Text analysis for knowledge graphs

Popping, Roel

Published in:
Quality & Quantity

DOI:
10.1007/s11135-006-9020-z

Publication date:
2007

Quality & Quantity (2007) 41:691–709
DOI 10.1007/s11135-006-9020-z

Text Analysis for Knowledge Graphs

ROEL POPPING
Department of Sociology, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands. E-mail: r.popping@rug.nl

Abstract. The concept of knowledge graphs is introduced as a method to represent the state of the art in a specific scientific discipline. Next the text analysis part in the construction of such graphs is considered. Here the 'translation' from text to graph takes place. The method that is used here is compared to methods used in other approaches in which texts are analysed.

Key words: text analysis, knowledge graphs

1. Introduction

The growth of scientific knowledge has drawn for a long time. It is usually considered at a high level of abstraction, where the discussion is about patterns of growth. Some considers these patterns as unstable; there are 'sudden' developments that change all that is known. This is even named a scientific revolution (Kuhn, 1962). Others see the growth patterns as a rather continuous development.

On a more concrete level one usually looks at persons or groups who are responsible for this growth. Here the focus is on certain people who have a group of scientists around them, who together are able to produce a lot of new knowledge. There has also been a lot of attention for communication among scientists. Here investigators discuss each other's ideas, comments on texts, and so on. This might happen at conferences, but it takes also place in another way. Investigators mail each other text (electronically or via the postman), or phone each other. These are the so-called invisible colleges, which play an important role in the development of science (Crane, 1972). The activities in these colleges may result in more official co-operation between investigators. Here the process of growth is relevant.

In the context of growth of knowledge networks have been used to represent citation networks (Hummon and Doreian, 1989) in order to see who is influenced by whom, or by a specific text of a person. The networks are also used to indicate persons co-operating together, often within specific research groups (De Haan, 1994).

One might also look at the knowledge that is already available in some field. It is available, but in a distributed way. State-of-the-art articles usually try to overcome this problem. Representing this knowledge in a simple and clear way might also contribute to the growth of knowledge. At this point, however, a distinction must be made between theories or, more concrete, hypotheses that still have to be tested, and theories or hypotheses that have already passed a test. These latter ones are the interesting ones, as they denote the real empirically proven knowledge. Representation of the state of the art of this knowledge in a certain field or discipline of science allows getting a picture of gaps in the knowledge or areas where conflicting results have been found. This all makes clear where further research is needed, and what knowledge can safely be used in policy-making or decision-making processes. For the representation, it is necessary to combine the results of several investigations. This is accomplished in so-called knowledge graphs. A problem, however, is which knowledge to use and how to represent this knowledge. This text deals with these problems.

The organisation of the text is as follows. First a short description of knowledge representation is given; necessary to place the subject of knowledge graphs. Next these graphs are introduced. The remaining text deals with text analysis. The main questions concern which texts should be used, how they should be coded and which problems have to be overcome.

2. Knowledge Representation

In knowledge representation a distinction is made between procedural and declarative knowledge (Ryle, 1949). Procedural knowledge ('knowing how') is described as a set of prescriptions for actions. Usually these prescriptions are referred to as situation-action or if-then rules. Declarative knowledge ('knowing that') is given as a set of assertions about a certain subject. Conclusions can be drawn from these assertions by inference methods. Such inference methods can be based both on formal logic and on graph theory. Conceptual knowledge is declarative knowledge in the form of law-like assertions. Its core consists of explicitly defined types of relations between concepts.

The application of a knowledge representation for a specific task is referred to as exploitation. An example is decision support.

The process of obtaining a representation is called the structuring of knowledge. Here the acquisition of knowledge is necessarily combined with the integration of this knowledge.

Based on use and form Stokman and De Vries (1988) present a systematic evaluation of systems in the field, see Table I.

The exploitation of procedural knowledge is accomplished by chaining of rules; usually these rules have an if-then structure. Forward chaining is

Table I. Classification of knowledge-based activities

Use of knowledge	Form of knowledge Procedural knowledge	Conceptual knowledge
Exploitation	Decision-support and information retrieval on the basis of chaining rules	Decision-support on the basis of detection causality, information retrieval on the basis of definition relations
Structuring	Verification of rules Induction of rules	Integration of definitions and causal models

Source: Stokman and De Vries (1988: 189).

used in case an action is prescribed. In case one tries to find an explanation for what has happened backward chaining is used. The structuring of procedural knowledge takes place when a chain of rules is inspected and updated that is triggered by a particular problem presented to a rule based system. The exploitation of conceptual knowledge is met in the application of inference procedures to semantic networks for the retrieval of information. The knowledge graphs, to be introduced hereafter, are part of the structuring of conceptual knowledge which leads the integration of definitions and causal models.

3. Knowledge Graphs

A knowledge graph is a network in which knowledge in some field is represented by labelled nodes and labelled links between these nodes. For the links only a few types of relations are used (James, 1992: 98).

The construction of knowledge graphs starts with the extraction of information from texts. This is called *text analysis*. Here subject-verb-object (SVO) syntactic relations are encoded as $concept_1$-link-$concept_2$ relations.[1] The result is a list of concepts, represented as labelled points, and a list of typed links between the points. These form the so-called author graph.[2] A concept is a unit of meaning. It is used as the basic unit for the meaning content of what it refers to (Popping, 2000: 17). The most important type of link between points is the causal relation. This relation refers to a structural relation between concepts.

The next step is called *concept identification*. Here the various author graphs are combined into one graph by identifying points with each other. When the texts that were the basis of the graphs deal with the same concept, points with the same label are identified. An author may use synonyms for a concept; therefore, points with different labels should be identified too. This is done by comparing the neighbourhoods of points

to identify the potentially identical pairs. An index has been developed for measuring the similarity between two points. The value this index takes, in combination with a threshold value, can be used to decide upon identification of two concepts. In the same way it is possible to detect points with the same label, but referring to a different content, the so-called homonyms. For example, a chair is something one sits on, but it can also refer to one's position (e.g., a committee leader or a professor). One of these points should receive another label. A compiled graph results which is free of ambiguity of language. This graph is further investigated in procedures called *concept integration* and *link integration*. The first procedure tries to find interesting substructures; the second procedure infers new links from the given ones. The result is called the integrated graph.

In order to represent the structure of knowledge sometimes a complex relation is necessary: the frame relation. This relation combines a number of concepts and relations that are inseparably connected into a single concept. These concepts and relations together ensure that the frame functions as it is supposed to. Often it holds for specific types of variables like constructs and theoretical terms (Kaplan, 1964: 54–57). An example of a frame is the measurableness of quality of work. A bicycle might also be regarded as a frame; it consists of the frame, the wheels, the handlebar, and so on. Together all these parts enable the bike to work. Concept integration aims at determining those subgraphs that are candidates for contraction into a frame. (Note, the term is used in another sense from that generally used in artificial intelligence.) In link integration relations are combined to deduce new relations. If there exist relations between points A and B as well as between B and C, there may be reasons to infer a relation between A and C. To find these new relations, a path-algebra (Carré, 1979: 84–85) is used. Relations can be based on multiplication, for the serial combination, and on addition, for the parallel combination. The whole process of knowledge integration and graph construction is summarised in Figure 1.

Four characteristics are distinguished with respect to relations between concepts: directionality, meaning, sign, and strength (Carley, 1993; Popping, 2000: 99). First only directionality and meaning have been used in the theory on knowledge graphs. All relations are unidirectional, and the meaning, or the semantic relation, is denoted by using types (see below). A meaning like "is friends with" is not used, but "is a kind of" is used. The characteristics sign (positive or negative) and strength (usually a value on a 0–1 scale) are added in a later stage (Popping, 2003).

Originally the idea was to represent knowledge by using as few meaning types as possible. First only the types CAU, PAR, and AKO were used. The CAU relation denotes a cause–effect relation. This can refer to a structural relation between concepts ("The occurrence of unstable market

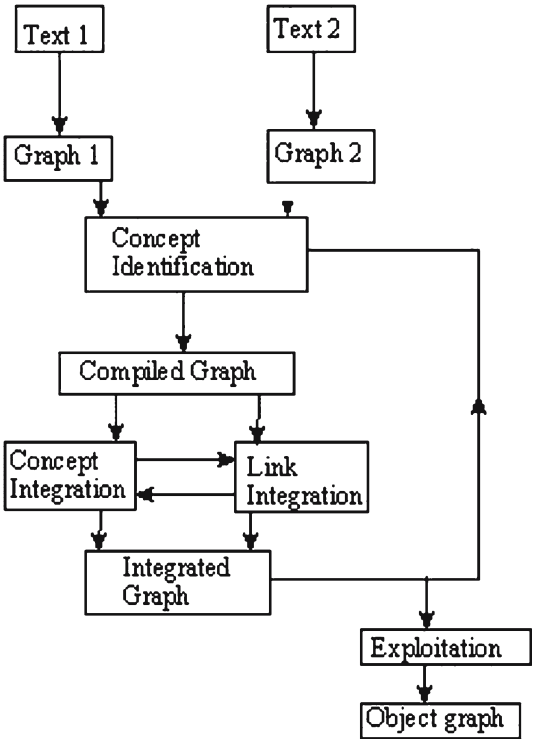

Figure 1. Process of knowledge graph construction.

positions causes polarisation"), but also to a process ("An increase in X causes an increase in Y"). The relation is asymmetric and transitive. In all methods using networks based on text the causal relation is read as *might* cause.[3] PAR stands for the *is part of* relation ("Having relations with high status is a part of social capital"), it characterises a property of an attribute. AKO refers to *is a kind of* ("A married man is a kind of man; a low educated woman is a kind of woman"). Here something is exemplified, a concept or class of concepts is considered as a special case of another concept or class of concepts. The latter two relations are transitive and asymmetric.

Inverse relations are also distinguished: CBY (is caused by), HAK (has as kind), and HAP (has as part) (Stokman and De Vries, 1988). Today different types of concepts are distinguished: types and tokens. Tokens play a role similar to that of a variable in logic. Types are labelled points representing generic concepts that are determined by their attribute sets. Types can be seen as giving schema information, whereas tokens represent arbitrary instantiations of types. A token denotes an individual that can be chosen from a universe given by the discourse. As an example,

"Pluto" is a token, and "dog" is a type. The choice of the individual might be restricted, and the restriction follows from the relations attached to the token. The relation between token and type is denoted by ALI (alike). There are seven relations between types: PAR, CAU, AKO, ORD (ordering), ASS (association; symmetric), EQU (equal; symmetric), DIS (distinct).

4. Mathematical Elaboration

From a mathematical point of view the knowledge graphs are thoroughly elaborated (Bakker, 1987). The process of concept identification is well defined. The outcomes found by using similarity indices suggest that different concepts might be identical, or in reverse that identical looking concepts are different. In the first situation, the investigator finds concepts that are synonyms. In the second situations, there are homonyms. It is up to the investigator to decide which concepts have to be taken together, and which should be split. The link integration process is also clear. A limited number of relation types is used now, as was indicated above. By constructing knowledge graphs in different fields of application we will learn whether these are the necessary types or whether some types have to be added or can be dropped.

Concerning concept integration some discussion is needed. Measures are available to find concepts that might be integrated, but should one do this? Usually concepts are linked to one or more specific concepts within a frame (Popping, 2003). Figure 2 contains the concept 'career perspective', which is linked by a PAR relation to the concepts 'employee' and 'job'. The concept might be considered as a frame, as it consists of several concepts itself and the relations between these concepts. In Figure 3 the concept is unravelled into the concepts 'career' and 'perspective' that are linked by means of a PAR relation. One can say this is the content of the frame. The concept 'employee' is no longer linked to the complete concept, but just to the 'career' part in it. The same holds for 'job', which is now only linked to 'perspective'. This new representation gives another view on the concepts, and probably allows more links between concepts.

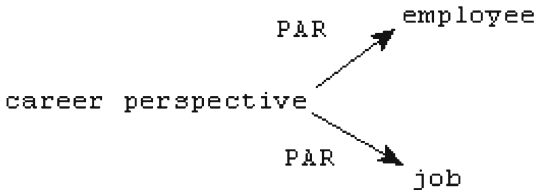

Figure 2. The concept "career perspective".

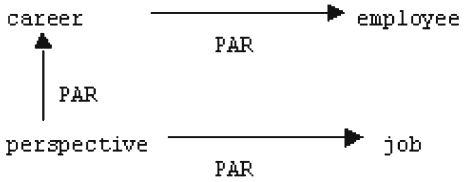

Figure 3. Unravelled concept "career perspective".

A balance is to be found between what is included in a frame and which concepts must be maintained in order to allow linking to other concepts. The unravelling looks very useful in case the frame represents mainly a state of the art, and not when it represents a process. This can be verified after several knowledge graphs have been constructed.

The problems just mentioned are already related to the content, to that what is represented in a knowledge graph. Here more discussions are needed. These discussions are centred on two main questions: (1) Which knowledge is used, where does it come from?; and (2) how should this knowledge be represented in graphs?

5. The Construction of a Knowledge Graph

Knowledge graphs have been developed with the intention to represent scientific texts in some field of application (James, 1992). Especially sciences using pre-paradigmatic knowledge (Kuhn, 1962), like social and medical sciences,[4] seem relevant fields. Theories in these sciences are empirically oriented, rather than deductive systems built upon a small number of premises as in, for example, physics. Scientific knowledge in the sciences mentioned first is oriented towards explanation and prediction of empirical phenomena by means of theories, in which covariances between classes of phenomena are ordered in a logically consistent and coherent system. The building stones of these theories are concepts of which at least some are related to empirical phenomena. For testing a scientific theory two submodels are distinguished: a measurement model that specifies the relation between manifest behaviour and latent concepts; and a structural model that specifies the direction and type of association between the different concepts and that, as a consequence, specifies the structure of the phenomena to which these concepts refer. This implies that concepts can be distinguished at different theoretical levels.

In the past social scientists performed content analysis. Today the term thematic text analysis is used. Here theme occurrences (or counts) are examined. Furthermore some other views on text analysis have been developed. The first one is the semantic text analysis, in which theme usage is examined (cf. Kelly and Stone 1975; Roberts 1989; Franzosi 1990). Here

SVO-statements are examined. From here it is a small step to the next view, the network text analysis where the SVO-statements are taken together in networks. In a way these three ways of text analysis are cumulative. The construction of a knowledge graph resembles the network text analysis. If necessary the text analysis for knowledge graphs can be compared to these more general methods of text analysis.

Knowledge graphs should be based on knowledge that is empirically tested. This is true factual knowledge, knowledge valid at a certain time and place. Time and place are the only restrictions posed to the knowledge available in an author graph. In case such a graph is integrated with another author graph the resulting graph applies to the population that is the intersection of the populations the original author graphs are based on (Popping, 2003).

The benefit of only using empirically tested knowledge is that there will be no interpretation and no evaluation by the original investigator or the investigator or coder responsible for constructing the knowledge graph. When one starts from theories or hypotheses that are not empirically tested, it is possible that what is stated in a clause is not correct. After (link) integration this might result in a graph, which contains information at several places that is not correct and that is not or hardly traceable to the source from where it came. An example of a knowledge graph based on not yet proved theories is found in Popping and Strijker (1997). They present two theories in the field of labour market research. Popping (2005) discusses a graph based on some empirical studies in this same field. Hoede and Weening (1999) present knowledge graphs based on definitions of the concept 'imperialism', as given by Marx and Schumpeter. The only interesting thing is that later such 'theoretical' graphs can be compared to the knowledge graphs based on empirical data. This tells how good the theory did predict reality. The graphs do not show the actual state of the art, however.

An enormous difference with the common qualitative research and quantitative text analysis research is that the investigator does not have to try and find relevant concepts and processes. These are mentioned by the author of the text that is at the basis of a knowledge graph. Another difference is that no discourse or natural language is investigated, although this has been related to knowledge graphs (Van den Berg, 1993); only empirical relations are investigated.

The choice for the field to be covered in a knowledge graph is up to the investigator who will construct this graph. This investigator has to inform on which texts the graph is based, and why on these texts. Here the demarcation of the field becomes visible, and also the criteria the investigator used for selecting a text or not.

The goal for which the graph should be used has its effects on the choices that are to be made now. The graph may be used to have a description of the field; here CAU, ASS, AKO and PAR relations are very important. In case explanation is a main function, CAU relations will get a lot of attention. Here it is also relevant to distinguish between independent and dependent concepts. The concepts and relations are mainly part of the structural model. But, one has to look at the measurement model in the sense that the strength of the CAU (and ASS) relations in the graph are based on statistical parameters. These relations are no deterministic relations. The parameter is the one as used by the original investigator. However, in one study this is a product moment correlation coefficient, in an other study it is a B or beta coefficient found in some type of regression analysis. Here some problem may rise. PAR and AKO relations usually are not based on empirical results, but are based on definitions. These will be part of the structural model.

It was mentioned before that the graph holds for a certain population. It is possible however to consider a graph from a certain perspective. A graph about labour markets will be different for employees compared to employers. This will give rise to conditional graphs (Popping, 2003). Now one has to ask for the conditions under which different representations of empirical scientific knowledge might be integrated.

The population is restricted by time and place. Time might also be considered in another sense. Comparing knowledge graphs on two moments in time will allow investigating scientific growth. The part that is added (a possibility that is removed) shows the scientific growth in the time between the two moments. Knowledge graphs do not tell why it was decided to do investigations with respect to specific new issues.

5.1. SELECTION OF RELEVANT TEXTS

Texts in which reports of empirical studies in the field selected are presented are at the basis of knowledge graphs. Even now there might be a lot of candidate texts. It is not possible to say in advance which ones are the relevant texts. One criterion for deciding about the relevance of a text is the fact that the text is cited very often. One can even use that it has been cited that many times in acknowledged systems like the Social Science Citation Index within a certain number of years. Another criterion is that experts in the field mention a text. Finally a criterion based on a completely different principle. The investigation that is reported in the text covers a very broad population. Therefore, the restriction of time and place as mentioned before is limited as much as possible. The two criteria mentioned first contain some quality element; the last mentioned criterion is lacking this argument.

Instead of analysing texts one might also interview the authors of these texts. I would only do this in addition to the text analysis. The author can assist in interpreting the text and in defining the concepts and their relations. This task might be compared to protocol analysis (Ericsson and Simon, 1994). A backdraw might be that the author only mentions a number of relations or gives an interpretation that is different from the one in the text. This is relevant in case one wants to control the source of a relation.

Texts may show significant differences in (1) the terminologies, (2) the level of detail of the descriptions, and (3) the relations distinguished (Bakker, 1987: 29). The first difference refers to the problem of synonyms and the problem of different definitions by authors. This is solved in the process of concept identification. The second and third difference go back to the investigation that is at the basis of the text. For the one constructing the knowledge graph these are facts. Especially the second difference might cause an imbalance in the details in the knowledge graph.

5.2. SELECTION OF RELEVANT SENTENCES

In most thematic text analysis studies, all sentences from a text are coded. Here in fact a computer does the job. In case manual coding is necessary, the sentences to be coded are usually selected in a random procedure or they have to meet some specified criterion (like, the first sentence is a paragraph containing...). With respect to a knowledge graph all sentences are needed that contain crucial information. This is information containing results of the empirical research (by preference presented as CAU relations), or containing context information necessary to understand the results (usually PAR or AKO relations). Bakker (1987: 37) writes about law-like sentences. The sentences consist of one or more clauses. Each clause contains a positive assertion. Empirically tested knowledge does not contain negative assertions (like: "a house is not a tree").

In practice the main conclusions as drawn by the author are incorporated in the knowledge graphs. Problematic decisions concern the marginal looking conclusions, remember the person constructing the knowledge graph is not the original investigator. Should these marginal looking conclusions be included or not? Also the question on the strength of the relation between two concepts in order to consider this relation as relevant might give rise to discussions.

Roberts (1989), doing semantic text analysis, distinguishes four kinds of sentences, based on the meaning that they intend to convey (namely, as a description or as a judgement of a process or of a state-of-affairs). As knowledge graphs are based on the results of empirical research, the judgement of a process and the judgement of a state-of-affairs will not occur in

the texts that are at the basis of such a graph. This implies a simplification compared to the text analysis as is performed in more general methods.

5.3. CODING A CLAUSE

This section deals with the question: How can empirical scientific knowledge be extracted from its carrier and represented in a computer system?

In several handbooks on text analysis, a distinction is made between the instrumental and representational approach to coding. At issue is 'whether' one chooses to apply one's own theory or one's sources' theories to the texts under analysis. This distinction was originally made by Osgood and has recently been refined by Shapiro (1997). In the instrumental view texts are interpreted according to the researcher's theory. The approach ignores the meanings that the texts' authors may have intended. When the representational perspective is applied, texts are used as a means to understand the author's meaning. For more details the reader is referred to Popping (2000: 59).

The coding for knowledge graphs always starts from the representational view. This means that coding is in line with what the original investigator had to say. As results of empirical research are presented one refers to facts and not to opinions or evaluations.

For the coding it is relevant that the investigator concentrates on clauses, those parts of a sentence with their own inflected verb and associated subject and object. A sentence might consist of several clauses. Inflected verbs can be recognised as any words that change form when the person and/or tense of the clause are changed. An example: The word, 'go', in 'I go', and the word, 'goes', in 'He goes', are inflected verbs because they change form when the subject changes from first to third person.

The coding of simple clauses will, technically speaking, hardly give problems. The concepts representing subject and object are given in the text. These are manifest concepts that were already defined by the original investigator, on the place of the verb one of the relation types should come that has been accepted for entrance in the knowledge graph. An example of a clause is: "The employee has essential skills." Here we recognise subject, verb and object. For the coding the relation type is most important. 'Has' denotes a quality. This clause is best coded in reversed order as the concept 'essential skills' is part of the concept 'employee'. It is up to the coder to decide whether the genitive that belongs to the subject or object part of the clause ("essential skills") must be taken into account. If so, one might have to distinguish two concepts: "skills" and "essential skills". This way of coding resembles very much the ways followed by other investigators in the field of network text analysis (Carley, 1993; Van Cuilenburg et al., 1988).

A big difference with coding in qualitative research is that in that type of research the investigator has to find and define the concepts (see e.g. Contas, 1992), and that these even might be latent concepts. In knowledge graphs the concepts are determined by the original investigators on whose texts the graphs are based.

Real problems may rise when the concept is a complex concept. I will come to this later. Problems might also rise when the clause is considered. This works in two ways: (1) when does one have a relation that is relevant for the knowledge graph, and (2) the relation might also be complicated.

The question on the relevance of a relation holds especially in the situation of a causal relation. In the empirical research the strength of the relation is indicated as mentioned before by some statistical index. The investigator has to decide which value such a coefficient must have at least in order to include the relation in the knowledge graph. The criterion depends on the actual value, but also on characteristics of the sample on which the investigation is based.

The complexity of clauses is met when relative, proxy or conjunctive clauses are found. A relative clause is a clause that modifies a noun or noun phrase (e.g., "the employee, who has a high level of training"). A proxy clause stands proxy for the subject ("that he had a high level of education pleased them") or object ("they knew that a high level of education is necessary"). Relative and proxy clauses will not be found too often when one concentrates on empirical results. Conjunctive clauses on the other hand, will be found rather often. These clauses have a relation to another clause that is defined in terms of a conjunction ("the employee was happy, because the candidate had a high level of education").

The reliability of the coding task can be controlled for in case the coding is at least performed twice. Now intercoder reliability indices can be computed. These indices are available for the situation where concepts are named in different ways (Popping, 1992). Adaptations must be looked for in case one coder selects other sentences from a text than another coder does.

6. Theoretical Levels of Abstraction of Concepts

From a theoretical point of view concepts can be distinguished at three different levels of abstraction. The highest level is constituted by the *theoretical construct*. Constructs can not completely be reduced to variables that can be observed, as they contain some additional systemic meaning (Kaplan, 1964: 58). An example of a construct is social economic status. Constructs are measured via *dimensions*. These dimensions constitute the second level. Dimensions do not fully cover the complete meaning of the construct. The construct social economic status is usually measured by

using dimensions referring to income, education, and occupational status. Dimensions are not directly recognised in the empirical world, but only by *operational definitions*. An operational definition of a specific issue is a description of the operations, which allow determining whether and eventually how this issue occurs in the empirical world; it is a perceptible indicator of that issue. Gross and net income are operational definitions, they refer to the dimension income.

Knowledge graphs on the level of operational definitions are very detailed. As indicators might be changed (e.g., not gross but net income is measured), it is possible that several indicators can be used to represent one dimension.

Problems may rise when one switches to another level. Entering a relation 'income PAR social economic status' is correct, the investigator using this statement will loose some detail information however in replacing 'x relation income' by 'x relation social economic status'. Replacing one level by another however really gives problems. Assume someone found: '(low) income CAU poverty', but wants to have this on the level of the construct. Now 'low social economic status CAU poverty' would be found, which is not correct.

The issue becomes urgent in case in one study income is used as a variable and in another study the social economic status. The problem is to be solved by the investigator who constructs the knowledge graph. The issue might be considered as part of the problem of implicit knowledge (Popping, 2003), knowledge that in fact everybody possesses and that therefore never is communicated. In that case one will get income PAR social economic status. Information about the context in which the concepts are used is relevant here. I do not see other general solutions, so the only remark left is that the investigator should be careful.

7. Ambiguity

Ambiguity refers to the situation in which an expression can have more than one meaning. 'Spring' denotes a place where water flows, but also a season of the year. From Schrodt et al. (1994) we learn that the words 'accuse' and 'deny' lack ambiguity, but that 'force' and 'attack' are very problematic. This is because the latter terms can be used both as nouns and as verbs. Ambiguity complicates the encoding process in any analysis of texts. Roberts and Popping (1996) distinguish three types of ambiguity (and related methodological problems) that arise in text analysis during three necessary steps in constructing networks from texts: idiomatic ambiguity (in the identification of concepts [or points]), illocutionary ambiguity (in the identification of syntactic links [or relations]), and relevance ambiguity (in the identification of network characteristics). As text analysis for

knowledge graphs includes only a limited part of the general text analysis, several problems do not occur. Next follows a treatment that is specialised on text analysis for knowledge graphs.

As the knowledge graphs to be built are based in results of empirical research a lot of ambiguity will already be overcome in reporting these results. This however, does not have to take away all ambiguity.

7.1. IDIOMATIC AMBIGUITY IN IDENTIFYING CONCEPTS

In the classical thematic text analysis, a computer program links words to concepts. Here an instrumental approach to understanding text is taken; the analyst's theoretical perspective is used to 'decode' meanings of which the text's manifest form is symptomatic. A human coder performs the kind of text analysis that is discussed here. Now a representational view to understanding text is used, the analyst uses *Verstehen* to encode the texts according to the meanings their sources intended. The issue is no longer 'how' to encode text, but 'whether' one chooses to apply one's own theory or one's sources' theories to the texts under analysis.

The one who performed the empirical investigation identified the concepts that are used. In the process of constructing a knowledge graph the coder might run into problems when confronted with concepts referring to the same characteristic but still defined in another way. An example is found in the different definitions of 'imperialism' as investigated by Hoede and Welling (1999). But concepts as used in empirical research might also have a different content. Many latent concepts are measured by applying some scaling procedure. These might be different, based on variables included and on scaling method used.

This problem of a different content becomes especially relevant when graphs are integrated. Sometimes the coder is aware of these differences; otherwise they should become visible when procedures for testing on synonyms and homonyms are applied. Here the correct theoretical level of the concept is also relevant. Besides, does the concept contain all relevant information? Or, is 'essential skills' identical to 'skills', so what is the role of the genitive? These are questions that must be answered by the coders (and investigator) in a clear way.

7.2. ILLOCUTIONARY AMBIGUITY IN IDENTIFYING CONCEPT RELATIONS

After removing idiomatic ambiguities a coder in a traditional text analysis study may become confronted with illocutionary ambiguity. This is because such coders must accommodate linguistic ambiguities that arise between the concepts under analysis and the intentions of authors who uttered them – ambiguities that are grounded the very structure of language itself

and that can only be clarified in the light of the context in which they are expressed. As indicated before such ambiguities are not expected with regard to knowledge graphs based on empirical texts.

Most relevant with respect to the knowledge graphs is whether the coder is able to choose the correct relation type. This might become especially relevant when the concepts are of a different theoretical level. Does the coded SVO-statement still have the same meaning as the original clause, especially after inversion? Also a conjunctive clause may take care of ambiguity, it should be clear whether it refers to a concept or to a whole clause.

A distinction can be made between grammatical and semantic complexity. Grammatical complexity refers to the situation in which it is hard to find out the basic structure of a sentence. Semantic complexity is met when problems rise in interpreting sentences that are connected. This is for example when in one sentence a reference is made to another sentence by using a word like 'it' or 'that'. Easier is the conjunction that usually refers to a word but sometimes also to a clause.

7.3. RELEVANCE AMBIGUITY IN APPLYING A NETWORK GRAMMAR

In addition to performing idiomatic and illocutionary disambiguation, the network text analyst in a traditional study must also specify the *nature* of the networks being encoded. For example, the network might represent the 'mental model' of an individual or the 'logical structure' of a debate (Kleinnijenhuis 1990). Moreover, syntactic links in any sample of networks must be defined in some consistent manner (e.g., as relations of identity as in Carley (1986), or of causality as in Kleinnijenhuis et al., 1997). After all, it is only insofar as one understands how concepts are linked within a network that one can meaningfully interpret measures of the 'positions' that specific concepts or relations hold within it.

Thus in addition to taking idiomatic and illocutionary ambiguity into account, the network text analyst must also apply what might be called a 'network grammar' to texts. For example, if a network were to be constructed entirely of identity relations (i.e., if a network grammar of identity relations were to be applied to a text), the researcher would need to develop 'relevance rules' for deciding which statements in the texts do indicate identity relations among themes (e.g., "the government's fiscal policy is the reduction of spending"), and which do not (possibly, "the government's fiscal policy has resulted in the reduction of spending"). Only statements (i.e., concept-relations) relevant to the grammar of the network under construction are then encoded according to this grammar.

To our knowledge no network text analyst has specified the relevance rules applied in constructing her or his networks. As a consequence, there is no precedent for outlining how one might resolve the 'relevance

ambiguity' (i.e., the relevance of some, but not all theme-relations) that will inevitably arise as network grammars are applied to samples of texts. Whatever the method of its resolution, relevance ambiguity further complicates the researcher's encoding task by adding to the previously discussed problems of resolving idiomatic and illocutionary ambiguities in the texts under analysis.

The nature of the knowledge graph denotes here especially the population the graph is referring to and also the point of view from which it is considered. In several studies on labour markets, we found a description as seen from the point of view of the employees, but also from the point of view from the employers. What does one have, especially after integrating several author graphs?

In the network statements are related. A concept having an incoming arrow can only serve as concept with outgoing arrow, when in both statements the concept is at the same theoretical level.

7.4. CONSEQUENCES

Ambiguity must be overcome. In case of just homonyms or synonyms this should not be too complicated. It has consequences however for the process of concept identification, where one is confronted with differences in definitions used by different investigators. It is possible to catch these differences by unravelling the concepts referring to these definitions.

8. Representation of Knowledge in Graphs

The representation of knowledge in graphs can be considered from two points of view: what is represented and how is it represented? In the what question the conditions as mentioned before play their role; this refers especially to the population for which the knowledge is to be represented.

8.1. WHAT IS REPRESENTED?

Apart of the conditions posed by the populations, the original empirical studies are based on knowledge that can be represented from the point of view of theory development or use in the field of policy making. This point of view is not relevant for the construction of the graph. In the representation one might want to investigate the graph at a macro level and at a micro level. The macro level deals with the social conditions that give rise to collective effects, while on the micro level actors and goals are investigated that lead to individual effects. In the representation one might be interested in only one of the two.

8.2. HOW IS IT REPRESENTED?

Available knowledge can be represented in several ways. Two ways that will often come to pass are the following. The knowledge is represented in at least three columns in a text or in a spreadsheet. The columns contain the first concept, the relation type, and the second concept. Additional columns can be reserved for sign and valence. The knowledge can also be shown in a plot. Here concepts are represented by points and relations between concepts as lines. This is on a computer screen or on paper. Figures 2 and 3 in this text are examples of such plots. A problem can be that, as one soon has a lot of points, this plot becomes disordered. Now one will have to use techniques for optimal ordering. Popping (2005) presents information based on six studies. His final graph contains 194 points. This is at least disorderly. Therefore it should be possible to represent only a part of a graph or to zoom in on a part of a graph.

The computer program KnowJoke has been developed for constructing and representing knowledge graphs.

9. Conclusions

Knowledge graphs have been introduced and the process of text analysis has been described. Problems an investigator might become confronted with are mentioned, where possible (directions to) answers are given.

Knowledge graphs are based on texts by authors. The 'translation' from text to graph however is performed by coders. The investigator responsible for the construction of the knowledge graph may want to go to the author and ask for comments, or even a verification of correctness. One step further is that this process is repeated. Now one comes close to protocol analysis that is used for the construction of knowledge or expert systems.

Finally two tests must be standed. The first one must answer the question whether the theory is represented in a correct way in the knowledge graph. The second one should inform about the usefulness of the knowledge graph for other investigators, but also for policy makers.

Notes

1. Wrightson (1976: 292) indicates a relation as cause concept – linkage – effect concept. This suggests that all relations are causal relations, which is not the case.
2. The text-analysis process itself is not described here. In this process, however, one should stay with the concepts and relations as used by the original investigator. This implies that, when a computer program is be used for this coding, this program must allow the user to follow the representational view on coding (Popping, 2000: 26). Here coding is performed according to the intended view of the original investigator. This is the opposite of the instrumental view, where in fact the process is automated and is performed according to the view of the investigator constructing the graph.

Rahmstorf (1983) has defined 39 relations to handle nominal phrases. They allow exact representations, but, due to their number, they make the knowledge graphs unworkable.

3. In empirical research the strength of a causal relation is usually denoted by a correlation coefficient or a regression coefficient. These usually do not have the value 1.
4. Kuhn (1962) distinguishes between paradigmatic and pre-paradigmatic knowledge. The former refers to a small number of premises (e.g., mathematics, physics). In sciences based on pre-paradigmatic knowledge these formal systems are not predominant.

References

Bakker, R. R. (1987). *Knowledge Graphs: Representation and Structuring of Scientific Knowledge*. Ph.D. thesis, Twente University of Technology.

Carley, K. M. (1986). An approach for relating social structure to cognitive structure. *Journal of Mathematical Sociology* 12: 137–189.

Carley, K. M. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. In: P. V. Marsden (ed.), *Sociological Methodology 1993*. Cambridge, MA: Basil Blackwell, pp. 75–126.

Carré, B. (1979). *Graphs and Networks*. Oxford: Clarendon Press.

Contas, M. A. (1992). Qualitative analysis as a public event: The documentation of category development procedures. *American Educational Research Journal* 29(2): 253–266.

Crane, D. (1972). *Invisible Colleges. Diffusion of Knowledge in Scientific Communities*. Chicago: The University of Chicago Press.

De Haan, J (1994). *Research groups in Dutch Sociology*. Ph.D.-thesis, University of Utrecht.

Ericsson, K.A. & Simon, H. A. (1994). *Protocol Analysis, Verbal Reports as Data*. Cambridge, MA: MIT Press.

Franzosi, R. (1990). Computer-assisted coding of textual data. *Sociological Methods and Research* 19(4): 225–257.

Hoede, C. & Weening, H. M. (1999). Graph theoretical analysis as tool for validation and conceptualization. In: C. Van Dijkum, D. De Tombe & E. Van Kuijk (eds.), *Validation of Simulation Models*. Amsterdam: SISWO, pp. 70–87.

Hummon, N. P. & Doreian, P. (1989). Connectivity in a citation network: the development of DNA theory. *Social Networks* 11(1): 39–63.

James, P. (1992). Knowledge graphs. In: R.P. Van de Riet & R.A. Meersman (eds.), *Linguistic Instruments in Knowledge Engineering*. Amsterdam: Elsevier, pp. 97–117.

Kaplan, A. (1964). *The Conduct of Inquiry*. Scranton: Chandler.

Kelly, E. F. & Stone, P. J. (1975) *Computer Recognition of English Word Senses*. Amsterdam: North-Holland.

Kleinneijenhuis, J., de Ridder, J. A. & Rietberg, E. M. (1997). Reasoning in economic discourse. An application of the network approach to the Dutch press. In: C.W Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Manuscripts*. Mahwah, NJ: Lawrence Erlbaum Associates, pp.191–207.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Popping, R. (1992). In search for one set of categories. *Quality and Quantity* 25(1): 147–155.

Popping, R. (2000). *Computer-assisted Text Analysis*. London: Sage.

Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information* 42(1): 91–106.

Popping, R. (2005). Representation of developments in labour market research. *Quality & Quantity* 39(3): 241–251.

Popping, R. & Strijker, I. (1997). Representation and integration of sociological knowledge by using knowledge graphs. *Social Science Information* 35(4): 731–747.

Rahmstorf, G. (1983). Die semantische Relationen in nominalen Ausdrücken des Deutschen. [The semantic relations in nominal expressions in German language]. Unpublished doctoral dissertation, Johannes Gutenberg University Mainz.

Roberts, C. W. (1989). Other than counting words: A linguistic approach to content analysis. *Social Forces* 68(1): 147–177.

Roberts, C. W. & Popping, R. (1996). Themes, syntax, and other necessary steps in the network analysis of texts. *Social Science Information* 35(4): 657–665.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Schrodt, P. A., Davis, S.G. & Wedle, J. L. (1994). Political science: KEDS – A program for the machine coding of event data. *Social Science Computer Review* 12(4): 561–587.

Shapiro, G. (1997). The future of coders: Human judgments in a world of sophisticated software. In C. W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 225–238.

Stokman, F. N. & De Vries, P. H. (1988). Structuring Knowledge in a Graph. In: G. C. Van der Veer & G. Mulder (eds.), *Human-Computer Interaction: Psychonomic Aspects*. Berlin: Springer, pp. 186–206.

Van Cuilenburg, J. J., Kleinnijenhuis, J. & De Ridder, J. A. (1988). Artificial intelligence and content analysis. *Quality and Quantity* 22(1): 65–97.

Van den Berg, H. (1993). *Knowledge Graphs and Logic: One of Two Kinds*. Ph.D. thesis, Twente University.

Wrightson, M. T. (1976). The documentary coding method. In: R. Axelrod (ed.), *Structure of Decision: The Cognitive Maps of Political Elites*. Princeton, NJ: Princeton University Press, pp. 291–332.