Induction of Classifications from Linguistic Data

Rainer Osswald[1] and **Wiebke Petersen**[2]

Abstract. We present a flexible approach for extracting hierarchical classifications from linguistic data. To this end, the framework of observational logic is introduced, which extends the logic that underlies standard Formal Concept Analysis by allowing disjunctive rules and exclusions. We give a rigorous mathematical characterization of how the chosen rule type affects the structure of the induced hierarchy. The framework is applied to the induction of hierarchical classifications from linguistic databases. The pros and cons of several types of hierarchies are discussed in detail with respect to criteria such as compactness of representation, suitability for inference tasks, and intelligibility for the human user.

1 THE LOGIC OF LINGUISTIC CLASSIFICATION

A simple method for classifying (linguistic) data is provided by *taxonomic trees*, which are ubiquitous in linguistic textbooks. For example, nominal words are traditionally subdivided into pronouns, nouns, adjectives, etc; pronouns are further subdivided into interrogative pronouns, personal pronouns, etc, etc. From a logical point of view each concept of a taxonomic tree *implies* its superordinate concept; e.g. *pronoun* implies *nominal word*. Furthermore, any two subconcepts of the same concept are *incompatible*, as e.g. *noun* and *adjective*. In addition, classification by taxonomic trees is often assumed to be *exhaustive* in the sense that every concept implies the disjunction of its immediate subconcepts.

Systemic networks, which have their roots in systemic grammar (e.g. [10]), provide a more sophisticated formalism for presenting linguistic classification. Figure 1 shows a small fragment of such a network. The classifiers aligned to the right of a bar constitute a *choice system*; its members are assumed to be pairwise incompatible. Choice systems have *entry conditions* which are determined by the network structure: brace and bar correspond respectively to conjunction and disjunction. Entry condition and choice disjunction are assumed to imply each other. So, *third* $\wedge$ *singular* is equivalent to *feminine* $\vee$ *masculine* $\vee$ *neuter*, and every two members of $\{$*feminine*, *masculine*, *neuter*$\}$ are incompatible.

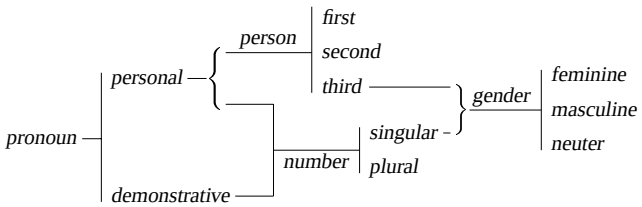


Figure 1. Part of a systemic classification of English pronouns (after [21])

Both types of classifications can be regarded as *theories* consisting of (universally quantified) conditionals whose premises and conclusions are built by finite conjunction and disjunction from primitive predicates or concepts (see Section 2.1). Given such a classification, it is natural to ask for the *conjunctive concepts* or *entity types* that are compatible with the classification. Roughly speaking, a conjunctive concept is a set of primitive concepts that is *consistently closed* with respect to the classificational theory in question (see Section 2.3 for details).[3]

The rest of the paper is organized as follows: In Section 2, observational theories are introduced, which subsume simple inheritance networks as well as Horn theories. In addition, it is shown how the canonical universe of such a theory depends on the class the theory belongs to. This result is used in Section 3 for inducing different types of conceptual hierarchies from a formal context. In the case of Λ-free Horn theories, the induced hierarchies essentially coincide with the concept lattices of Formal Concept Analysis. In Section 4, we apply the framework to the induction of classifications from linguistic data. We discuss the effects of varying the underlying theory class. Moreover, we consider the selective addition of disjunctive rules. One possible application is the construction of classification trees.

2 CLASSIFICATION AS OBSERVATIONAL THEORY

2.1 Observational theories

Linguistic classifications of the sort presented in Section 1 can be regarded as first-order theories consisting of universally quantified conditionals of the form $\forall x(\phi x \rightarrow \psi x)$, henceforth written as $\phi \subseteq \psi$. The one-place predicates ϕ and ψ are inductively built by $\wedge$ and $\vee$ from members of a set Σ of *primitive* predicates plus two predicates V and Λ, which stand respectively for $\lambda x(x = x)$ and $\lambda x(x \neq x)$.[4] Following [20] we call predicates constructed that way *observational*.

Let $\mathrm{T}[\Sigma]$ be the *(free) term algebra* of observational predicates over Σ. *Observational statements* over Σ are statements of the form $\phi \subseteq \psi$, with $\phi, \psi \in \mathrm{T}[\Sigma]$; *observational theories* are sets of such

[1] Applied Computer Science VII, University of Hagen, Germany
[2] Institute of Language and Information, Department of Computational Linguistics, University of Düsseldorf, Germany

[3] See [2] for an early approach of this kind, which employs *simple inheritance theories with binary exclusions*.

[4] The λ-notation indicates predicate abstraction; $\phi \wedge \psi$ stands for $\lambda x(\phi x \wedge \psi x)$, etc.

statements. In the following, 'classification' and 'observational theory' are used interchangeably. Notice that the lack of an explicit negation operator does not restrict the logical expressivity of observational statements. The reason is that an arbitrary universally quantified monadic predicate built by conjunction, disjunction, negation, and conditional from primitive predicates has a conjunctive normal form and is thus equivalent to a conjunction of observational statements.[5] For instance, $\phi \subseteq \neg\psi$ is equivalent to $\phi \wedge \psi \subseteq \Lambda$ and $\phi \wedge \neg\psi \subseteq \chi$ is equivalent to $\phi \subseteq \psi \vee \chi$.

A *Horn statement* over Σ is an observational statement $\phi \subseteq \psi$, where ϕ and ψ are free of disjunctions. In case ϕ and ψ belong to Σ, we speak of a *simple inheritance statement*. Statements of the form $\phi \subseteq \Lambda$ are referred to as *exclusion statements*. A *Horn theory* is an observational theory consisting solely of Horn statements. Similarly we speak of *simple inheritance theories*, etc.

2.2 Interpretations and models

Interpretations and models of observational theories are defined as usual in standard (first-order) predicate logic: a (set-valued) *interpretation* of Σ consists of a *universe* U and an *interpretation function* M from Σ to $\wp(U)$. The interpretation M can be inductively extended to $\mathrm{T}[\Sigma]$ by $M(\mathrm{V}) = U$, $M(\Lambda) = \varnothing$,

$$M(\phi \wedge \psi) = M(\phi) \cap M(\psi), \quad M(\phi \vee \psi) = M(\phi) \cup M(\psi).$$

An interpretation M *models* a statement $\phi \subseteq \psi$ iff $M(\phi) \subseteq M(\psi)$. A *model* of an observational theory Γ over Σ is an interpretation that models each of the statements of Γ. An interpretation corresponds to a *satisfaction relation* $\models_M$ from U to Σ, which can be extended to one from U to $\mathrm{T}[\Sigma]$ such that $x \models \phi$ iff $x \in M(\phi)$. The set $M(\phi)$ is called the *extent* of ϕ.

Consider an interpretation of Σ with universe U and satisfaction relation $\models$. Given two members x and y of U we say that x *is specialized by* y (notation: $x \sqsubseteq y$) if y satisfies every member of Σ that is satisfied by x. It follows by term induction that

$$x \sqsubseteq y \qquad \text{iff} \qquad \forall \phi \in \mathrm{T}[\Sigma]\,(x \models \phi \to y \models \phi). \tag{1}$$

If the specialization preorder $\sqsubseteq$ is antisymmetric and thus a partial ordering, we say that the interpretation satisfies *identity of indiscernibles*. (For the sake of the conventions of Formal Concept Analysis, more special elements will be graphically depicted *below* less special ones.)

2.3 The canonical universe

There is a standard way to associate with each observational theory Γ over Σ a *canonical model* $\mathrm{M}(\Gamma)$ of Γ. Its universe $\mathrm{C}(\Gamma)$ consists of the Γ-*closed, consistent subsets* of Σ, which are specified as follows: for each $X \subseteq \Sigma$ and $p \in \Sigma$ define $X \models p$ iff $p \in X$; extend $\models$ inductively to $\mathrm{T}[\Sigma]$, i.e. $X \models \mathrm{V}$ always, $X \models \Lambda$ never, $X \models \phi \wedge \psi$ iff $X \models \phi$ and $X \models \psi$, and $X \models \phi \vee \psi$ iff $X \models \phi$ or $X \models \psi$. Now, let $\mathrm{C}(\Gamma)$ be the set of all X such that for every statement $\phi \subseteq \psi$ of Γ, if $X \models \phi$ then $X \models \psi$.[6] Specialization on $\mathrm{C}(\Gamma)$ is set inclusion and hence a partial order, which is easily seen to be directed complete.

[5] Essentially the same observation is made in [5].

[6] Alternatively, one can take the set of Γ-models with values in $\mathbf{2} = \{0, 1\}$, where $\mathbf{2}$-valued interpretations and models are defined as in standard propositional logic: a $\mathbf{2}$-valued interpretation v of Σ is a model of Γ iff $v(\varphi) \leq v(\psi)$ for every statement $\varphi \subseteq \psi$ of Γ.

For each interpretation M of Σ with universe U let ε_M be the function from U to $\wp(\Sigma)$ such that

$$\varepsilon_M(x) \;=\; \{p \in \Sigma \,|\, x \models_M p\}. \tag{2}$$

By definition of specialization, $x \sqsubseteq y$ iff $\varepsilon_M(x) \subseteq \varepsilon_M(y)$. So ε_M is an order embedding of U into $\wp(\Sigma)$ if M satisfies identity of indiscernibles. Moreover, it follows by term induction that

$$x \models_M \phi \qquad \text{iff} \qquad \varepsilon_M(x) \models \phi. \tag{3}$$

Consequently, if M is a model of an observational theory Γ then ε_M is a homomorphism of models from M to $\mathrm{M}(\Gamma)$. The canonical model is thus the "largest" Γ-model satisfying identity of indiscernibles in the sense that every other such model M is embedded in $\mathrm{M}(\Gamma)$ via ε_M.

Depending on the class of Γ, the canonical universe $\mathrm{C}(\Gamma)$ can be characterized as a subset system as follows:[7]

Theorem 1 *If an observational theory Γ over Σ belongs to one of the classes listed on the left of Table 1 then its canonical universe $\mathrm{C}(\Gamma)$ is closed with respect to the properties listed in the same row on the right. Conversely, if a subset system $\mathcal{U}$ over Σ has closure properties that are listed in the right column then $\mathcal{U}$ is the canonical universe of an observational theory over Σ of the corresponding class on the left.*

Class of Γ	Closure properties of $\mathrm{C}(\Gamma)$
observational	local membership
Horn	nonempty intersection + directed union
Λ-free Horn	intersection + directed union
simple inheritance	intersection + union
exclusion	subsets + finitely bounded union
simple inheritance + exclusion	nonempty intersection + finitely bounded union

Table 1. Relationship between Γ and $\mathrm{C}(\Gamma)$

The reader is referred to [13] for a proof, where in addition an order-theoretic characterization can be found. (For instance, the subset systems that are closed with respect to nonempty intersection and directed union, also known as *inductive intersection systems*, correspond to the *bounded-complete algebraic dcpos*, or *Scott domains*, for short.)

Given class $\mathcal{C}$ of observational statements over Σ (e.g. the class of Horn statements) and a subset system $\mathcal{U}$ over Σ, let $\Gamma_{\mathcal{C}}(\mathcal{U})$ be the set of all $\mathcal{C}$-statements $\phi \subseteq \psi$ such that

$$\forall X \in \mathcal{U}\,(X \models \phi \to X \models \psi).$$

[7] A subset system $\mathcal{U}$ over Σ is *locally closed* if it contains every subset X of Σ which is *locally a member of* $\mathcal{U}$ in the sense that for every finite subset F of Σ there is a member Y of $\mathcal{U}$ such that $X \cap F = Y \cap F$. *Directed union* is short *union of (upwards) directed subsets*, and *finitely bounded union* means *union of subsets whose every finite subset has an upper bound*. Notice that every *finite* subset system is locally closed and closed with respect to directed union.

We call $\Gamma_{\mathcal{C}}(\mathcal{U})$ the *canonical $\mathcal{C}$-theory* associated with $\mathcal{U}$. The theory $\Gamma_{\mathcal{C}}(\mathcal{U})$ is of course highly redundant since it is closed with respect to entailment. (See [5] for the definition of a *nonredundant basis* of a theory.)

Let us say that $\mathcal{U}$ is *$\mathcal{C}$-definable* if $\mathcal{U}$ is the canonical universe of a $\mathcal{C}$-theory (which is the case, for instance, if $\mathcal{C}$ is the class of Horn statements and $\mathcal{U}$ is an inductive intersection system). It is easy to see that $\mathcal{U}$ is $\mathcal{C}$-definable just in case $\mathcal{U} = \mathrm{C}(\Gamma_{\mathcal{C}}(\mathcal{U}))$. In general, $\Gamma_{\mathcal{C}}(\mathcal{U})$ is the *least $\mathcal{C}$-definable subset system containing $\mathcal{U}$*. Consequently, by Theorem 1:

Theorem 2 *The canonical universe of $\Gamma_{\mathcal{C}}(\mathcal{U})$ is the closure of $\mathcal{U}$ with respect to the properties of Table 1 that correspond to class $\mathcal{C}$.*

3 FORMAL CONCEPT ANALYSIS

3.1 Complete theories of formal contexts

Consider the situation that a certain set U of objects is classified with respect to a set Σ of properties (or attributes). In other words, we are given a satisfaction relation $\models$ from U to Σ, i.e. an interpretation function M from Σ to $\wp(U)$. In the terminology of *Formal Concept Analysis* ([7]), the triple $\langle U, \Sigma, \models \rangle$ is called a *formal context*.[8]

Given a formal context one can ask for a theory that explains the data. To make this precise, we need to fix the type of theory we are interested in. For example, one can ask for a simple inheritance theory with or without exclusions, a Horn theory with or without Λ, or an observational theory in general.

Let $\mathcal{C}$ be a class of observational statements over Σ. We call a $\mathcal{C}$-theory Γ a *complete $\mathcal{C}$-theory of M* if, first, every statement of Γ is true with respect to M, i.e. if M is a model of Γ, and, second, if Γ entails every $\mathcal{C}$-statement that holds in M, that is, if for all $(\phi \subseteq \psi) \in \mathcal{C}$,

$$\text{if} \quad M(\phi) \subseteq M(\psi) \quad \text{then} \quad \Gamma \vdash \phi \subseteq \psi.$$

It is an immediate consequence of definitions that a complete $\mathcal{C}$-theory of M is unique up to logical equivalence. Moreover, there is a trivial way to get a complete theory: take the set $\Gamma_{\mathcal{C},M}$ of all $\mathcal{C}$-statements that are true with respect to M:

$$\Gamma_{\mathcal{C},M} \;=\; \{(\phi \subseteq \psi) \in \mathcal{C} \mid M(\phi) \subseteq M(\psi)\}.$$

Let us explore more closely the relation between a given formal context and the canonical universe of its complete $\mathcal{C}$-theory. As shown in Section 2, a formal context, i.e. a satisfaction relation $\models$ from U to Σ, determines a specialization relation $\sqsubseteq$ on U; see (1). In addition, the (pre)order-preserving function ε_M from U to $\wp(\Sigma)$ defined by (2) takes $x \in U$ to $\{p \in \Sigma \mid x \models p\}$. Let $\mathcal{U}_M$ be the image $\{\varepsilon_M(x) \mid x \in U\}$ of ε_M. In general ε_M is not one-to-one because there is no guarantee of *identity of indiscernibles*, i.e. different elements of U may satisfy exactly the same members of Σ. We have $\mathcal{U}_M \simeq U/\!\sim$ instead, with $x \sim y$ iff $\varepsilon_M(x) = \varepsilon_M(y)$.

Now notice that the canonical $\mathcal{C}$-theory $\Gamma_{\mathcal{C}}(\mathcal{U}_M)$ associated with $\mathcal{U}_M$ coincides with $\Gamma_{\mathcal{C},M}$; for by (3), $\varepsilon_M(x) \models \phi$ iff $x \models \phi$. So we can apply Theorem 2 to characterize the canonical universe of a complete $\mathcal{C}$-theory of M. For instance, if Γ is a complete Horn theory of M then $\mathrm{C}(\Gamma)$ is the closure of $\mathcal{U}_M$ with respect to nonempty intersection and directed union; similarly, if Γ is a complete simple inheritance theory of M then $\mathrm{C}(\Gamma)$ is the closure of $\mathcal{U}_M$ with respect to intersection and union.

[8] Beware, $\langle U, \Sigma, \models \rangle$ is called a *classification* in [1].

Example: Let Σ be $\{a, b, c, d, e\}$. Suppose U consists of the seven elements $x_1, x_2, \ldots, x_7$ which are classified according to the table of Figure 2. In addition, the figure shows the specialization order on

	a	b	c	d	e
x_1	X	X			X
x_2	X				
x_3		X			
x_4	X	X	X	X	
x_5	X		X		
x_6	X	X	X	X	
x_7	X		X	X	

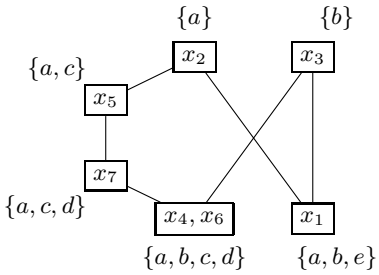


Figure 2. Classification table and induced specialization order

$U/\!\sim$ induced by the given formal context (where x_4 and x_6 are indiscernible, i.e. $x_4 \sim x_6$), as well as the corresponding subset system $\mathcal{U}_M$ over Σ. Figure 3 provides an overview of the canonical universes of several complete $\mathcal{C}$-theories of M, with varying $\mathcal{C}$. At the top of the figure there is the canonical universe of a complete simple inheritance theory of M; it is the closure of $\mathcal{U}_M$ with respect to intersection and union. A (nonredundant) complete simple inheritance theory of M is given by the statements $d \subseteq c$, $c \subseteq a$, $e \subseteq a$, and $e \subseteq b$. The diagram below the top on the left depicts the closure of $\mathcal{U}_M$ with respect to intersection of nonempty subsets and union of bounded subsets. It is the canonical universe of the extension of the above simple inheritance theory by the exclusion statement $c \wedge e \subseteq \Lambda$. Addition of the Horn statement $b \wedge c \subseteq d$ further weakens the closure properties of the associated canonical universe. If the statement $b \wedge c \subseteq d$ is added to the simple inheritance theory before the exclusion statement $c \wedge e \subseteq \Lambda$, the resulting effect on the respective canonical universes is as depicted by the right branch of Figure 3. Finally, adding the statements $\mathrm{V} \subseteq a \vee b$ and $a \wedge b \subseteq c \vee e$ leads to a complete observational theory of M, whose canonical universe consequently is $\mathcal{U}_M$.

From the viewpoint of *machine learning* (e.g. [11]), the problem of inducing theories from formal contexts can be characterized as follows: The $\mathcal{C}$-theories consitute the *hypothesis space* $\mathcal{H}$ of the learning problem, whereas the *version space* with respect to $\mathcal{H}$ and M consists of all $\mathcal{C}$-theories with model M. The commitment to statement type $\mathcal{C}$ determines the *inductive bias*: one can fit the data only as well as $\mathcal{C}$ permits. On the other hand, if $\mathcal{C}$ is too expressive, *overfitting* can occur: the induced theory explains the given data perfectly but does not allow generalizations. See Section 4.2 for a more thorough discussion of this problem.

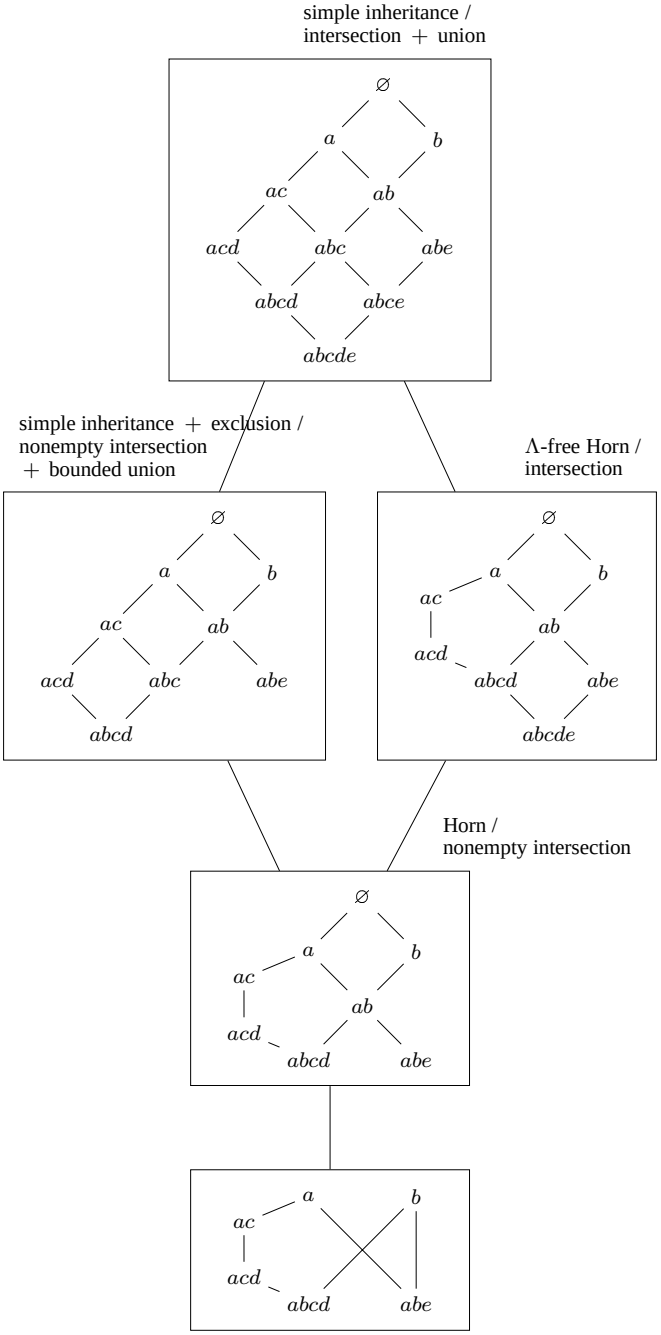


Figure 3. Canonical universes of complete $\mathcal{C}$-theories of formal context

3.2 Concept lattices

Formal Concept Analysis associates with each formal context a (complete) lattice of formal concepts. A *formal concept* of a context $\langle U, \Sigma, \vDash \rangle$ is a pair $\langle V, X \rangle$ consisting of a set $V \subseteq U$ of objects (the *extent*) and a set $X \subseteq \Sigma$ of attributes (the *intent*) such that X is the set of those attributes that are shared by all objects of V, whereas V consists of all objects that have all attributes of X. So $\langle V, X \rangle$ is a formal concept just in case $V^{\blacktriangleright} = X$ and $X^{\blacktriangleleft} = V$, where

$$V^{\blacktriangleright} = \{p \in \Sigma \mid \forall x \in V (x \vDash p)\} = \bigcap \{\varepsilon_M(x) \mid x \in V\},$$
$$X^{\blacktriangleleft} = \{x \in U \mid \forall p \in X (x \vDash p)\} = \bigcap \{M(p) \mid p \in X\},$$

and M is the interpretation function associated with the formal context. Clearly $\langle (V^{\blacktriangleright})^{\blacktriangleleft}, V^{\blacktriangleright} \rangle$ is a formal concept for each $V \subseteq U$. Furthermore, every formal concept of the context is of the form $\langle (V^{\blacktriangleright})^{\blacktriangleleft}, V^{\blacktriangleright} \rangle$. The set of all formal concepts is partially ordered by the *subconcept-superconcept-relation* $\leq$, which is defined as follows:

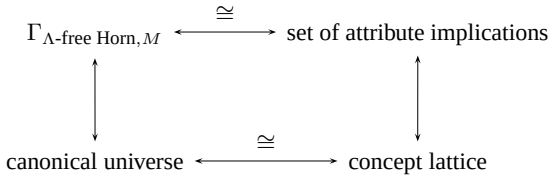
$$\langle V_1, X_1 \rangle \leq \langle V_2, X_2 \rangle \quad \text{iff} \quad V_1 \subseteq V_2 \quad \text{iff} \quad X_1 \supseteq X_2.$$

The set of formal concepts ordered by $\leq$ forms a complete lattice, the so-called *concept lattice* of the formal context.

By definition, the set $\{V^{\blacktriangleright} \mid V \subseteq U\}$ of intents is the closure of $\mathcal{U}_M = \{\varepsilon_M(x) \mid x \in U\}$ with respect to intersection. Since in the finite case, the set of intents is trivially closed with respect to directed union, it follows by Theorem 2:

Theorem 3 *The set of intents determined by a finite formal context is precisely the canonical universe of the complete Λ-free Horn theory of that context.*

The following diagram summarizes the correspondence between (finite) concept lattices and canonical universes of Λ-free Horn theories (see also [5], [6]).



4 INDUCTION OF HIERARCHICAL CLASSIFICATIONS FROM LINGUISTIC DATA

4.1 Applying Formal Concept Analysis to the induction of monotonic linguistic hierarchies

Modern linguistic theories regard linguistic knowledge to a large part as being lexical (e.g. HPSG [15], [18]). The lexicon is hierarchically structured in order to capture generalizations over linguistic entities. In general, these lexical hierarchies are constructed manually by linguists using linguistic knowledge and theory-driven hypotheses. However, in order to be independent of any specific theory, most of the linguistic databases contain purely unstructured data. For example, the lexical database CELEX, compiled by the Dutch Center for Lexical Information, consists of three large electronic databases and provides users with detailed English, German and Dutch lexical data. The German database, which serves us as a test database, holds 51.728 lemmas with 365.530 corresponding word forms.

The automatic induction of linguistic hierarchies is desirable both from a practical and a theoretical point of view. On the one hand, it makes the processing of large amounts of data possible and provides fast results. On the other hand, it is of theoretical interest to compare an automatically induced classification with existing linguistic descriptions, in order to reveal the linguistic assumptions made by the human experts. Furthermore, an automatically induced hierarchy can guide the linguist in analyzing new linguistic data. To this end, the induced hierarchy should exhibit as much of the implicitly given information as possible, and the original flat input data should always be reconstructible from the induced hierarchy. Formal Concept Analysis satisfies these demands. It has already been applied to the following linguistic areas: meronymy ([17]), WordNet ([16]), semantics of speech-act-verbs ([9]), and verb paradigms ([8]).

Table 2 shows a many-valued formal context based on CELEX, that models a part of the German nominal inflection. It contains the gender information and the inflectional paradigms of eight German nouns (*Herr* 'mister', *Name* 'name', *Staat* 'state', *Hemd* 'shirt', *Farbe* 'color', *Bett* 'bed', *Onkel* 'uncle', *Ufer* 'bank'/ 'shore').[9] If the features of this context are scaled with respect to the nominal scale (see [7]), one ends up with a one-valued context consisting of eight objects and 19 attributes. This will be our example context in the remaining part of the paper.

	gender	sing nom	sing gen	sing dat	sing acc	plur nom	plur gen	plur dat	plur acc
Herr	masc	*	*_n	*_n	*_n	*_n	*_n	*_n	*_n
Name	masc	*	*_ns	*_n	*_n	*_n	*_n	*_n	*_n
Staat	masc	*	*_s	*	*	*_n	*_n	*_n	*_n
Hemd	neut	*	*_s	*	*	*_n	*_n	*_n	*_n
Farbe	fem	*	*	*	*	*_n	*_n	*_n	*_n
Bett	neut	*	*_s	*	*	*_n	*_n	*_n	*_n
Onkel	masc	*	*_s	*	*	*	*	*_n	*
Ufer	neut	*	*_s	*	*	*	*	*_n	*

Table 2. Example data: derivational paradigms of eight German nouns

Figure 4 shows the concept lattice which corresponds to our example context. As usual, only the attribute and the object concepts are labeled.[10] The concept lattice represents a monotonic multiple inheritance hierarchy, where a node inherits all the attributes labeled to its supernodes.[11] Notice that conflicting attributes cannot be inherited, since the hierarchy is constructed on the base of the subset relation of concept intents.

Let us focus more closely on the four unlabeled nodes of the example concept lattice.[12] In an inheritance hierarchy nodes have essentially two roles: first, they can introduce new information, which will be inherited by subnodes and second, they "collect" information from their supernodes and transmit it "bundled up" to their subnodes. Unlabeled nodes are nodes which only perform information bundling and not information introduction. Nodes that do not bundle up information are necessarily labeled. (They are $\wedge$-irreducible in that they have less than two direct upper neighbors.) Altering the hierarchy by varying the underlying theory which models the data of the context changes the proportion between the information introducing and the information bundling nodes.

4.2 Extensions and restrictions of concept lattices

Among the different hierarchical representations of a given data set there is none which is optimal in every respect. Rather, the question is to find the most appropriate representation depending on the task for which the hierarchy is built. Two criteria must be met by any reasonable representation: it must be complete and consistent with respect to the data. Furthermore, a good representation is maximally informative, maximally compact, and avoids redundancies by capturing generalizations. Unfortunately, it is not possible to construct an inheritance hierarchy which is optimal with respect to each of these criteria.

What does it mean to say that an hierarchical network is maximally informative? In principle, every hierarchy which is consistent and complete with respect to the data is equally informative in the sense that the original context can be reconstructed from the hierarchy. But consider the hierarchy in Figure 5, which corresponds to the complete observational theory of the example context: Since two of the objects in the example are either indiscernible or incommensurable, the hierarchy is *flat*; only the fact that indiscernible objects are merged discriminates this representation from the one in Table 2. For the observer the flat hierarchy is less informative than the concept lattice, although from the viewpoint of the underlying theories, the Horn theory is a subtheory of the observational theory and therefore less informative. Since we are interested in the induction of hierarchical representations, we record that hierarchies differ with regard to the amount of information they exhibit explicitly. If the hierarchy is designed to be viewed by human beings it should maximize this amount of information.

The compactness of a network can be measured in several respects, but in what follows we will only look at the number of nodes. The compactness criterion clearly favors the network of the observational theory.

A good representation avoids redundancy by capturing generalizations. In the flat hierarchy determined by the complete observational theory (see Figure 5) no generalizations are captured and therefore, several attributes have to be stated more than once (e.g. "gender:masc"). In other words, the complete observational theory leads to "overfitting". In the concept lattice (see Figure 4) all generalizations are captured and every attribute and every object occurs exactly once; such a representation is said to be free of redundancy.

Is there any representation that has this desirable property but is more compact than the concept lattice? It follows from Section 2.3 that such a representation is the canonical universe of an extension of a complete Λ-free Horn theory describing the concept lattice by disjunctive rules that are consistent with respect to the data. Recalling the two different roles of nodes in inheritance hierarchies we can dispense with the four nodes which only bundle up information. This results in the inheritance network in Figure 6, which is the partially ordered set of the attribute and object concepts (*AOC-poset*), bounded

[9] '*' represents the root of the derived word form. For example, if the feature "sing dat" has the value "*_n" at the object "Name", that means that the singular dative form of *Name* is *Namen*. As usual, the unstressed vowel (the so-called *schwa*) is disregarded since its occurrence is determined by phonological rules.

[10] The *attribute concept* associated with an attribute p is the greatest concept whose intent contains p and the *object concept* of an object x is the smallest concept whose extent contains x.

[11] An inheritance hierarchy is said to be multiple if it is not excluded that a node has more than one supernode from which it inherits.

[12] We will disregard the unlabeled bottom node.

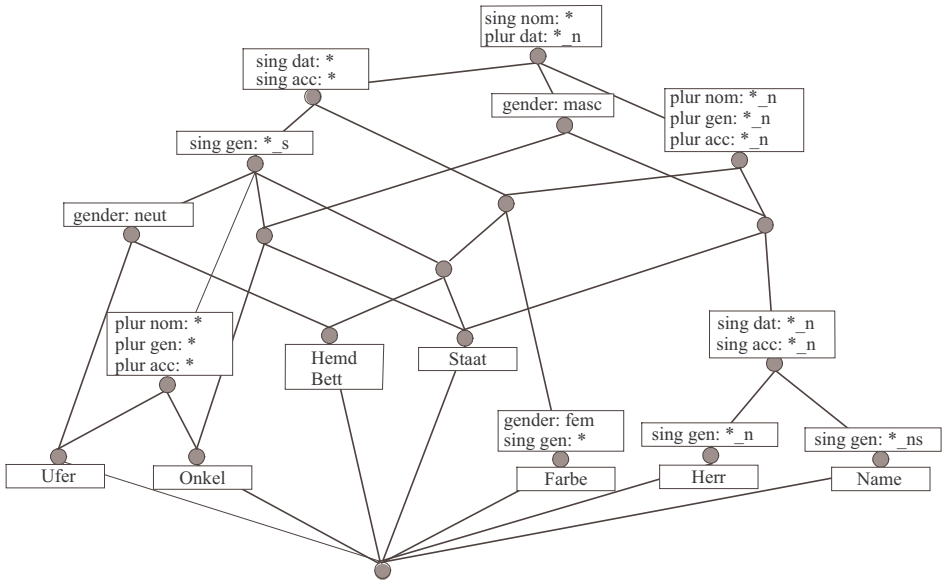


Figure 4. Concept lattice corresponding to Table 2

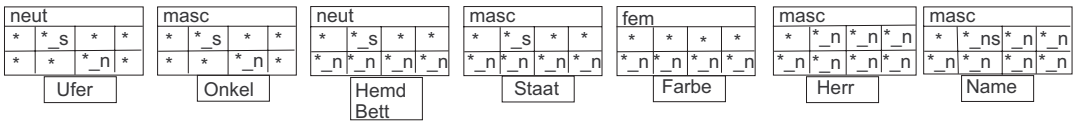


Figure 5. Generic universe of the complete observational theory

by a top and a bottom node. (Notice that in the case of a reduced context the AOC-poset consists exactly of the $\wedge$-irreducible and the $\vee$-irreducible nodes of the concept lattice.) The theory describing the AOC-poset consists of all the rules from the Horn theory corresponding to the concept lattice and an additional disjunctive rule for each concept which is neither an object nor an attribute concept. These rules are obtained in the following way: For each concept node in question, the conjunction of its intent forms the premise, and the disjunction of the intents of its subconcepts without the intent of the concept itself forms the conclusion of the new rule. In our example, one has to add the following rules in order to get the AOC-poset:

sing gen:*_s $\wedge$ plur nom: *_n		
$\wedge$ sing dat: *	$\subseteq$	gender: neut $\vee$ gender: masc
sing gen:*_s $\wedge$ gender: masc	$\subseteq$	plur nom:* $\vee$ plur nom:*_n
sing dat:* $\wedge$ plur nom: *_n	$\subseteq$	gender: fem $\vee$ sing gen: *_s
plur nom:*_n $\wedge$ gender: masc	$\subseteq$	sing dat:*_n $\vee$ sing gen: *_s

Compared to the concept lattice, the AOC-poset is more compact and also free of redundancy. But it is not as informative as the concept lattice, since the information about common attributes is not captured in single nodes anymore. In the worst case, the AOC-poset has only four levels: the top and the bottom nodes, the level of the attribute nodes, and the level of the object nodes. This happens if, first, all objects intents and, second, all attribute extents are pairwise incomparable with respect to set inclusion. Nevertheless, the AOC-poset is more informative than the complete observational theory since it simplifies the access to the information to which objects an attribute applies and it shows the hierarchical relations between the attributes.

The number of nodes in an AOC-poset is bounded by the sum of the number of attributes and the number of objects plus two. In realistic data sets the difference in compactness between AOC-posets and concept lattices can be dramatic. For instance, the number of nodes in the concept lattice capturing the derivational information of German lemmas contained in the lexical database CELEX is greater than 72.000, whereas the number of nodes in the corresponding AOC-poset is less than 4.000. (The underlying formal context consists of 9.567 objects and 2.032 attributes.) Hence, switching to the AOC-poset reduces the memory requirements. Moreover, since the AOC-poset is just the partial order of the attribute and object concepts, there is an efficient construction algorithm. To summarize, compared to concept lattices, AOC-posets provide a very simple method to induce redundancy-free inheritance hierarchies from huge databases. Inference tasks, however, are better supported by concept lattices, due to the explicit representation of shared attributes.

Having discussed the case of adding rules to a complete Horn theory, it remains to consider the omission of rules. Switching to the complete simple inheritance theory without exclusions seems to be overdone, because for the example context of Table 2 the resulting lattice has 78 concepts. Since the attributes of the example are feature-value pairs, where the values of each feature are incompatible, it makes sense to take the complete simple inheritance theory with exclusions instead. The corresponding hierarchy has 21 elements, witness Figure 7, and is hence less compact than the concept lattice. The simple inheritance theory is weaker than the one describing the AOC-poset or the concept lattice; it is thus more likely that a new object can be inserted without serious changes to the structure of the lattice.

4.3 Classification trees

All hierarchical representations presented so far make use of multiple inheritance, whereas in traditional linguistics, hierarchical classifications are usually presented in form of taxonomic trees. In modern linguistic theories, multiple inheritance is included but in general restricted to special cases like multi-dimensional inheritance (e.g. HPSG, [15]). Even in these approaches, tree-like hierarchies play a prominent role. Their characteristic property is that the subclassification at each node is based on the different values of a single feature.

Let us briefly indicate by example how to "cut out" classification trees of this type from concept lattices by adding disjunctive rules. Figure 8 shows such a tree for the data of Table 2. It first classifies the nouns with regard to their gender, which defines the full inflectional paradigm of the feminine nouns. The neuter nouns are then further classified according to their plural marking strategy, while the masculine nouns are first specified with respect to their singular accusative forms and then their plural and their singular genitive forms respectively. The tree of Figure 8 is the canonical universe of the theory consisting of (a) the conjunctive statements corresponding to the concept lattice, (b) exclusionary statements which ensure the incompatibility of the feature values, and (c) the following disjunctive rules:

sing acc:*	$\subseteq$	gender: fem $\vee$ gender: masc $\vee$ gender: neut
plur nom:*_n	$\subseteq$	sing acc:* $\vee$ sing acc:*_n

Notice that the conclusions of these rules specify the selectable values for a single feature.

Of course, the form of the classification tree is not determined by the given formal context. Figure 9 shows a different classification tree, where the sorting decisions are done in another order. It has one node less than the first tree, since, after choosing the singular accusative form of the masculine nouns, there are only two nouns left which have to be further distinguished by fixing their plural forms. In order to get this tree, one can employ the following disjunctive rules:

sing acc:*	$\subseteq$	gender: fem $\vee$ gender: masc $\vee$ gender: neut
plur nom:*_n	$\subseteq$	sing gen:* $\vee$ sing gen:*_s $\vee$ sing gen:*_n $\vee$ sing gen:*_ns

The indeterminateness of classification trees is the main argument against them. But one should keep in mind that trees are much easier to read than multiple inheritance networks, because they do not have crossing lines. Therefore it would be interesting to have a system which allows to switch between different classification trees and the concept lattice or the AOC-poset.

Finally, it has to be emphasized that we do not propose to construct decision trees from concept lattices by adding rules to the underlying theory, because for inducing decision trees a lot of efficient tools are available. The purpose of presenting classification trees is solely to show that besides AOC-posets trees can also be characterized as an extension of concept lattices.

5 OUTLOOK

In addition to purely monotonic inheritance hierarchies, nonmonotonic approaches are becoming more and more important in linguistic theories (e.g. [4], [3]). In [14] one finds a first discussion of the

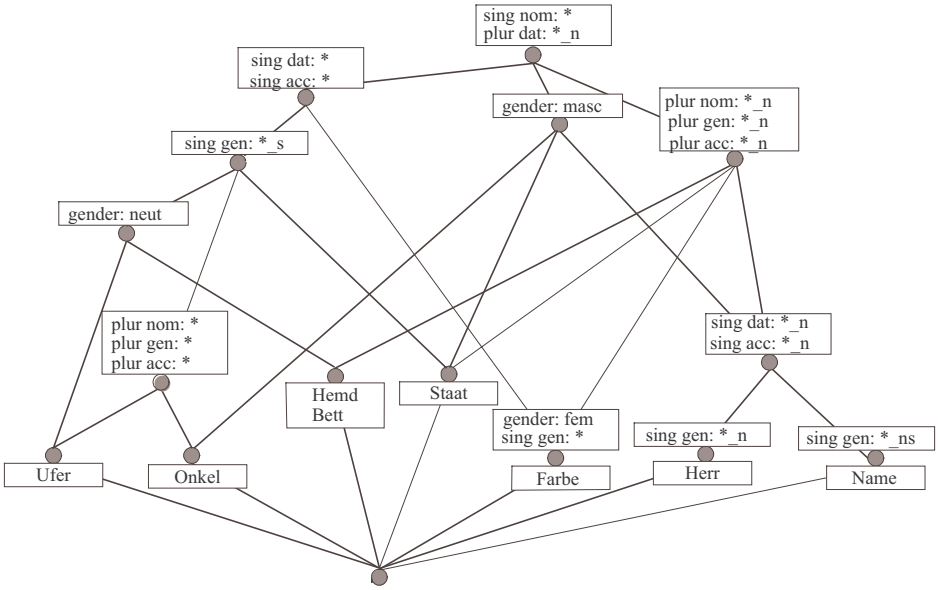


Figure 6. AOC-poset corresponding to Table 2

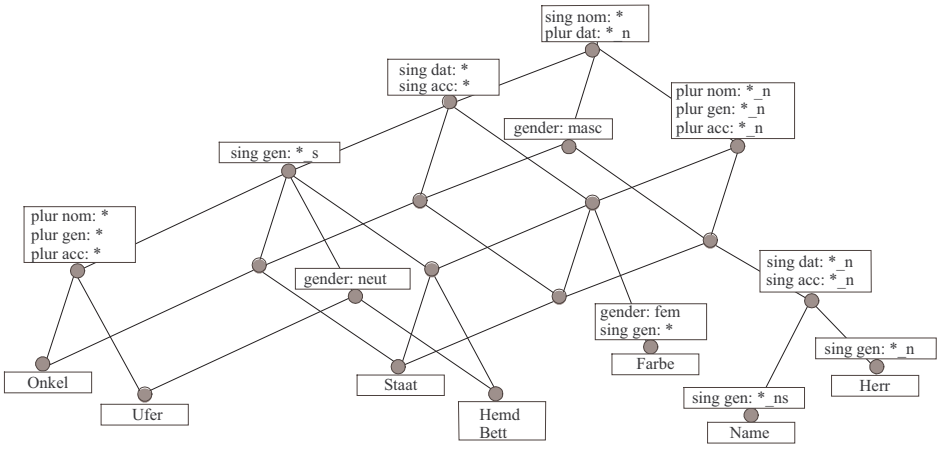


Figure 7. Simple inheritance hierarchy with exclusions

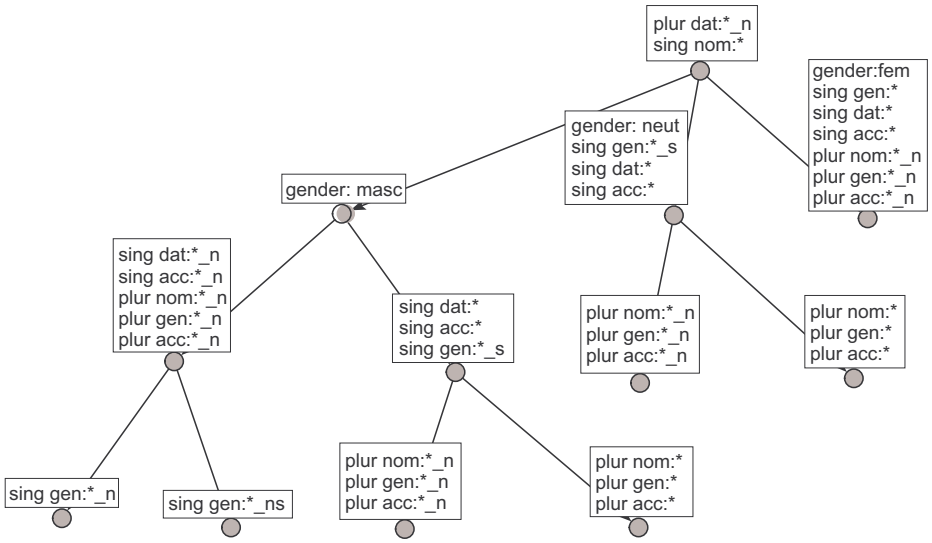


Figure 8. A possible classification tree

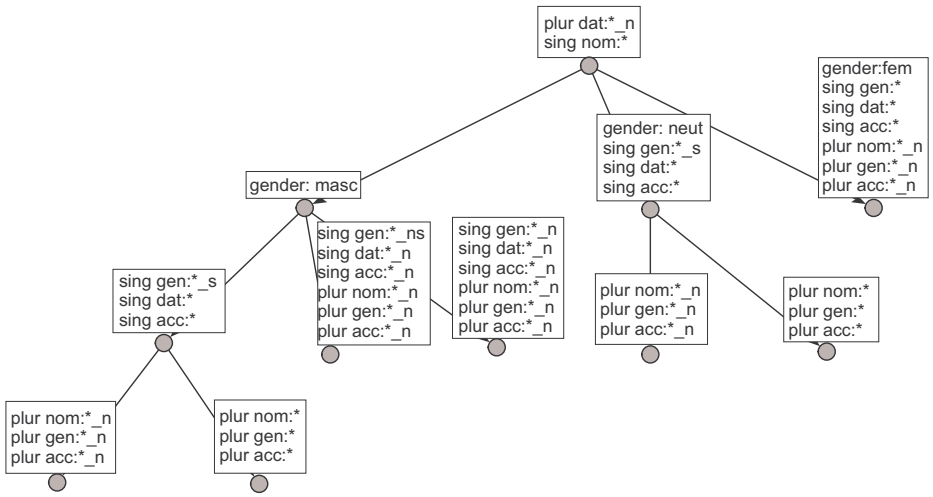


Figure 9. Another classification tree

potentials of applying Formal Concept Analysis to the induction of regularities, subregularities, and exceptions in order to obtain reasonable nonmonotonic inheritance hierarchies. This is a topic of current research.

Another possible application of the presented approach is to allow disjunctive rules in attribute exploration tasks. As discussed in Section 4.2, the problem is to avoid accepting too many disjunctive rules, since otherwise, in the case of incommensurable objects the exploration would always end in a flat hierarchy like that of Figure 5. One way to prevent this could be to introduce two steps: first, the standard attribute exploration is performed and second, each concept which is not yet an attribute or an object concept is tested to determine whether there is any object in the universe to which exactly the attributes of its intent apply. If so, the object is added to the context and if not, a disjunctive rule is added which excludes the concept from the canonical universe. In an exploration tool the concept could be tested by presenting the corresponding disjunctive rule (see Section 4.2) and asking if there is any known counter example.

Furthermore it would be interesting to explore possible ways to automatically shift from one theory to another, based on parameters like compactness monitored during incremental construction of the inheritance hierarchy.

REFERENCES

[1] Jon Barwise and Jerry Seligman, *Information Flow. The Logic of Distributed Systems*, Cambridge Tracts in Theoretical Computer Science 44, Cambridge University Press, Cambridge, 1997.

[2] Bob Carpenter and Carl Pollard, 'Inclusion, disjointness and choice: The logic of linguistic classification', in *Proceedings of the 29th Annual Meeting of the ACL*, pp. 9–16, (1991).

[3] Ann Copestake and Alex Lascarides, 'Default representation in constraint-based frameworks', *Computational Linguistics*, **25**(1), 55–105, (1999).

[4] Roger Evans and Gerald Gazdar, 'DATR: A language for lexical knowledge representation', *Computational Linguistics*, **22**(2), 167–216, (1996).

[5] Bernhard Ganter, 'Attribute exploration with background knowledge', *Theoretical Computer Science*, **217**(2), 215–233, (1999).

[6] Bernhard Ganter and Rüdiger Krausse, 'Pseudo models and propositional Horn inference', Technical Report MATH-AL-15-1999, Technische Universität Dresden, (1999).

[7] Bernhard Ganter and Rudolf Wille, *Formal Concept Analysis. Mathematical Foundations*, Springer, Berlin, 1999.

[8] Anja Großkopf, 'Formal concept analysis of verb paradigms in linguistics', in *Ordinal and Symbolic Data Analysis*, eds., E. Diday, Y. Lechevallier, and O. Opitz, 70–79, Springer, Berlin, (1996).

[9] Anja Großkopf and Gisela Harras, 'Begriffliche Erkundung semantischer Strukturen von Sprechaktverben', in *Begriffliche Wissensverarbeitung: Methoden und Anwendungen*, eds., Gerd Stumme and Rudolf Wille, 273–295, Springer, Berlin, (2000).

[10] Michael A. K. Halliday, *An Introduction to Functional Grammar*, Edward Arnold, London, 2 edn., 1994.

[11] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[12] Rainer Osswald, 'Classifying classification', in *Proceedings of the Joint Conference on Formal Grammar and Mathematics of Language (FG/MOL-01)*, eds., Geert-Jan Kruijff, Larry Moss, and Dick Oehrle, Electronic Notes in Theoretical Computer Science 53, (2001).

[13] Rainer Osswald, *A Logic of Classification – with Applications to Linguistic Theory*, Ph.D. dissertation, FernUniversität Hagen, Praktische Informatik VII, 2002.

[14] Wiebke Petersen, 'A set-theoretic approach for the induction of inheritance-hierachies', in *Proceedings of the Joint Conference on Formal Grammar and Mathematics of Language (FG/MOL-01)*, eds., Geert-Jan Kruijff, Larry Moss, and Dick Oehrle, Electronic Notes in Theoretical Computer Science 53, (2001).

[15] Carl Pollard and Ivan Sag, *Information-Based Syntax and Semantics, Vol. 1*, CSLI Lecture Notes, No. 13, CSLI Publications, Stanford, CA, 1987.

[16] Uta Priß, 'The formalization of WordNet by methods of relational concept analysis', in *WordNet: An Electronic Lexical Database and Some of its Applications*, ed., C. Fellbaum, 179–196, MIT-Press, (1998).

[17] Uta Priß, *Relational concept analysis: semantic structures in dictionary and lexical databases*, Shaker Verlag, Aachen, 1998.

[18] Ivan A. Sag and Thomas Wasow, *Syntactic Theory: A Formal Introduction*, CSLI Publications, Stanford, CA, 1999.

[19] Gerd Stumme, 'Distributive concept exploration – a knowledge acquisition tool in formal concept analysis', in *KI-98: Advances in Artificial Intelligence*, eds., Otthein Herzog and Andreas Günter, Lecture Notes in Artificial Intelligence 1504, pp. 117–128, Berlin, (1998). Springer.

[20] Steven Vickers, 'Topology via constructive logic', in *Logic, Language, and Computation, Vol. 2*, eds., Jonathan Ginzburg, Lawrence S. Moss, and Maarten de Rijke, CSLI Publications, Stanford, CA, (1999).

[21] Terry Winograd, *Language as a Cognitive Process: Syntax*, Addison-Wesley, Reading, MA, 1983.