1 TOWARDS LINGUISTIC KNOWLEDGE DISCOVERY IN LANGUAGE VARIATION DATABASES

MARCO R. SPRUIT : m.r.spruit@cs.uu.nl

This paper introduces an industry-standard knowledge discovery process into the realm of language variation research. The Linguistic Knowledge Discovery (LINKDIS) approach advocated here integrates exploratory data mining techniques and related statistical procedures in a manner which inherently emphasises the importance of linguistic interpretability of uncovered results. In addition the proposal raises four interrelated research questions to complement and improve current research strategies with respect to syntactic and morphological variation and the relation between language variation and geography.

1.1 INTRODUCTION

This paper presents a new methodological approach which aims to contribute to language variation research in general, and syntactic and morphological variation research in Dutch Low Saxon language varieties in particular. The advocated Linguistic Knowledge Discovery (LINKDIS) approach integrates and extends two related research strands. First, it improves upon the exploratory data mining techniques in language variation research and the linguistic interpretation of uncovered variable associations which SPRUIT (2007) introduced. Second, it investigates quantitative methods to help assess the relationships between linguistic levels and the role of geography as an underlying factor of influence in language variation. The latter research strand improves upon SPRUIT, HEERINGA AND NERBONNE (IN PRESS).

The suggested approach introduces an industry-standard and human-centered knowledge discovery process into the realm of language variation research and aims to provide an answer to the research question of how linguistically relevant associations between syntactic variables can be discovered. It should be noted that a nearly identical question was already raised in SPRUIT (2007), but with one minor though very significant difference: the question raised here specifically explores the l i n g u i s t i c relevance of the results.

In addition, the strategy advocated provides a best-of-practice, methodological framework to help structure language variation research which aims to uncover linguistically interesting variable associations at the syntactic level, linguistically interesting associations between syntactic and morphological variables, and the

extent to which the role of geography as an extralinguistic factor of influence can be assessed and controlled for, with respect to linguistic knowledge discovery in Dutch Low Saxon language variation databases.

The scientific contribution of the proposed research approach is twofold. First, it may contribute to a better understanding of syntactic variation, as well as the relationship between syntactic and morphological variation, in the Dutch Low Saxon language area. Ultimately, a better insight into the interrelationships between language variation patterns within and among the linguistic levels might even help uncover the grammatical dependencies within the human language system in general. Second, it may contribute to a better understanding of the relationship between geographical patterns of language variation and variation patterns outside the realm of linguistics. Ultimately, a better insight into appropriate methods to help identify and unravel interrelationships between human aspects across research area boundaries might help focus the future research agenda in humanities computing.

1.2 A QUESTIONABLE OUTLINE

The presentation structure of the linguistic knowledge discovery approach described in this paper is inspired by "The Three W's" and "The Five W's (and one H)" information-gathering concepts which were first popularised in TRUMBULL (1888:120). The "Three W's"—What? Why? What of it? —method was soon expanded into "The Five W's"—When? Where? Whom? What? Why? —and has for long been considered an influential, inspirational and imaginative checklist in journalism and research, among others (FIVE WS, 2009). The research approach presented in this paper does not explicitly investigate the "When?" question, but it is applicable to both synchronic and diachronic language variation research. This leaves the following "W's" and one "H" as structuring aspects of this paper: "Where?" "Why?" "What?" "How?" "Whom?" and "What of it?".

The introductory "Where?" section briefly indicates the geographical boundaries of the Dutch Low Saxon language varieties under discussion, Then, the section "Why?" describes the overall aims of the proposed methodological approach from a strategic perspective. The section "What?" discusses the key objectives from a tactical perspective and includes formulations of four relevant research questions. The section "How?" introduces the linguistic knowledge discovery process and elaborates on its six phases from an operational perspective. The section "With whom?" briefly reviews collaboration opportunities through an examination of desirable expertise. The "What of it?" section concludes this paper with a discussion of the potential impact of the envisioned line of research.

1.3 WHERE? THE DUTCH LOW SAXON AREA

Even though this paper focuses on a methodological approach, it frequently refers to Dutch Low Saxon as one particularly interesting group of language varieties to which the research approach could well be applied to. Its interest revolves around the availability of several syntactic and morphological variation databases which detailedly document Dutch Low Saxon dialects (as well as many other language varieties in the Dutch language area). Dutch Low Saxon is especially suitable for the proposed knowledge discovery process because the LINKDIS method requires a "critical mass" of statistically interpretable dialect data. Section 5 discusses the SAND and MAND atlases which provide this "critical mass".[1]
The Dutch Low Saxon dialect varieties under discussion are a diverse group of Low Saxon—*i.e.* Low German—dialects spoken in the northeastern area in the Netherlands. Figure 1 shows that varieties occur in the Dutch provinces of Groningen, Drenthe, Overijssel, the Veluwe and Achterhoek regions in Gelderland, and the Stellingwerven areas in southern Friesland (BLOEMHOFF, 2009). Even though Low Saxon dialects are found throughout northern Germany in an area approximately three times as large as the entire Netherlands, Dutch Low Saxon dialects are considered to be varieties of Dutch (NIEBAUM AND MACHA, 2006:221).

Figure 1: The Dutch Low Saxon dialect area (GRÖNNEGER1 (2009)).

[1] Other groups of language varieties within the Dutch language area, such as Limburgish and West Flemish, seem less likely candidates for the proposed knowledge discovery process because less data are available.

1.4 WHY? OVERALL AIMS

From a strategic perspective, the envisioned methodological approach aims to contribute to language variation research in general, and syntactic and morphological variation research in Dutch Low Saxon language varieties in particular. The motivation for this research is twofold:

(I) Contribute to better understanding of syntactic variation, as well as the relationship between syntactic and morphological variation, in the Dutch Low Saxon language area. It may help determine whether there might be structural, typological constraints linking variation at the linguistic levels. Ultimately, a better insight into the interrelationships between language variation patterns within and among the linguistic levels may help uncover the grammatical dependencies within the human language system in general.

(II) Contribute to a better understanding of the relationship between geographical patterns of language variation and variation patterns outside the realm of linguistics. This line of thought assumes that geographical patterns of syntactic variation may reflect residues of political, social and cultural changes over time. Ultimately, a better insight into appropriate methods to help identify and unravel interrelationships between human aspects across research area boundaries may help focus the future research agenda in humanities computing.

Regarding the latter ultimate goal, SPRUIT (2008:15) provides two examples of geographical patterns of syntactic variation which may reflect residues of political, social and/or cultural changes over time. First, a correlation between dialect borders and the political history of the province of Friesland is discussed as documented in BREE (1994). Second, a striking correlation is uncovered between a syntactic dialect border and the social-cultural Catholic-Protestant boundary which HEEK (1954) first described, around the northern border of the Noord-Brabant area. The advocated research approach might provide comparable insights into the relatively unchartered expanse of potentially relevant interdisciplinary variable relationships.

1.5 WHAT? KEY OBJECTIVES

From a tactical perspective, the LINKDIS method approaches the overall aims by integrating and extending two related research strands in the field of Dutch language variation as introduced in SPRUIT (2007) and SPRUIT, NERBONNE AND HEERINGA (IN PRESS) regarding:

(A) Exploratory data mining techniques in language variation research and the linguistic interpretation of uncovered variable associations (SPRUIT, 2007).

(B) Quantitative methods to assess the associations between linguistic levels and the role of geography as an underlying factor of influence (SPRUIT ET AL., IN PRESS).

Regarding research strand (A): SPRUIT (2007) introduces a data mining technique in the realm of Dutch language variation based on the association rule mining method developed by AGRAWAL ET AL. (1993). It examines the first volume of the Syntactic Atlas of the Dutch Dialects (SAND1; BARBIERS ET AL., 2005) which contains 485 geographical distributions of syntactic phenomena in 267 Dutch dialect varieties, 66 of which—*i.e.* around 25%—are considered to be Dutch Low Saxon language varieties. SAND1 covers syntactic variation related to the left periphery of the clause and pronominal reference and includes variation with respect to complementisers, subject pronouns and expletives, subject doubling and subject cliticisation following yes/no, reflexive and reciprocal pronouns, and fronting phenomena.

The technique in SPRUIT (2007) contributes to the global linguistic research effort of parameterisation of the structural diversity of language varieties by proposing a computational method to discover syntactic variable associations automatically. The rule induction system based on geographical overlap uncovers several potentially interesting relationships between syntactic microvariables in Dutch dialects. The applied concept of proportional overlap originally stems from research areas such as ecology and biogeography and is notably examined in HORN (1966).

Table 1: An example of a SAND1 association rule.

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 group 'g-lieden' belongs to the same paradigm as the singular pronoun 'gij'.

To clarify the key concept of a rule induction system based on geographical overlap, Table 1 shows the highest ranked association rule in SAND1 according to the data mining procedure in SPRUIT (2007). It should be interpreted as follows. One

of the variables in map A on page 46 in SAND1 is associated with a variable in map B on page 38. The rule states that, in the context of a strong *plural* subject pronoun in second person, IF one of the pronouns in the 'g-lieden' group occurs THEN the strong *singular* subject pronoun in second person 'gij' (or 'gie') nearly always occurs as well, as indicated by the accuracy value of 99 percent. A linguistic interpretation of this association rule could be that the pronouns belong to the same paradigm.[2]

However, in contrast with the association rule shown in Table 1, a frequently recurring problem with such a computation-oriented data mining approach is that a large number of uncovered variable associations remain hard to interpret linguistically. The results in SPRUIT (2007) indicate that a more focused knowledge discovery process is required to maintain an effective balance between linguistic and technological needs, limitations and opportunities. PIATETSKY-SHAPIRO (1991) first emphasised that knowledge should be the end product of data-driven discovery by introducing the phrase "Knowledge Discovery in Databases" (KDD) to describe a more complete data mining process.

The current proposal embraces the CRoss Industry Standard Process for Data Mining (CRISP-DM) knowledge discovery model (CHAPMAN ET AL., 2000) to effectively guide the linguistic knowledge discovery process. The CRISP-DM model incorporates and extends the human-centered KDD process as first described by BRACHMAN AND ANAND (1996) and popularised by FAYYAD ET AL. (1996) in an industry- and tool-neutral manner. HIPP, GÜNTZER AND NAKHAEIZADEH (2002) notably describe an enhanced version of the CRISP-DM model to optimise the association rule mining process. CAO AND ZHANG (2006) provide a similar method referred to as domain-driven data mining. MARBÁN ET AL. (2009) suggest to revise the CRISP-DM model by aligning software development processes with data mining processes, which results in a data mining engineering viewpoint. This brings about the first research question of the proposed research approach:

(i) How can linguistically relevant associations between syntactic variables be discovered?

The research question above aims to innovate language variation research by introducing a proven methodological process into the field of linguistics which focuses on actionable linguistic knowledge discovery.[3] The proposed method is outlined in the "How?" section below. It is fundamentally different from the computational approach in SPRUIT (2007) which focuses on the exploration of algo-

[2] The descriptive statistics confirm the strength of the association rule in Table 1 as follows. The antecedent occurs in 105 dialects, whereas the consequent was recorded in 116 locations. The intersection of these two sets of locations still contains 104 dialects. In only (117-max(116-105)=) one dialect the two syntactic variables do not co-occur, resulting in 99% accuracy.

[3] Actionable knowledge refers to potential linguistic discoveries which "[…] provide the ability to take a *proactive* stance rather than a *passive* or, at best, a *reactive* approach" (Thierauf, 2001:4).

rithmically interesting co-occurrences of syntactic variables. A second, inherently innovative aspect of this research approach would be the first-time inclusion of the second volume of the Syntactic Atlas of the Dutch Dialects (SAND2; BARBIERS ET AL., 2008) in a data mining analysis. The SAND2 database contains 697 geographical patterns of syntactic variables and focuses on syntactic variation in the right periphery of the clause. SAND2 includes variation related to verbal clusters, cluster interruption, morphosyntactic variation, the negative particle, and negative concord and quantification. Therefore, it seems highly probable that a joint analysis of the complementary SAND1 and SAND2 data will result in the discovery of various associations between syntactic variables among different syntactic subdomains. Thus, the combination of the proposed CRISP-DM process and the integrated investigation of both SAND databases should provide the optimal conditions to help answer the second research question:

(ii) What are linguistically interesting variable associations at the syntactic level?

The research question above does not limit itself to the applicability and interpretability of various data mining techniques as documented by WEIS AND INDURKHYA (1997), BERRY AND LINOFF (1997) and FAYYAD ET AL. (1996), among others. One novel data mining idea which this type of research enables—when applied to the SAND atlases—involves the exploration of linguistically interesting associations at a categorisation level using available part-of-speech (POS) tags. These word-category disambiguation markers, which have been added to the SAND data semi-automatically, might provide additional insights into syntactic variable relationships. The knowledge discovery process needs to be guided through suggestions and hypotheses provided by project-external language variation experts. They should be interviewed individually before data mining commences.

Regarding research strand (B): SPRUIT *ET AL.* (IN PRESS) introduce a quantitative method to investigate the extent of the associations between the linguistic levels of syntax, pronunciation and vocabulary based on their geographical variation patterns. One key finding of this research is that associations between all three linguistic levels are measurable, even when the influence of geography as an underlying factor is accounted for. These modest but substantial associations might indicate structural constraints between the linguistic levels. However, one of the more interpretational complexities regarding the pronunciational differences under investigation involved the potential interplay of phonetic, phonological and morphological variation.

The current research proposal proceeds from these remarks by incorporating another set of Dutch Low Saxon language variation databases into the linguistic knowledge discovery process: the first and second volumes of the Morphological Atlas of the Dutch Dialects (MAND1, SCHUTTER ET AL., 2005; MAND2, GOEMAN ET AL., 2008). Based on morphological data from the Goeman-Taeldeman-Van Reenen project (GTRP; Goeman and Taeldeman, 1996), the MAND databases contain 1876 geographical distributions of purely morphological phenomena in 613 dialects in the Netherlands and Flanders, about 140 places of which—*i.e.*

around 23%—are Dutch Low Saxon language varieties. The morphology-syntax interface has been widely studied during past decades in attempts to better understand the relation between word formation and sentence formation (LI, 2005; ACKEMA AND NEELEMAN, 2004; BOOIJ, 2002; CHOMSKY, 1995; BAKER, 1988). The present approach introduces quantitative techniques to guide data-driven searches for structural constraints between the two interrelated linguistic levels. Note that this process requires an investigation of several methodological challenges regarding combinations of association rules on multiple datasets as well (*e.g.* ZHAO ET AL., 2007). To summarise, the third research question can be stated as follows:

(iii) What are linguistically interesting associations between syntactic and morphological variables?

Both the quality and impact of any answers to research questions (i), (ii) and (iii) will depend on an adequate assessment of the role of geography and other extralinguistic factors. Even though, in principle, research questions (i-iii) may be entirely legitimate in linguistics in itself, without any relation to extralinguistic factors, the geographical dependency arises in the current context from the fact that the rule induction system is based on geographical co-occurrences between linguistic variables. Further unraveling the influence of geographic distance as an underlying factor in language variation may even be more important when considering that geographical patterns of variation may also reflect residues of political, social and cultural changes over time (NERBONNE AND HEERINGA, 2007). In other words, if it were possible to analyse language variation patterns separately from geographical distances (and, ultimately, from other extralinguistic factors as well) to a certain extent, then this would greatly facilitate linguistic analyses in identifying implicational variable chains and other association patterns in both space and time (*i.e.* language variation and historical linguistics). One innovative perspective which this research proposes, is to explore language variation classifications based on Minimum Description Length (MDL) algorithms such as KRIMP (SIEBES, VREEKEN AND LEEUWEN, 2006; VREEKEN AND SIEBES, 2008). Since these classification techniques are not based on distance metrics but on aggregate attribute compressions instead, a comparison between various classification techniques may provide additional insights into the role of geography as a factor in linguistic knowledge discovery processes. TATTI AND HEIKINHEIMO (2008) notably provide a possibly interesting alternative through their probabilistic model which implicitly induces a tree-based representation of the mining results. Concludingly, the fourth and final research question addresses the following fundamental issue in language variation research:

(iv) To what extent can the role of geography as an extralinguistic factor of influence be assessed and controlled for—with respect to linguistic knowledge discovery in language variation databases?

1.6 HOW? RESEARCH PLAN

From an operational perspective, the LINKDIS method introduces the interactive and iterative CRISP-DM process to the language variation research domain, fine-tuning the technique where appropriate. Figure 2 shows the main steps that compose the LINKDIS process to be executed:

1. Linguistic understanding
2. Data understanding
3. Data preparation
4. Data mining
5. Evaluation
6. Knowledge deployment

The arrows illustrate the cyclical nature of the LINKDIS process which revolves around the linguistic datawarehouse. The method itself may be repeated as well, resulting in a LINKDIS process cycle. The practical steps of the research approach are briefly described below.

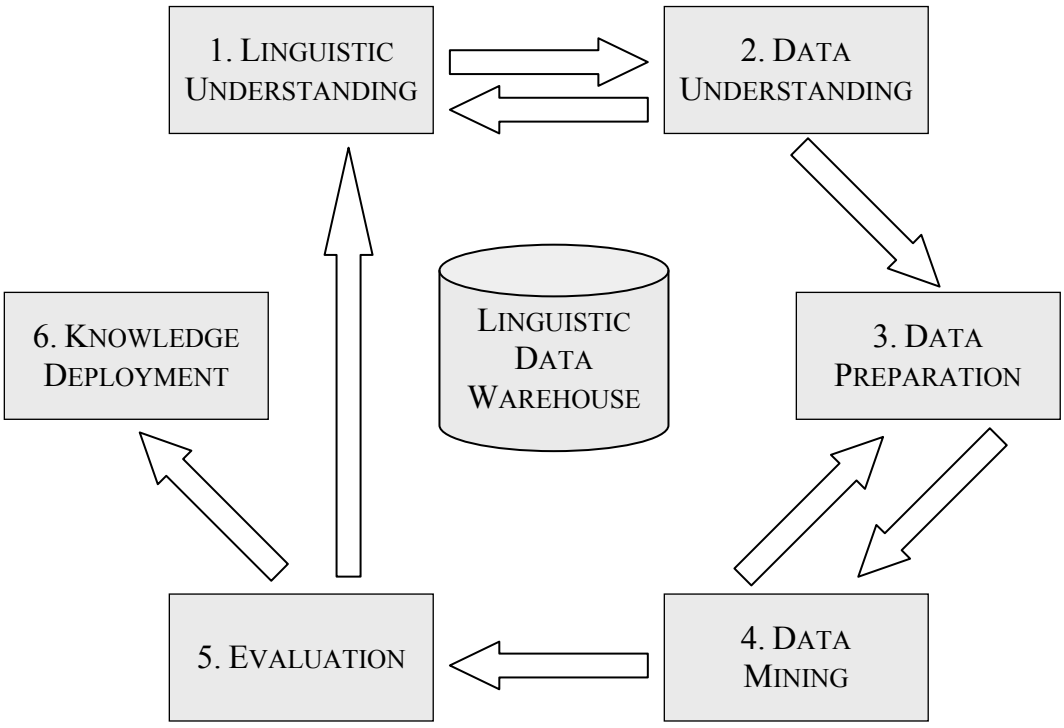

Figure 2: The LINKDIS process.

1.6.1 Linguistic understanding

The first goal is to uncover important factors at the beginning which might influence the outcome of the knowledge discovery process. Therefore, relevant prior linguistic knowledge needs to be be assessed through a series of semi-structured interviews with a representative sample of experts in the field of Dutch Low Saxon language variation. This step should result in a compilation of relevant syntactic and morphological variation insights which, then, need to be turned into data mining problem definitions—*i.e.* hypotheses.

1.6.2 Data understanding

Once the initial requirements are formulated, an intimate understanding of the available Dutch Low Saxon attributes and values in the SAND and MAND language variation databases becomes essential. Do the SAND and MAND databases have the potential to answer the linguistic questions derived in the first step? The linguistic data mining problem definitions which the available SAND and MAND data cannot possibly answer sufficiently, are filtered out of the set of language variation problems which will be investigated. SHARMA AND OSEI-BRYSON (2008) notably describe the role of human intelligence in this phase.

1.6.3 Data preparation

The relevant subset of the SAND and MAND data will need to be further preprocessed to improve the overall quality of the knowledge discovery process. Additionally, new useful attributes will be derived from the data in this step, which contain higher level information only implicitly contained in the original data. Examples include the derivation from the encoded location occurrence of the proper "province" name and the Dutch Low Saxon dialect group (such as "Westerkwartiers" or "Stellingwerfs"). Finally, the smallest subset of the available syntactic and morphological variables, which are deemed relevant to the knowledge discovery problem definitions as formulated in the first step, can be organised into an online data warehouse. The data warehouse integrates the SAND and MAND datasets into one uniformly structured language variation database which can be made publicly accessible upon process completion.

1.6.4 Data mining

The central step of data mining consists of matching the formulated linguistic goals with appropriate data mining methods and machine learning techniques (*e.g.* WITTEN AND FRANK, 2005). For example, research question (iv) might best be answered by exploring language variety classifications based on MDL algorithms

such as KRIMP (VREEKEN AND SIEBES, 2008). In constrast, SPRUIT (2007) shows that association rule mining methods—of the form "IF variable A THEN variable B"—are able to help answer research questions (ii) and (iii). Other data mining techniques such as clustering and regression are to be investigated as well. Finally, special attention needs to be given to effective visualisation techniques and tools in the spirit of FAYYAD, GRINSTEIN AND WIERSE (2001), among others. The results need to be collected in a linguistic model base.

1.6.5 Evaluation

In this interpretation step the model is evaluated thoroughly from both methodological and linguistic perspectives. The sequence of steps which the data mining method executed in order to construct the model needs to be reviewed in order to ensure that it properly achieves the linguistic objectives. Following the Delphi research approach (LINSTONE AND TUROFF, 1975), this validation step will include presentations of the findings to the language variation experts which were interviewed in the first step. Do the results provide linguistically relevant answers to the hypotheses formulated in phase one? The step ends with a decision on whether the data mining results are valid and thus usable. Otherwise, the negative feedback is incorporated through an improved reiteration of all preceeding steps. This integration of linguistic expertise in the key input, output and feedback phases—in combination with relevant linguistic findings in literature—should ensure linguistic relevance of all discovered knowledge.

1.6.6 Knowledge deployment

All discovered and verified knowledge models should be made publicly available through a unified knowledge base in appropriate formats, including rules, lists, groupings, question-answer pairs and visualisations. The knowledge base can be considered the external user interface part of the language variation datawarehouse. Other deployment steps include application guidelines for further research and a monitoring and maintenance plan for the knowledge base. Additionally, relevant findings should be presented as well at various conferences and published in relevant publications in order to collect as much feedback as possible regarding all uncovered and discovered linguistic knowledge.

1.7 WITH WHOM? COLLABORATION OPPORTUNITIES

Due to the highly multidisciplinary nature of this line of research, the LINKDIS process needs to be carried out in collaboration with at least three types of research groups. First, continous guidance is required from language variation researchers to ensure linguistic relevance throughout the process. Second, the ap-

proach needs algorithmic expertise to help customise various data mining algorithms for optimal linguistic exploitation. Third, highly iterative IT implementation methods as described above are far from trivial to execute successfully and require IT methodological expertise to safeguard and finetune a consistent and successful LINKDIS process.

1.8 WHAT OF IT? POTENTIAL IMPACT

The research results of the described LINKDIS project can directly benefit the Dutch Low Saxon language variation research community by providing empirically-derived answers to current issues in language variation research. It can also benefit language variation researchers around the world by outlining the required steps to effectively perform data-driven discovery in language variation databases, and by providing specialised language variation algorithms. Furthermore, the proposed LINKDIS process can improve the overall data quality of language variation databases, enabling an increased overall quality of related linguistic research. Any, even tentative, answers are likely to trigger new and more focused research questions as a consequence, further illustrating the potential impact of uncovering answers to the four main questions specified earlier. They might even influence the direction and preferred choices of method of future language variation research. Ultimately, this research approach might be scaled up to supersede language variation databases and include demographical, social and political variation data as well.

The advocated LINKDIS process improves upon and customises the KDD and CRISP-DM process methods. The former has already been proven successful in diverse fields such as genetics, astronomy and marketing, among many others (FAYYAD ET AL., 1996), whereas the latter epitomises the current industry-standard guidelines in business implementations around the world (KDNUGGETS, 2007). A successful implementation of this process in linguistics may well result in subsequent applications of the methodological framework in future knowledge discovery projects within this field, further propagating data-driven research in linguistics, and perhaps even throughout the humanities and arts.

In this respect, the current LINKDIS research approach can also be envisioned as a relevant method in the context of the prestigious Common Language Resources and Technology Infrastructure (CLARIN) project, which, in November 2008, was granted nine million euros by the Dutch Ministry of Education, Culture and Science, based on an ESFRI Advisory Report (DUINEN ET AL., 2008). The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable (CLARIN, 2008). Interestingly, the ESFRI report advised against inclusion of image and sound resources and advised against an integration of social sciences infrastructure. Instead, the report advised to focus on a language resource infrastructure only. This is perfectly in line with the current research approach.

The LINKDIS method outlined above aims to provide a methodology-sound knowledge discovery process specifically designed for linguistic research, while anticipating numerous additional exploitation opportunities with respect to the upcoming CLARIN language resource infrastructure. Similarly, the LINKDIS approach may also be relevant to other recently started projects such as the European DARIAH and the Dutch AlfaLab initiatives, which focus on developing a common data and tools research environment for the humanities and arts. All in all, the LINKDIS process provides a best-practice-based research approach which effectively channelises the many emerging and exciting IT opportunities within dialectal variation research in general and Dutch Low Saxon variation research in particular.

1.9 REFERENCES

ACKEMA, P., NEELEMAN, A. (2004). Morphology ≠ Syntax. In G. Ramchand and C. Reiss (eds.) *Oxford Handbook of Linguistic Interfaces*. Oxford: Oxford University Press.

AGRAWAL, R., IMIELINSKI, T., SWAMI, A. (1993). Mining association rules between sets of items in large databases. In P. Buneman, S. Jajodia (eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26–28* (pp. 207–216), New York: ACM Press.

BAKER, M. (1988). *Incorporation: a theory of grammatical function changing*. Chicago: University of Chicago Press.

BARBIERS, S., BENNIS, H., VOGELAER, G. DE, AUWERA, J. VAN DER, HAM, M. VAN DER (EDS.) (2008). *Syntactic Atlas of the Dutch Dialects, Volume 2*. Amsterdam: Amsterdam University Press.

BARBIERS, S., BENNIS, H., DEVOS, M., VOGELAER, G. DE, HAM, M. VAN DER (EDS.) (2005). *Syntactic Atlas of the Dutch Dialects, Volume 1*. Amsterdam: Amsterdam University Press.

BERRY, M., LINOFF, G. (1997). Data mining techniques: for Marketing, Sales, and Customer support. Toronto: Wiley.

BLOEMHOFF, H. (2009). *Nedersaksisch*. Retrieved April 8, 2009, from http://taal.phileon.nl/pdf/nedersaksisch.pdf.

BOOIJ, G. (2002). *The Morphology of Dutch*. Oxford: Oxford University Press.

BRACHMAN, R., ANAND, T. (1996). The process of knowledge discovery in databases: A human-centered Approach. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 37–57). Cambridge: AAAI/MIT Press.

BREE, C. VAN (1994). The development of so-called Town Frisian. In Bakker, P., Mous, M. (eds.), *Mixed Languages. 15 Case Studies in Language Intertwining* (pp. 69–82), Studies in Language and Language Use, Volume 13, Amsterdam:

IFOTT.

CAO, L., ZHANG, C. (2006), Domain-driven data mining: A practical methodology. *International Journal of Data Warehousing and Mining*, 2(4), 49–65.

CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C., WIRTH, R. (2000). *CRISP-DM 1.0. CRoss Industry Standard Process for Data Mining*. Retrieved January 7, 2009, from http://www.crisp-dm.org/.

CHOMSKY, N. (1995). *The Minimalist program*. Cambridge: MIT Press.

CLARIN (2008). *Common Language Resources and Technology Infrastructure.* Retrieved January 8, 2009, from http://www.clarin.eu/.

DUINEN, R. VAN, BOOIJ, G., OOSTERLINCK, A. BARON, WITHOLT, B. (2008). *ESFRI Advisory Report*. Retrieved January 8, 2009, from http://www.minocw.nl/documenten/NWO_ESFRI_compleet.pdf.

FAYYAD, U., GRINSTEIN, G., WIERSE, A. (2001). *Information Visualization in Data Mining and Knowledge Discovery.* San Francisco: Morgan Kaufmann.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine*, 17, 37–54.

FIVE WS. (2009, February 14). In *Wikipedia, The Free Encyclopedia*. Retrieved 15:49, February 17, 2009, from http://en.wikipedia.org/w/index.php?title=Five_Ws.

GOEMAN, T., OOSTENDORP, M. VAN, REENEN, P. VAN, KOORNWINDER, O., BERG, B. VAN DEN (EDS.) (2008). *Morphological Atlas of the Dutch Dialects, Volume 2*. Amsterdam: Amsterdam University Press.

GOEMAN, A., TAELDEMAN, J. (1996). Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval* 48, 38–59.

GRÖNNEGER1 (2009): *Koart Leegsaksisch*. Retrieved April 24, 2009, from http://commons.wikimedia.org/wiki/File:Koart_Leegsaksisch.png.

HEEK, F. VAN (1954). Het geboorteniveau der Nederlandse Rooms-Katholieken. Leiden.

HIPP, J., GÜNTZER, U., NAKHAEIZADEH, G. (2002). Data Mining of Association Rules and the Process of Knowledge Discovery in Databases. In Perner, P. (ed.), *Advances in Data Mining 2002* (pp. 15–36). Berlin Heidelberg: Springer-Verlag.

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

LI, Y. (2005). A Theory of the Morphology-syntax Interface. Cambridge: MIT Press.

LINSTONE, H., TUROFF, M. (1975). *The Delphi method: Techniques and applications*. Reading: Addison-Wesley.

NERBONNE, J., HEERINGA, W. (2007). Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In S. Featherston, W. Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base* (pp. 267–297), Ber-

lin: Mouton De Gruyter.

NIEBAUM, H., MACHA, J. (2006). Einführung in die Dialektologie des Deutschen, Volume 2, neubearbeitete Auflage, Tübingen: Max Niemeyer.

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

KDNUGGETS (2007). *Poll: Data Mining Methodology*. Retrieved 16:50, February 23, 2009, from http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.

SCHUTTER, G. DE, BERG, B. VAN DEN, GOEMAN, T., JONG, T. DE (EDS.) (2005). *Morphological Atlas of the Dutch Dialects, Volume 1*, Amsterdam: Amsterdam University Press.

SHARMA, S., OSEI-BRYSON, K. (2008). Role of Human Intelligence in Domain Driven Data Mining. In L. Cao, S. Yu, C. Zhang, H. Zhang (eds.), *Data Mining for Business Applications*. (pp. 53–61). New York: Springer.

SIEBES, A., VREEKEN, J., LEEUWEN, M. VAN (2006). Item Sets that Compress. In J. Ghosh, D. Lambert, D. Skillicorn, J. Srivastava (Eds.), *Proceedings of the SIAM Conference on Data Mining* (pp. 393–404). SIAM.

SPRUIT, M., HEERINGA, W., NERBONNE, J. (IN PRESS). Associations among linguistic levels. *Lingua, Special issue on Syntactic databases*.

SPRUIT, M. (2008). *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD thesis, University of Amsterdam, LOT Dissertation Series 174, Utrecht: Utrecht Institute of Linguistics.

SPRUIT, M. (2007). Discovery of association rules between syntactic variables. Data mining the Syntactic Atlas of the Dutch dialects. In P. Dirix, I. Schuurman, V. Vandeghinste, F. van Eynde (eds.), *Computational Linguistics in the Netherlands 2006. Selected papers from the seventeenth CLIN meeting* (pp. 83–98), Utrecht: Utrecht Institute of Linguistics.

TATTI, N., HEIKINHEIMO, H. (2008). Decomposable Families of Itemsets. In *Machine Learning and Knowledge Discovery in Databases* (pp. 472–487), Lecture Notes in Computer Science 5212, Berlin / Heidelberg: Springer.

THIERAUF, R. (2001). *Effective Business Intelligence Systems*. Westport: Quorum Books.

TRUMBULL, H. (1888). Teaching and Teachers: Or, The Sunday-school Teacher's Teaching Work and the Other Work of the Sunday-school Teacher, Philadelphia: J. Wattles.

VREEKEN, J., SIEBES, A. (2008). Filling in the Blanks - Krimp Minimisation for Missing Data. In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*, Pisa, Italy.

WEIS S., INDURKHYA, N. (1997). *Predictive datamining, a practical guide*. San Francisco: Morgan Kaufmann.

WITTEN, I., FRANK, E. (2005). *Data Mining: Practical Machine Learning Tools*

and Techniques. Second edition, San Francisco: Morgan Kaufmann.

ZHAO, Y., ZHANG, H., FIGUEIREDO, F., CAO, L., ZHANG, C. (2007). Mining for Combined Association Rules on Multiple Datasets. *Proceedings of the 2007 international workshop on Domain driven data mining* (pp. 18–23), New York: ACM Press.