestingness. An increasing number of variable components in a rule decrease its comprehensibility and, therefore, its interestingness. Coincidentally, the application of the complexity factor in the sample data does not actually change the rule order. The list of association rules in Table 8 is ternarily sorted on descending accuracy. However, it would be equally valid to apply descending completeness as an alternative ternary sort factor. Favouring accuracy over completeness simply signifies that it is considered more important that a rule is correct than it is to discover the degree to which the consequent variables are predicted by the antecedent variables. The definitions of accuracy and completeness in Table 6 also illustrate these alternate perspectives on rule importance quite evidently. The first two rules in Table 8 demonstrate the effect of choosing completeness over accuracy to optimally sort the association rules. The rules have identical levels of interestingness and complexity but differ in the degree of accuracy and completeness. The first rule states that *if* variable B occurs in a dialect *then* variable A or D always occur as well; the rule is 100 percent accurate. However, it does not imply that the inverse is true as well. Indeed, in dialects one and two either variable A or D occurs but not variable B. This is specified in the second rule which states that if either variable A or D occurs in a dialect, then there is a 60 percent certainty that variable B occurs as well. This example adequately illustrates the asymmetric nature of the relationship between the antecedent variables and the consequent variables of an association rule. Furthermore, an asymmetric variable association may be interpreted as a variable dependency with potentially hierarchical implications.

[5]The list of potentially interesting association rules can be sorted interactively using an external software programme such as Excel or SPSS.

6.7 Data mining the Syntactic Atlas of the Dutch Dialects

The following pages highlight a small selection of potentially interesting association rules between the 485 syntactic variables in the SAND1 database based on their geographical co-occurrences in 267 Dutch dialects. The algorithm evaluated 234,740 rules without any variable disjunctions, i.e. all antecedents and consequents consist of only one variable, and found 10,730 interesting associations with an accuracy value of 90 percent or higher. This observation manifests the considerable proportional overlap between the syntactic variables in SAND1. Additionally, it could arguably be interpreted as an indication that highly interesting association rules with high coverage and high accuracy values effectively reduce the importance of the geographical occurrences in the data set. The information value of geography—by definition—becomes limited to generic density and distributional information when variable distributions overlap nearly perfectly. Ascending from the observational level of geographical distributions to more abstract variable associations would facilitate syntactic analyses to identify implicational chains and other association patterns.

Table 9. Example of a highly ranked association rule in SAND1 with one variable disjunct: "if either antecedent variable A1 or A2 occurs, then it is certain that the consequent variable also occurs".

Antecedent A1:	p46b:julle(n)/jullie (Subject pronouns 2 plural, strong forms, complex) We geloven dat julle(n)/jullie niet zo slim zijn als wij. we believe that you$_{plural,strong}$ not so smart are as we. 'We believe that you are not as smart as we are.'
Antecedent A2:	p46b:julder/jielder (Subject pronouns 2 plural, strong forms, complex) We geloven dat julder/jielder niet zo slim zijn als wij. we believe that you$_{plural,strong}$ not so smart are as we. 'We believe that you are not as smart as we are.'
Consequent:	p46a:j-[lieden-compositum] (Subject pronouns 2 plural, strong forms) We geloven dat j-lieden niet zo slim zijn als wij. we believe that you$_{plural,strong}$ not so smart are as we. 'We believe that you are not as smart as we are.'
Statistics:	Rank=9, Combination=5,327,848, Interestingness=61.31, Accuracy=100%, Coverage=40%, Completeness=93%, Complexity=1, A-Locations=107, C-Locations=114, AC-Overlap=107, AC-Disjunction=114
Interpretation:	The infrequent pronoun 'julder/jielder' perfects the implicational association of the frequent 'julle(n)/jullie' variant with the pronoun 'j-lieden'.

The number of variable combinations rises to 113,614,160 candidate rules as soon as either the antecedent or consequent of a rule may include one variable disjunction. No less than 56,267,729 generated association rules are at least 90 percent accurate.[6] This is to be expected since the algorithm disjunctively combines variables. Once a strong association between two variables has been found, any disjunctively added variable will further strengthen the association.

Table 9 presents an association rule with one variable disjunction as an example of a potentially interesting rule with a higher complexity. However, higher complexity association rules become exceedingly more difficult to interpret linguistically. As a matter of fact, it can already be quite challenging to linguis-

[6]The corresponding output file is 33 GB. The programme execution time was around 18 hours on a MacMini PowerPC G4 (1.5 GHz) computer.

tically interpret rules without variable disjunctions. Interactive explorations can only partly facilitate the evaluation process. Therefore, the remainder of the current paper concentrates on association rules without variable disjunctions.

Table 10. The most interesting rule in SAND1 without variable disjuncts.

Antecedent:	p46a:g-lieden (Subject pronouns 2 plural, strong forms) We geloven dat g-lieden niet zo slim zijn als wij. we believe that you$_{plural,strong}$ not so smart are as we. 'We believe that you are not as smart as we are.'
Consequent:	p38b:gij/gie (Subject pronouns 2 singular, strong forms) Ze gelooft dat gij/gie eerder thuis bent dan ik. she believes that you$_{singular,strong}$ earlier home are than I 'She thinks that you'll be home sooner than me.'
Statistics:	Rank=1, Combination=10,321, Interestingness=58.38, Accuracy=99%, Coverage=39%, Completeness=89%, Complexity=0, A-Locations=105, C-Locations=116, AC-Overlap=104, AC-Disjunction=117
Interpretation:	The plural pronoun 'g-lieden' belongs to the same paradigm as the singular pronoun 'gij'.

Table 10 shows the potentially most interesting association rule in SAND1 without variable disjunctions. The rule associates one of the variables in map A on page 46 in SAND1 with a variable in map B on page 38. It states that, in the context of a strong *plural* subject pronoun in second person, if the complex pronoun 'g-lieden' occurs, then the strong *singular* subject pronoun in second person 'gij' (or 'gie') nearly always occurs as well. This is indicated by the accuracy value of 99 percent. This value is calculated using the definition in Table 6 as follows: $|A\&C| / |A| * 100$ = AC-Overlap / A-Locations * 100 = 104 / 105 * 100 = 0.99 * 100 = 99 percent. Similarly, the interestingness value results as follows: $|A\&C| - |A||C|/N$ = AC-Overlap - (A-Locations * C-Locations / 267) = 104 - (105 * 116 / 267) = 104 - 45.62 = 58.38.

The geographical distributions of the rule variables in Table 10 are patterned quite coherently (not shown). All occurrences are found in the southern half of the Dutch language area. Although it may not be particularly surprising to discover a strong association between two typically southern word forms, it does not automatically follow that it may not be considered interesting or even significant to discover that the geographical overlap between, specifically, these two southern word forms is nearly all-inclusive. It is sufficient to interactively sort all association rules on antecedent name, descending interestingness and descending accuracy, respectively, to verify this hypothesis. This action reveals that only nine potentially interesting association rules exist with the complex pronoun 'g-lieden' as their antecedent and which also have an accuracy of 90 percent or higher.

The top six 'g-lieden' rules state that if in a dialect people can say *We geloven dat g-lieden niet zo slim zijn als wij* 'we believe that you$_{strong}$ not so smart are as we', then people in that dialect can also say, in descending degree of certainty, (a) *Ze gelooft dat gij/gie eerder thuis bent dan ik* 'she believes that you earlier home are than I', (b) *Ik denk da Marie hem zal moeten roepen* 'I think that Mary him will must call', (c) *U [niet-beleefdheidsvorm] gelooft dat Lisa even mooi is als Anna* 'you [non-honorific] believe that Lisa as beautiful is as Anna', (d) *Fons zag een slang naast hem* 'Fons saw a snake next to him', (e) *Erik liet mij voor hem*

werken ‘Erik let me for him work’ and (f) *De jongen wie/die z’n moeder gisteren hertrouwd is* ‘the boy who/that his mother yesterday remarried is’.

Rules (d) and (e) also strongly indicate a relationship between the second person, plural complex pronoun ‘g-lieden’ and the third person, singular, reflexive pronoun ‘hem’. It is unclear how this association should be interpreted linguistically. Although the rules might describe a previously unknown linguistic relationship, it could also merely reflect that the variables are geographically clustered. The latter case would signify the methodological reminder that a strong variable association does not necessarily imply a linguistic causation. All in all, the analysis above adequately illustrates how exploration of one association rule may easily trigger interactive investigations of several more potentially interesting rules and may raise new questions to answer.

Another approach of interactively exploring the result set of rules focuses on the examination of implicational chains between syntactic variables. Table 11 lists the highest ranked implicational chain of four syntactic variables in the set of association rules without variable disjunctions to illustrate this phenomenon. First, rule six states that if subject doubling occurs after V in second person singular, then it also appears after V in second person plural. Second, the third highest rule asserts that if subject doubling occurs after V in second person plural, then the second person plural pronoun ‘g-lieden’ nearly always arises as well. As an aside, this rule effectively demonstrates the implicit capacity to discover variable associations across syntactic domains. Third, the highest ranked rule convincingly associates the second person plural pronoun ‘g-lieden’ with the second person singular pronoun ‘gij/gie’. Finally, rule eight confirms the transitive nature of the rules with the association between subject doubling after V in second person singular and the second person singular pronoun ‘gij/gie’.

From a statistical perspective many more linguistically interesting variable associations can be expected to surface upon closer investigation. The explorations described above merely attempt to indicate the great potential of association rule mining as a meaningful contribution to linguistic theory in general and syntactic theory in particular. Another promising approach could employ association rule mining to quantitatively validate existing and new typological hypotheses. This is in contrast with the current approach which focuses on exploration and identification of variable patterns. However, every approach will require extensive consultation with syntactic theorists to meaningfully interpret the data. SAND1 provides geographical maps of many individual variable distributions to facilitate interpretation and validation of potentially interesting association rules. The generated sets of induced association rules and the rule induction programme are publicly available for interactive exploration at http://dialectometry.net/syntax/.

6.8 Conclusions

This research has successfully demonstrated how associations between syntactic variables in Dutch dialects can be discovered computationally using an association rule mining technique based on proportional overlap. The rule induction system

Table 11. The most interesting implicational chain of association rules between four syntactic variables: d54a:after_v → d55a:after_v → p46a:g-lieden → p38b:gij/gie.

Variable 1/4: d54a:after_v (Subject doubling 2 singular)

As	gij	gezond	leeft,	leef- de	gij	langer.
if	$you_{singular}$	healthily	live,	live- $you_{singular,weak}$	$you_{singular,strong}$	longer

'If you live healthily you will live longer.'
Rank=6, Combination=6,509, Interestingness=52,78, Accuracy=92

Variable 2/4: d55a:after_v (Subject doubling 2 plural)

As	gulder	gezond	leeft,	leef- de	gulder	langer.
if	you_{plural}	healthily	live,	live- $you_{plural,weak}$	$you_{plural,strong}$	longer

'If you live healthily you will live longer.'
Rank=3, Combination=7.503, Interestingness=54,07, Accuracy=93

Variable 3/4: p46a:g-lieden (Subject pronouns 2 plural, strong forms)

We	geloven	dat	g-lieden	niet	zo	slim	zijn	als	wij.
we	believe	that	$you_{plural,strong}$	not	so	smart	are	as	we.

'We believe that you are not as smart as we are.'
Rank=1, Combination=10,321, Interestingness=58,38, Accuracy=99

Variable 4/4: p38b:gij/gie (Subject pronouns 2 singular, strong forms)

Ze	gelooft	dat	gij/gie	eerder	thuis	bent	dan	ik.
she	believes	that	$you_{singular,strong}$	earlier	home	are	than	I

'She thinks that you'll be home sooner than me.'
Rank=8, Combination=6,552, Interestingness=52,73, Accuracy=98

facilitates identification and exploration of previously unknown variable relationships and validation of existing parametric generalisations. The ability to define variable associations asymmetrically is considered to be an important property of the technique in the syntactic domain. The analysis of the sample data has indicated that the Piatetsky-Shapiro measure of interestingness adequately formulates the relationships between the evaluation factors of accuracy, coverage and completeness.

The application of the association rule mining technique to the Syntactic atlas of the Dutch dialects has revealed the existence of many potentially interesting associations with high accuracy and coverage values and showed considerable overlaps between the geographical distributions of syntactic variable pairs. The exploratory review has examined the highest ranked association rules and also discussed an implicational chain of variable associations. The results strongly indicate that many more potentially interesting associations between syntactic variables are likely to be uncovered upon further investigation.

6.9 Discussion

The approach presented in this paper to discover associations between syntactic variables can be extended and refined in several ways. For example, the candidate generation algorithm listed in Table 5 could be extended to incorporate exception rules as well. These are rules which cannot be predicted from existing knowledge and typically combine high accuracy with poor coverage values. Further refinements of the data mining procedure may include experimentation with alternative measures of interestingness and incorporation of additional rule quality evaluation

factors such as surprisingness, among others.

An interesting property of data mining applications such as association rule mining arises as more variables become available to the procedure. The formula in Figure 3 shows that the number of generated candidate association rules increases factorially with the number of variables. Also, increasing complexity is another source of combinatory explosion. These observations are relevant in the current context because the second volume of the SAND (SAND2) is due to appear at the end of 2007. Incorporation of the SAND2 data into the association rule discovery process will result in a linguistic database containing around 750 syntactic variables and covering all major syntactic microvariation domains. Although the linguistically trained mind may be extremely effective in heuristically associating variables, the astronomical SAND combination space will undoubtedly exceed human limits of association precision and capacity. Additionally, the compartmented and repetitive nature of data mining algorithms makes them good candidates for computational scaling and parallellisation using grid computing techniques. Therefore, a combination of the unsurpassed human heuristic capabilities with the verifiable precision and processing power available to data mining tools may well contribute to the understanding of the structural diversity of language varieties. There is, of course, no reason to stop incorporating more data into the procedure. For example, it could be really interesting to combine available phonological data with these syntactic data to discover potential associations between variables among linguistic levels (Spruit et al. n.d.).

An entirely different application of association rule mining analyses the set of variable associations to define clusters of geographically overlapping variables known as composite variables (Spruit 2006). This application assumes that if a group of variables nearly always occur together, then a single variable of such a group does not add to the variation between two language varieties by itself. Therefore, from a quantitative perspective the cluster of variables can be interpreted as one entity which should more accurately quantify syntactic variation. Preliminary visualisations of the distance relationships between Dutch dialects based on the Jaccard distance between composite syntactic variables appear to classify the Dutch dialect areas quite accurately. The dialect maps appear to be in line with expert opinion and correspond with dialect distance visualisations (cf. (Spruit 2006, Spruit et al. n.d.)) but require further research.

Finally, it would be interesting to compare the discovered variable associations with results based on more classic statistical methods such as Cramér's V or correspondence analysis. Cramér's V is a statistic which measures the strength of association between two categorical variables based on the χ^2-statistic. Time permitting, this approach could be well worth investigating. One of the method's attractive benefits is that it calculates the statistical significance of each variable pair association. Another statistical technique which may hold promise is correspondence analysis. This method resembles the factor analysis technique but has specifically been designed to help explore associations between categorical variables. However, the interpretability of the resulting correspondence visualisations may become an issue given the considerable geographical overlaps between the

syntactic variable distributions. Furthermore, a more fundamental shortcoming of the two alternative approaches described above is the inherent symmetric nature of the discovered variable associations.

References

Agrawal, R., Imielinski, T. and Swami, A.(1993), Mining association rules between sets of items in large databases, *in* P. Buneman and S. Jajodia (eds), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM Press, Washington, D.C., pp. 207–216.

Barbiers, S., De Vogelaer, G. and Devos, M.(2005), *Syntactic Atlas of the Dutch Dialects*, Vol. 1, Amsterdam University Press, Amsterdam.

Cornips, L. and Jongenburger, W.(2001), Elicitation techniques in a Dutch syntactic dialect atlas project, *in* H. Broekhuizen and T. van der Wouden (eds), *Linguistics in the Netherlands*, John Benjamins, Philadelphia/Amsterdam, pp. 53–63.

Frawley, W., Piatetsky-Shapiro, G. and Matheus, C.(1992), Knowledge discovery in databases: An overview, *AI Magazine* **13**, 213–228.

Freitas, A.(1999), On rule interestingness measures, *Knowledge-based Systems* **12**, 309–315.

Gianollo, C., Guardiano, C. and Longobardi, G.(to appear), Three fundamental issues in parametric linguistics, *in* T. Biberauer (ed.), *The Limits of Syntactic Variation*, John Benjamins, Philadelphia/Amsterdam.

Hand, D., Mannila, H. and Smyth, P.(2001), *Principles of Data Mining*, The MIT Press, Cambridge, MA.

Haspelmath, M.(to appear), Parametric versus functional explanations of syntactic universals, *in* T. Biberauer and A. Holmberg (eds), *The Limits of syntactic variation*, Benjamins, Amsterdam.

Horn, H.(1966), Measurement of overlap in comparative ecological studies, *The American Naturalist* **100**, 419–424.

Lenca, P., Meyer, P., Vaillant, B. and Lallich, S.(to appear), On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid, *European Journal of Operational Research*, Elsevier.

McGarry, K.(2005), A survey of interestingness measures for knowledge discovery, *The Knowledge Engineering Review* **20**, 39–61.

Newmeyer, F.(2005), *Possible and probable languages: a generative perspective on linguistic typology*, Oxford University Press, Oxford.

Piatetsky-Shapiro, G.(1991), Discovery, analysis and presentation of strong rules, *in* G. Piatetsky-Shapiro and W. Frawley (eds), *Knowledge Discovery in Databases*, AAAI/MIT Press, pp. 229–248.

Rizzi, L.(1986), Null objects in Italian and the theory of pro, *Linguistic Inquiry* **17**, 501–557.

Spruit, M.(2006), Measuring syntactic variation in Dutch dialects, *in* J. Nerbonne and W. J. Kretzschmar (eds), *Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation*, Vol. 21, Oxford University Press, Oxford, pp. 493–506.

Spruit, M., Heeringa, W. and Nerbonne, J.(n.d.), Associations among linguistic levels, Presented at a special session at Digital Humanities, Paris, 6 July 2006.