Formal Conceptual Structure in Language

Janos J. Sarbo
University of Nijmegen, Toernooiveld 1
6525 ED Nijmegen, The Netherlands
e-mail: janos@cs.kun.nl

Abstract. Conceptual structures like conceptual graphs and concept lattices are formal representations of knowledge, e.g. structure or information content. In this paper, it is argued that, based on Peirce's semeiotic theory of signs and linguistic knowledge about syntactic clusters, conceptual structures can be derived in natural language automatically. This is incorporated in a novel model of language called Natural Language Concept Analysis (NLCA).
Keywords conceptual structure, Peirce, semeiotic, language analysis, NLCA.

I INTRODUCTION

Conceptual structures like conceptual graphs [13] and concept lattices [3] are formal representations of knowledge, e.g. structure or information content. However, concept lattices (CLs), which are yielded by formal concept analysis (FCA), have better computational properties, which make them more attractive for practical applications.

After some preliminary results with CLs [12], [11] we looked for some more realistic applications, for example, in natural language (NL). Finding such applications turned out not to be an easy task. This led us to re-examine the basics of the theory.

FCA and conceptual graphs (CGs) have their foundation in Peirce's doctrine of categories [14]. On the basis of this philosophical theory we found a more appropriate characterisation of the basic constituent of FCA, the formal context. According to the revised characterisation, most domains can be characterised by clusters, the source of which are relations. A suitable representation of the clusters yields the formal context, eventually. By this clarification of the genealogy of formal contexts a realistic application of knowledge representation in NL became viable.

Our interpretation of the formal context is based on the notion of a cluster. An explanation for the latter, in the case of NL, is found in Peirce's semeiotic. NL is a system of signs, and semeiotic is concerned with determining the formal conditions for signs as such [9]. By analysing the structure of signs, conditions for signs representing structure, in particular syntactic clusters were developed. This 'top-down' analysis of signs, and language as a system of signs, yields the specification of three types of syntactic sign. These signs find their implementation in a model, called Natural Language Concept Analysis (NLCA), introducing three relation schemes, one for each type of sign. The relation schemes are developed 'bottom-up' on the basis of our linguistic knowledge of syntactic clusters. Linking top-down and bottom-up developed specifications is common in most programming methodologies. In this sense, we can say that by specifying NLCA we are doing natural language *programming.*

II A GENEALOGY OF FORMAL CONTEXTS

Let us begin with a recapitulation of the basic definitions of FCA [14]. A *formal context* is a triple (O, A, R) where O is a set of (*formal*) *objects* and A is a set of (*formal*) *attributes*, while R is a binary relation between O and A indicating when an object o of O has an attribute a of A (the relationship oRa holds, or, equivalently, $(o, a) \in R$). For constructing formal concepts within a given formal context (O, A, R), the following operators are defined for $\mathsf{O} \subseteq O$ and $\mathsf{A} \subseteq A$: $\mathsf{A} \mapsto \mathsf{O}' := \{a \in A | oRa \text{ for all } o \in \mathsf{O}\}$, and $\mathsf{O} \mapsto \mathsf{A}' := \{o \in O | oRa \text{ for all } a \in \mathsf{A}\}$. A *formal concept* of a formal context (O, A, R) is defined as a pair (O, A) with $\mathsf{O} \subseteq O$, $\mathsf{A} \subseteq A$, $\mathsf{O} = \mathsf{A}'$ and $\mathsf{A} = \mathsf{O}'$. The ordering of formal concepts is defined by: $(\mathsf{O}_1, \mathsf{A}_1) \leq (\mathsf{O}_2, \mathsf{A}_2)$ iff $\mathsf{O}_1 \subseteq \mathsf{O}_2$ (or, equivalently $\mathsf{A}_1 \supseteq \mathsf{A}_2$). The set of formal concepts of a formal context together with the above defined order relation is called a *concept lattice*.

In the past years we developed various applications of formal concept analysis. In these, we made use of a simple extension of the theory that involves concept sublattices and their embedding relation [11]. In essence, the applications were based on the following scheme: 1) Specify a number of 'known worlds' and an 'unknown world' by their formal contexts and corresponding CLs; the CLs of the known worlds are called a knowledge base. 2) Show that the unknown world can be 'explained' by CLs of the knowledge base.

Most importantly, the specification of the unknown world does not have to contain references to the knowledge base. Technically, a solution to task (2) above involves the determination of an embedding relation on CLs. For example, in the application of [10] the knowledge base contained the CL of an angle (cf. example 1 below); the unknown world was given as a configuration of three intersecting lines. By identifying instances of the CL of the angle embedded in the CL of the unknown world it was shown that the unknown world is constituted from angles. Then, by abstraction, the unknown world was discovered as a triangle.

Example 1 The formal context of the angle is given by the sets: $O = \{e, f\}$, $A = \{0, 1, 2\}$, and $R = \{(e, 0), (e, 1), (f, 0), (f, 2)\}$. The concept lattice is shown in figure 1.

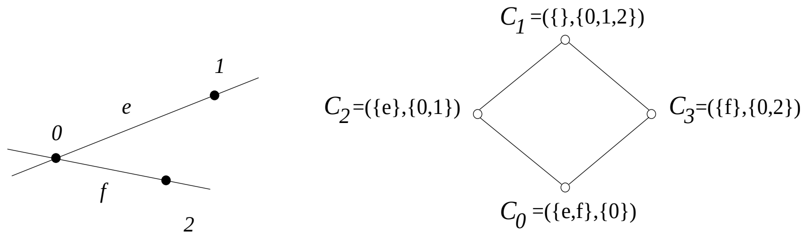


FIGURE 1. Diagram and concept lattice of the angle

In another application [12] it was shown that instances of the general notion of a card game (knowledge base) can be found in a detailed specification involving knowledge about various card games. Each such instance corresponds to particular classes of card games, e.g. to winning-oriented card games like poker etc.

From a logical point of view, each of these applications starts with a deduction step involving CL embedding, as described above. Then, a number of abduction steps follow. E.g. in the geometrical example, hypotheses are postulated and verified like 'the unknown world consists of angles' (i.e. the unknown world is some angular 'thing'); 'it contains three independent instances of the angle' (i.e. the unknown world is some countable 'thing'); and 'a cover of the CL of the unknown world is possible by CLs of the knowledge base' (i.e. the unknown world can be recognised as a single 'thing').

A Transparency

FCA aims at finding concepts and their hierarchy (i.e. a structure) in some domain. We may call such concepts and a structure *transparent* if they correspond with our intuitive conception of that domain. For example, the concepts of two different straight lines (C_2, C_3), and the intersection of these lines, or equivalently, the meet of these concepts (C_0), coincide with our conception of an angle, indeed.

In FCA, the formal context is assumed to be given. If for any reason the generated concepts and the CL are not transparent, then it is very difficult to know how the formal context should be adjusted such that the result will improve. So, an appropriate definition of the formal context is inevitably necessary.

B The inverse problem

Learning from the problems of interpretation of formal context, in our search for adequate examples we decided to take the unorthodox view of considering the inverse problem of FCA: assuming that a domain of concepts and structures is given, find the corresponding formal context. That the inverse problem is solvable follows from the fundamental theorem of FCA theory [3].

An appropriate domain of concepts and structures is found in natural language, in the domain of syntactic structure.[1] In the case of the inverse problem however, FCA theory is not helpful in determining the constituents of the formal context, except for some indications with respect to their size. Fortunately, linguistic modelling suggests that syntactic structure, in particular *syntactic clusters*, can be derived from the constituent lexical items based on their relations. This indicates that NL is not fundamentally different from other application domains of FCA.

So, at least we know where to start. But then, the question arises immediately: which lexical items will be objects, and which attributes? A simple answer does not seem to exist because a lexical item can have different roles in a sentence, context dependently. Although FCA theory, at least to our knowledge, does not treat the problem of dichotomisation (objects and attributes are given 'somehow'), it emphasises its strong relationship with Peirce's doctrine of categories [8], [14]. In this paper, we show that the philosophical theory can help to find a solution for dichotomisation, indeed.

C Peirce's categories

In its mature form, Peirce's doctrine of categories [9] states that all phenomena present three aspects, which are irreducible. The aspect of firstness is the aspect in virtue of which each phenomenon has an absolutely novel *quality*, unrelated to anything whatever. The aspect of secondness is the aspect in virtue of which each phenomenon involves an *interaction* (a relation, an event etc.). The aspect of thirdness is the aspect in virtue of which each phenomenon involves some *habit* (lawfulness, reasonableness, meaning etc.).

Though secondness cannot be reduced to firstness, it presupposes firstness, and, similarly, though thirdness cannot be reduced to either firstness or secondness, it presupposes both firstness (through secondness) and secondness. Conversely, the element of firstness remains a mere possible, unless it be actualised by some interaction; and the element of secondness remains brutal interaction unless it derives its meaning from thirdness. The important point here is that the categories are related to each other according to a relation of *subservience*. Thus, the categories must always be considered not only in themselves, but as they are relative to one another.

D Clusters and structure

We experience clusters of a domain as relations, as interactions between a group of items. In terms of Peirce's categories, clusters involve the category of secondness. But the mere fact that things are related says nothing whatever about the meaning of the relation. The latter aspect is the aspect of thirdness. If we abstract from the event aspect of clusters (cf. section III A), then the meaning of a cluster requires that the cluster be *embedded* in a class of clusters which are related to each other. Such web of related clusters is what is called a *structure*.

E Formal contexts revisited

In [8] it is argued that in a semantic understanding of *predication* the role of objects in the formal context correspond to the category of firstness, and that of the attributes to the category of secondness. We find this conclusion slightly counterintuitive, as the word 'attribute' denotes "a *quality* which is considered as belonging to, or inherent in, a person or thing" (Webster Dictionary, 1996).

But, why dichotomise in this way? Isn't it more plausible to define the formal context as a tuple consisting of a set A of all primary observations having the element of firstness (this corresponds, in terms of the traditional interpretation, to the set of attributes *and* (the firstness aspects of) objects), a set O of secondary observations having the element of secondness (this corresponds to a set of clusters, the origin of which is clarified below), and a relation R linking A and O (having the element of thirdness, potentially).

Unfortunately, such a definition is not always possible, because R, in general, is a relation on $A \cup O$. This implies that the formal context has to be derived from R, on the basis of the A and O sets. An additional

---

[1] Here, we assume that those concepts and structures are formal concepts and CLs, respectively. It appears later that this is not always the case.

problem is that clusters (elements of O) are *not* necessarily embedded in a class of other clusters, therefore clusters need not have meaning, and need not represent a structure, in general. A CL is, however, a structure (formal concepts are embedded according to their ordering), and formal concepts do have (formal) meaning. This shows that the CL discloses 'knowledge' hidden in the specification. We cannot be sure, however, that this knowledge meets our intuitive conception of the 'world' specified as a set of clusters.

We 'see' clusters via relation schemes ('knowledge base'), as *instances* of those schemes. Such instances are *subrelations* of R, the labels of which are elements of O (the label of a subrelation corresponds with a *rule* defining it). In this sense, relation schemes are a second form of knowledge, besides CLs. The difference between the two lies in their role: a CL represents (formal) knowledge yielded by FCA; a relation scheme generates input for FCA.

Example 2 The relation scheme of a point-blank (i.e. a straight line), can be defined as a configuration of two or more points and a line going through them. Instances of the scheme are the subrelations between line e, points 0 and 1, and, between line f, points 0 and 2 (cf. figure 1). Furthermore, if we assume that those instances are related by some other relation scheme and form a cluster (by a sort of 'and' coordination scheme, not specified here), eventually we get the relation (cf. figure 2): $A = \{0, 1, 2, e, f\}$, $O = \{01\text{e}, 02\text{f}, 012\text{ef}\}$, and $R = \{(01\text{e}, 0), (01\text{e}, 1), (01\text{e}, e), (02\text{f}, 0), (02\text{f}, 2), (02\text{f}, f), (012\text{ef}, 01\text{e}), (012\text{ef}, 02\text{f})\}$.

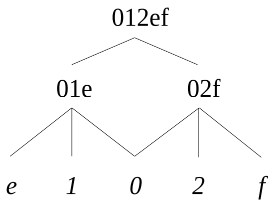


FIGURE 2. The relation of the angle

F From clusters to formal contexts

Clusters are relations; the set of clusters defines a graph $G = (V, E)$, where $V = A \cup O$ and $e_{ij} \in E$ iff $(v_i, v_j) \in R$. In the case of FCA we require (1) that a partitioning of V exists, $\pi = \{p_1, \ldots, p_k\}$, $0 < k \leq |V|$, (2) π is such that for each $a \in A$ there is a partition p_i $(0 < i \leq k)$ containing a, but no other elements of A, and (3) the graph $G' = (\pi, E')$, where $e_{rs} \in E'$ iff there are vertices $v_i \in p_r$, $v_j \in p_s$ $(p_r, p_s \in \pi)$, and an edge $e_{ij} \in E$, is a *bipartite* graph.

G', which is isomorphic to the formal context, induces two partitions that correspond to the attribute and object sets, in terms of the traditional interpretation. If π is such that the number of partitions containing a single element of O is maximal, then the generated concepts will be more transparent, and the CL more easily interpreted.

The partition π can be derived stepwise by merging vertices of G. The merging of a set of vertices $\mathsf{V} \subseteq V$ involves the deletion of the edges between vertices of V, and the generation of a new vertex which is labelled v, if there is a non-empty subset v of V which is a subset of A; and otherwise it is labelled V. The labelling of vertices by subsets of A can increase the transparency of formal concepts, especially if the elements of A involved denote 'major' information about clusters.

Example 3 A derivation of π for the clusters of figure 2 is shown in figure 3. We assume that e and f represent major information about 01e and 02f, resp. First, the vertices 01e and e, and 02f and f are merged, yielding new vertices labelled e and f, respectively; then, these vertices and 012ef are merged yielding a vertex labelled ef.

The formal context corresponding to the relation yielded is defined by the sets: $A = \{0, 1, 2\}$, $O = \{\text{e}, \text{f}, \text{ef}\}$, and R as shown in figure 3 (on the right hand side). The CL for this formal context is isomorphic to the one of example 1. This substantiates our claim that relation schemes are present in FCA, possibly without the user being aware of it.

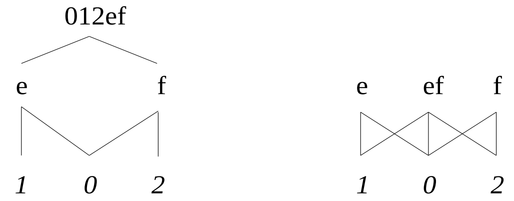


FIGURE 3. A partitioning of the relation of the angle

G Clusters and events

Clusters have the element of an event. The dynamic aspect of clusters may not be clear from the example of the angle, because its 'world' is given as a fact. In our experience, any non-trivial 'world' is perceived incrementally. We can clarify this by introducing a variant of our running example.

Assume that we observe 0, 1, e, 2, f incrementally, in this order, in an empty world. When the points 0, and later 1, appear, we 'see' nothing except for a potential for something to become a straight line (assume that the straight line is the only relation scheme applicable). Only upon the appearance of line e, and by establishing the relation, an event, between 0, 1 and e, are we able to see something: an instance of a straight line, 01e. By encountering point 2, and line f, a second event happens: 02f.

The dynamic nature of clusters is particularly important in the case of NL, which, due to topological and physiological restrictions of language processing, appears incrementally, always.

III FROM PEIRCE'S SEMEIOTIC TO SYNTACTIC SIGN

In this section an attempt is made to apply FCA to natural language. We have shown earlier that a prerequisite for such an application is a characterisation of the domain by a web of clusters.

In language modelling a common approach based on the division of input into clusters uses constituency relations (e.g phrase structure). Also, if clusters are regarded as relations, it can be argued that, from a mathematical point of view, dependency-based modelling takes precisely such a conception of language as its basis. But both models, at least to our knowledge, have serious difficulties with analysing actual language use, e.g. discontinuous modifications, coordination structures etc. [7].

We think that a reason for this lies in the nature of the relations assumed in these approaches. It will be argued that a more successful approach is possible on the basis of Peirce's semeiotic theory of signs.

A Peirce's semeiotic

Language is a system of symbols. According to Peirce, symbols share with icons and indices the structure of a sign. Every sign is constituted by the *triadic relation* between the sign itself, its object and its interpretant. Because signs are generated from signs, and in turn generate other signs, every sign must be an event. From this it follows that language consists of *sign-events* which, by virtue of their interpretants, are embedded within a web of sign-events.

The sign-events are generated according to rules which, in Peircean terms, are habits evolving from interaction with other signs. Language as a sign-event involves syntactic, semantic and pragmatic rules or habit.

Though Peirce distinguished different kinds of symbol-events according to their semantic value into quali-signs, sin-signs, and legi-signs, he did not, at least not to our knowledge, pursue the analysis of the function of symbol-events according to their syntactic value. But there is no reason why this cannot be done [1]. Indeed, from a syntactic point of view, symbol-events have a specific function, regardless of their semantic function. The syntactic value of the language symbols making up a unit of meaning (e.g. a syntactic cluster) may be seen in function of the value which they have in forming such a unit.

B Syntactic triad

By virtue of their secondness, events are marked by a binary relation which must be reflected in symbol-events. The syntactic value of symbol-events will therefore depend upon the sort of relation that obtains between two language symbols. If one of the symbols has by itself no information content and therefore is a mere quality, it will need another symbol to actualise its 'potential' content. Such nexus of two symbols, one of which has mere potential content, may be called a *proto-symbol* which corresponds to the category of firstness.

Similarly, when the nexus is constituted by an asymmetrical relation between one language symbol which itself has information content, but which derives its full content from its association with another language symbol which is in principle self-sufficient, it may be called a *deutero-symbol* which corresponds to the category of secondness.

Finally, when the nexus consists of language symbols which are self-sufficient but together generate the interpretant of the unit formed by the string it will be called a *trito-symbol* which, by its aspect of thirdness, mediates between the language symbols constituting a unit of meaning.

If in the above specification of syntactic signs, the internal structure of the subordinate signs is neglected, we get syntactic *sign schemes*.

C Clusters and syntactic signs

Syntactic signs, like any signs, are characterised by the triadic relation (by definition, lexical items are (degenerate) syntactic signs). What makes these signs syntactic is the very fact that they stand for specific *rules* or habits. Thus, the object of syntactic signs is the rule for which they stand. Their interpretant on the other hand is the generation of the selection of the next symbol-event. The interpretant of the entire string of language symbols is, from a syntactic point of view, the establishment of the correctness of the string, regardless of its semantic content.

D Economy of signs

Because syntactic symbol-events stand for rules governing the relationship between language symbols, the occurrence of one symbol will trigger the expectation of the occurrence of the other relevant symbol as required by the rule in question. Such expectation which may be regarded as the expression of a hypothesis (which according to Peirce is the conclusion of an abductive inference) regarding the nature of the next language symbol, must obey Peirce's principle of abduction. One of these may be called the principle of economy, which stipulates that all things being equal, a more plausible hypothesis must be tested before a less plausible one.

IV TOWARDS NLCA

NLCA is an abstract implementation model for Peirce's semeiotic applied to natural language, in particular, to syntactic clusters. A derivation of the principles of NLCA is the subject of this section.

A Symbols and topology

From a syntactic point of view, symbols have an expectation value, called their *combinatory need*. Syntactic signs in language are clusters which can be derived from the lexical items, based on their relations. Syntactic signs are recognised incrementally, from left to right. This, however, does not imply adjacency of the constituent lexical items. A cluster, in the above sense, denotes a group of *interdependent* items (in syntax, the term 'cluster' usually refers to a group of adjacent items).

From this it follows that, from an analysis' point of view, syntactic clusters are *tentative* signs, evoked on the basis of the combinatory need of the symbols recognised. A tentative sign can develop incrementally, by composition and re-composition. Re-composition is the process of first decomposing a sign to its (immediate) subordinate signs, and then composing a new sign from the constituent and other signs. In virtue of this capacity of self-organisation, language can be considered an *autonomous* system of signs.

B Semeiotic principles

Syntactic signs are composite signs which meet certain criteria. Thus, if a syntactic sign consists of a cluster of signs which are related to one another by a relation of subordination, the signs involved must in principle be *contiguous* to one another. This requirement is based upon the triadic structure of a syntactic sign the object of which is always a rule expressive of the expectation that a certain type of lexical items must be followed by another type of lexical item. We shall call a contiguous string a *range*. From this it follows that syntactic signs be necessarily *embedded* within ranges. Embedding allows a sign to have more than one role, in different signs.

By the principle of economy a more plausible hypothesis must be tested before a less plausible one. Clearly, in a left to right analysis, the shortest string representing a compatible candidate for a relation is the most plausible hypothesis. From this it follows that syntactic sign-events are established between nearest compatible candidates. Assume the contrary, then there is a sign which forms an event with not the nearest compatible sign. Then, by contiguity we have that the sign in question must have been involved already in some other sign-event.

Due to the tentativeness of clusters, the sign representing the 'nearest compatible candidate' may be subject to development. By forming a syntactic sign, that is, by establishing a syntactic relation, a new sign is generated. The properties of the new sign can be different from those of the signs involved in the relation. As a consequence, new sign-events may become possible. This interdependency of ranges and clusters implies that ranges are of a *dynamic* nature.

The combined effect of contiguity, embedding and the dynamic nature of ranges is that syntactic relations are established between nearest compatible candidates, always. This variant of the economy principle is called the principle of *greedy binding*. The above derivation from the semeiotic of syntactic signs to a range structure demonstrates that in Peirce's semeiotic theory of sign, meaning and form are related.

V RELATION SCHEMES

In NLCA we introduce three relation types (schemes) *qualification*, *minor predication* and *major predication* for the implementation of proto-, deutero- and trito-sign schemes, as follows. The mapping of sign schemes to relation schemes is based on their abstract properties and our linguistic knowledge of syntactic clusters. In so far the mapping is a faithful translation of those properties, we can postulate that the three relation schemes are necessary and sufficient for modelling syntactic clusters in language. The relation schemes are formalised in [2] and [6]. Below we briefly summarise the definitions; the underlying linguistic motivation is described in section V A.

A *qualification* (Q) consists of a qualifier and a core. The qualifier has no information content independent of the core; it makes the core more specific. The core fulfils the combinatorial need of the qualifier. A *minor predication* (mp) consists of a (minor) predicate and an argument. The predicate has information content independent of its argument, but its full content is derived from the relation with its argument. The relation between minor predicate and argument is asymmetric: the predicate needs its argument, but not the other way round. A *major predication* (MP) consists of a predicate and its argument(s). Both have information content, and the relation between predicate and argument(s) is symmetric (each requires the presence of the other). The predicate introduces an argument structure, and incorporates its arguments into a single relation.

The three aspects of a syntactic sign are present in the instances of each relation scheme, as follows. In each case, the thirdness aspect is represented by the relation (Q-, mp- or MP-relation). The secondness aspect is represented by the interaction between qualifier and core, or predicate and argument(s). The firstness aspect is represented by the constituent syntactic signs.

The constituents of Q, mp, and MP are realised by relation schemes, recursively. From this it follows that their instances, which are relations, define an embedded structure of ranges. In the construction of a relation functors (constructors) are involved. This implies that a relation can be uniquely deconstructed into its constituents (in the case of ambiguity we have a different relation for each analysis).

A Clusters linguistically

The properties of the relation schemes follow bottom up, based on our knowledge of syntactic clusters in language. Currently we have a model for English and Hungarian. In this paper, the language of illustration is English.

The different relation schemes reflect certain conceptual distinctions that may be expressed by means of language. The relation schemes *MP* and *mp* are referred to uniformly as *predication*, syntactically. The symmetrical instantiation of predication involves the distinction between an action/state and its participants: each requires the presence of the other. The asymmetrical variant distinguishes between an action/state or participant on the one hand, and its properties on the other (modification): the predicate requires the argument of which it is predicated, but the reverse does not hold. We call predicates of the first type *major predicates*, and predicates of the second type *minor predicates*.

There are various distinguishing factors between major and minor predicates. In English, major predicates (usually) relate to the noun-verb division; minor predicates do not. Major predicates are typically realized by verbs; minor predicates by adjectives, adverbs and prepositions. There is never more than one major predicate associated with an argument; there may be several minor predicates related to the same argument. (This reflects the possibility of having zero or more modifiers of an action or participant.)

The third type of relation, qualification, distinguishes between the core content of a linguistic expression and some qualification of it. At the level of an action and its participants, for example, this qualification may relate to referential status of NPs (e.g. definite vs. indefinite), or to tense and aspect information expressed by the verb.

The difference between a minor predicate and a qualifier is that the latter does not introduce a meaning that is independent of the element it qualifies. By contrast, a minor predicate has some aspect of meaning that is independent of the element it combines with. This is illustrated by the fact that minor predicates, not qualifiers, can be used in different contexts. For example, a prepositional phrase can modify an argument (e.g. noun) but also a predicate (e.g. verb). An adjective phrase can be used as a modifier of a noun, but also in the complementation of a verb.

The *Q*-, *mp*- and *MP*-relations can be realized in language on different levels, e.g. on the morphological level, on the level of syntax, or on the more abstract level of features.

Example 4 *Mary drank some wine.*
There are two *Q*-relations: drank, between *drink* and the feature PAST; and [some wine], between *some* and *wine*. There is an *MP*-relation between drank and its arguments, *Mary* and [some wine]. Each of these relations represent three aspects of a syntactic sign. E.g. in the case of [some wine], the *Q*-relation between *some* and *wine* represents the aspect of thirdness; the occurrence of *some* and *wine* together, the aspect of secondness; and the constituents, *some* and *wine*, the aspect of firstness.

Conform with our understanding of the term cluster, contiguity of signs does not imply that related items have to be contiguous. For example, in a discontinuous modification the intervening element contributes to the sign as a whole to which the remaining elements are related. This 'order creating' potential of syntactic signs is the key to NLCA's success in handling actual language use.

VI AN EXAMPLE

We develop an NLCA analysis of a sample sentence. Lexical items are elements of the set A. The set O (the set of clusters) is recognised incrementally, from left to right. Below we present those parts of the analysis that are essential for the derivation of the formal context. More details about algorithmic aspects of NLCA can be found in [4].

We assume that all information about lexical items is contained in the lexicon, such as type, number and location (pre- or post-) of argument(s), the relation scheme(s) they can evoke etc. In the example below, lexical information is specified on-the-fly. Items contributing to a relation (i.e. a cluster) are enclosed in brackets.

Example 5 *Mary drank some wine yesterday and a coke today.*
This example is problematic for the usual account of coordination, because the conjuncts *some wine yesterday* and *a coke today* do not form a single unit at any level of analysis; rather, they contain two independent constituents, one of which functions as an adverbial at clause level, and the other as direct object to the verb.

In the standard account, such coordination cannot be described, as there is no single unit of analysis that the rule describing the coordinate structure can refer to [5].

An NLCA analysis for the given sentence is as follows: *Mary* is a proper name which participates in an *MP*-relation. The item *drank*, which itself is a *Q*-relation between *drink* and PAST (drank), is a transitive verb. *Mary* is a compatible candidate (AGT) for drank (cf. greedy binding), but the combinatory need of the verb is not yet fulfilled, therefore no *MP*-relation is established between the two items. The item *some* is a compatible candidate for the second argument (PAT) of the verb. Therefore the *MP*-relation is formed: [Mary drank some]. The next item, *wine* can enter into a *Q*-relation with *some*. Therefore the *MP*-relation is re-composed yielding [Mary drank [some wine]].

The adverbial *yesterday* evokes an *mp*-relation with the verb, representing the adverbial's relation with the *MP*-relation as a whole. By re-composing the *MP*-relation we get: [Mary [drank yesterday] [some wine]].

The following item, *and*, is a coordinator. In general, coordination refers to the *relation* preceding the coordinator. The coordinator separates two ranges, one on the left and one on the right hand side, and coordinates compatible items. Briefly, two items are *compatible* when they may participate in the same relation scheme, in the same role (i.e. qualifier or core; predicate or argument), and relate to the same type of lexical unit. In this paper, the coordinator is considered a major predicate having two or more arguments.

The article *a* is a qualifier which must precede its core. There is no compatible item it can be coordinated with (we remark that non-syntactic factors are also involved in this matter). The noun *coke* participates in a *Q*-relation with *a*. The resulting cluster [a coke] is on the right hand side of the coordinator. There is an *MP*-relation preceding *and* which is decomposed. The item [a coke] can be coordinated with the compatible item [some wine], yielding the *MP*-relation: [[Mary [drank yesterday] [some wine]] and [Mary [drank yesterday] [a coke]]].

The adverbial *today* is on the right hand side of the coordinator. It can be coordinated with the compatible item *yesterday*. Again, by recomposition we get: [[Mary [drank yesterday] [some wine]] and [Mary [drank today] [a coke]]]. We emphasise that deconstruction and composition are operations at sign level, not on syntactic surface structure.

The embedded ranges are shown in figure 4. Items of outermost ranges are underlined; a marker c(d) indicates construction (deconstruction) of a relation. The resulting relation is shown in figure 5.

Input	**Step**	**Ranges and clusters**
Mary		Mary
drank	c	Mary [drink PAST]$_Q$
some	c	[Mary drank some]$_{MP}$
wine	d	Mary drank some wine
	c	Mary drank [some wine]$_Q$
	c	[Mary drank [some wine]]$_{MP}$
yesterday	d	Mary drank [some wine] yesterday
	c	Mary [drank yesterday]$_{mp}$ [some wine]
	c	[Mary [drank yesterday] [some wine]]$_{MP}$
and		[Mary [drank yesterday] [some wine]] and
a		[Mary [drank yesterday] [some wine]] and a
coke	c	[Mary [drank yesterday] [some wine]] and [a coke]$_Q$
	d	Mary drank [some wine] yesterday and [a coke]
	c	[[Mary [drank yesterday] [some wine]] and [Mary [drank yesterday] [a coke]]]$_{MP}$
today	d	Mary drank [some wine] yesterday and [a coke] today
	c	[[Mary [drank yesterday] [some wine]] and [Mary [drank today] [a coke]]]$_{MP}$

FIGURE 4. Ranges and clusters

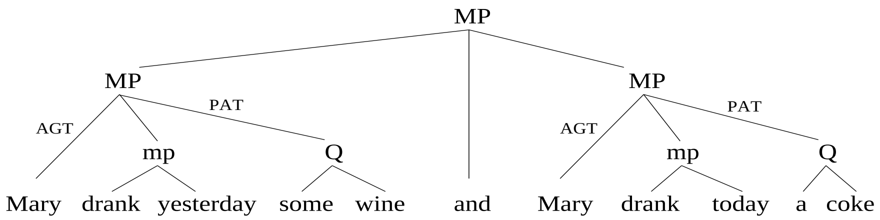
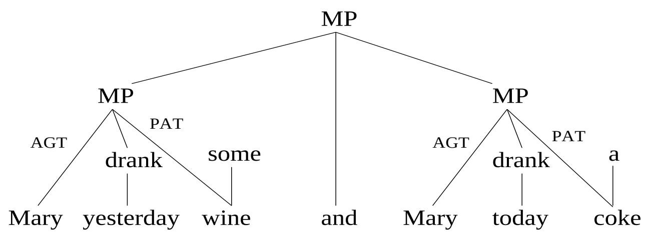


FIGURE 5. The derived relation

A A formal context

We develop a formal context from the relation of figure 5, stepwise. The resulting bipartite graph is shown in figure 6. Conform the definition of the three relation schemes, it can be assumed that 'major' information about the *Q*-, *mp*- and *MP*-relations is present in the core, the argument and the predicate, respectively.

In NLCA, the partitioning π is generated simultaneously with the analysis by a heuristic algorithm. In the 2-colouring of π we respect, when possible, the intuitive classification of items, according to which nouns are objects, verbs and adjectives are attributes. The resulting formal context is shown in figure 7. We conjecture that the complexity of finding a suitable π is NP-complete. NLCA, like FCA, applies heuristics, however, on a different level.



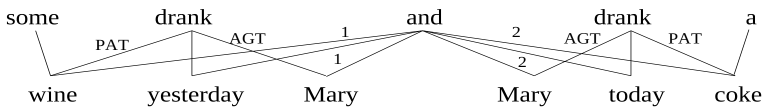


FIGURE 6. A partitioning of the derived relation

	drank(d)	AGT(a)	PAT(p)	some(s)	and(&)	1	2	a
Mary(M)	+	+			+	+	+	
wine(w)	+		+	+	+	+		
yesterday(y)	+				+	+		
coke(c)	+		+		+		+	+
today(t)	+				+		+	

FIGURE 7. The formal context

The CL of the formal context of figure 7 is depicted in figure 8. Interestingly, the formal concepts correspond to the focus of WH-questions: (C_1, C_3, C_5) What did somebody drink? (C_2) Who drank something? and (C_4, C_6, C_7) What happened? This suggests a potential correspondence between question formation in language, and the information reflected in formal concepts (cf. transparency). A sublattice also has information content, e.g. C_1-C_4-C_5-C_7 and C_3-C_5-C_6-C_7 represent the two sub-events, which are isomorphic, and the sublattice C_2-C_4-C_6-C_7 the coordination of the two events by their common agent.

C_0 =({},{d,a,p,s,&,1,2})
C_1 =({w},{d,p,s,&,1})
C_2 =({M},{d,a,&,1,2})
C_3 =({c},{d,p,&,2,a})
C_4 =({M,w,y},{d,&,1})
C_5 =({c,w},{d,p,&})
C_6 =({M,c,t},{d,&,2})
C_7 =({M,w,y,c,t},{d,&})

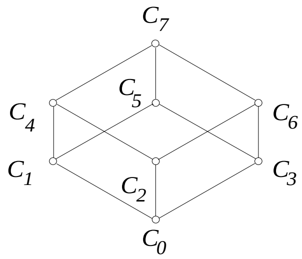


FIGURE 8. The concept lattice

VII A DIFFERENT ANGLE

In previous sections we have shown that, in the case of NL, a formal context can be derived automatically. We can try and apply this approach to our example of the angle. Indeed, the following sample sentence would do: *The straight line on* 0 *and* 1*, and the straight line through* 0 *and* 2 *is an angle.*

In order to get practical formal contexts, for example, in the case of information retrieval, the relation generated by the NLCA analysis may have to be simplified. Such a process, called *summarisation*, involves the systematic removal of *Q*-, *mp*- and *MP*-relations according to the order relation induced by the relation of subservience (recall that the relations are implementations of the syntactic signs, and syntactic signs correspond to the categories), and possibly, to domain specific information about lexical items and clusters. For example, we may assume that, in the domain of geometry, points and lines are important, unique entities. Therefore the treatment of relations involving such items must be subject to different rules.

Summarisation applied to the above sentence removes the *Q*-relations ([the line] and [an angle]) and two *mp*-relations (the two occurrences of [straight line]). In accordance with the domain specific information, the *mp*-relations evoked by the prepositions are not removed. Finally we get the 'telegram style' text: *line on* 0 *and* 1*, and line through* 0 *and* 2 *is angle.* The syntactic signs representing the subject (SU) and the subject complement (CS) are embedded in the *MP*-relation involving the verb *is*. Because *is* only mediates between SU and CS, we can say that SU and CS define each other. If we let, for example, SU define CS, then by taking the relation of SU, eventually, we get a formal context which is isomorphic to the one of the angle (cf. example 1).

VIII FURTHER RESEARCH

The summarisation potential of NLCA, and the capacity of CLs in reasoning can make NLCA attractive for information retrieval. A practical validation of this statement is part of an ongoing project.

IX CONCLUSION

Conceptual structure in natural language (NL) is found as a result of the application of formal concept analysis (FCA) to NL. This requires a re-definition of the basic notion of FCA, the formal context. Starting from Peirce's semeiotic theory of signs, three types of syntactic signs and corresponding relation schemes are developed. The relation schemes, which reflect conceptual distinctions that can be made in language, form the basis of a novel model of language, NLCA. A preliminary version of the language model is described in [7]. A proto-type implementation of NLCA is currently available.

A Acknowledgements

I am grateful to Vera Kamphuis for her indispensable support in this work, and for proof reading earlier versions of this paper. Valuable correspondence with József Farkas is acknowledged.

REFERENCES

1. Debrock G., and Sarbo J., "Towards a Peircean Model of Language", Technical Report CSI-R9802: University of Nijmegen (1998).
2. Farkas J., Kamphuis V., and Sarbo J., "Natural Language Concept Analysis: Towards a theoretical foundation", Technical Report CSI-R9717: University of Nijmegen (1997).
3. Ganter B., and Wille R., *Formale Begriffsanalyse – Mathematische Grundlagen*: Springer-Verlag (1996).
4. Grootjen F., "Natural Language Concept Analysis: Towards an algorithmic implementation", Technical Report CSI-R9811: University of Nijmegen (1998).
5. Kamphuis V., "Coordination and Surface Structure", *Actes du Congrès International des Linguistes*, Paris: Elsevier Sciences (to appear).
6. Kamphuis V., and Sarbo J., *Int. Conf. on New Methods in Language Processing and Computational Natural Language Learning*: ACL, 1998, pp. 205-214.
7. Kamphuis V., and Sarbo J., "NLCA – A truly relational model of language", Technical Report CSI-R9815: University of Nijmegen, (1998).
8. Lehmann F., and Wille R., "A triadic approach to formal concept analysis", *Third Int. Conf. on Conceptual Structures*: Springer-Verlag (1995).
9. Peirce C., *Collected Papers of Charles Sanders Peirce*: Harvard University Press (1931).
10. Sarbo J., "Lattice embedding", *Conceptual Structures: Knowledge Representation as Interlingua*, **1115**: Springer-Verlag, 1996, pp. 293-307.
11. Sarbo J., *The Computer Journal*, **39(10)**, pp. 868-875 (1996).
12. Sarbo J., and Farkas J., "A data representation for abstract reasoning", *Proc. of the Seventh BENELEARN Conference*, Tilburg, The Netherlands, 1997, pp. 99-108.
13. Sowa J., *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Boston, PWS Publishing Company (1998).
14. Wille R., *Logic and Algebra*, New York: Marcel Dekker, 1996, pp. 267-281.