A computer leaning approach to obtain safety information from multi-lingual accident reports

P. Hughes, M. Figueres-Esteban, R.A.H. El Rashidy, C. van Gulijk
Institute of Railway Research, University of Huddersfield, Huddersfield, UK

R. Slovak
Federal Office of Transport, Bern, Switzerland

Accident reports provide a valuable source of data for any safety management system. In multi-lingual jurisdictions, accident reports can be provided in more than one language. For example the Swiss transport authority collects accident reports that are written in either German, French, or Italian. The unstructured nature of free-text makes it difficult to extract information from large numbers of accident reports. Machine-reading of text is an emerging area of research, however there are few instances of information being extracted from text in more than one language.

This paper introduces an ontology-based interactive learning method between a human and computer software to identify safety-related information by analysing text written in three different languages. The results of the method were analysed by fluent speakers of each language, who rated the overall accuracy of the method to be 98.5%.

The method stores and processes the data in a NoSQL graph database, which provides a powerful tool to readily integrate the analysis with other data sources, for example train movement data, passenger census data, or even comparative data from other railways.

1 INTRODUCTION

A large amount of information that is useful to safety is contained in natural language text reports, for example accident reports, hazard reports, or safety audit reports. Whist these data sources can contain valuable information, it is not easy to extract the information. Human-reading of large amounts of textual data is slow and error-prone and, if the task is divided amongst a large number of readers, then differences in interpretation can occur. Machine reading of text is an emerging area of research which has the potential to provide useful information from large text sources, although there are still a number of problems to be overcome. No prior work has been found that has attempted to extract information from safety-related documents written in more than one language.

The Swiss Federal Office of Transport (FOT) collects textual data on safety incidents that occur on the nation's transport system. Switzerland is multilingual, and the text in the incident reports is provided in either German, French, or Italian. Each report contains information that could be useful to managing safety of the system, however no existing process is known whereby the information can be collated in a way that supports safety management.

This paper describes an ontology-based approach to obtain information from 5065 incident reports provided by the FOT. Incidents were classified into a number of categories including incidents that occurred: whilst passengers were boarding trains; whilst they were alighting trains; or as a result of passengers falling down stairs, caught by closing doors, or struck by falling baggage.

2 BACKGROUND

2.1 *Safety management of big data*

Modern approaches to safety management require organisations to collect information on accidents. It is very common for these data to include text that describes where and when the accident occurred; the context within which the accident occurred (for example weather conditions; activities that were taking place at the time); what injuries and damage resulted; what proximate and underlying causes led to the accident; what risk controls failed to allow the accident to occur; and recommendations to minimise and mitigate recurrence. The purpose of collecting such data is to support decision-making processes that consider information from different sources (for example safety-critical work procedures, budget data, or legislative requirements) and take action to optimise safety management.

Professional safety management systems often collect information not only on accidents that have been observed, but also on incidents where safety risk controls have broken down but no injury or damage occurred, so called *near-miss* or *close call* events (Gnoni, Andriulo et al. 2013, Andriulo and Gnoni 2014, Macrae 2014). These accident and incident reports themselves can amount to a large quantity of data (Hughes et al., 2016a). Extracting information from these large data sources can be a challenge by itself; the problem is further complicated when the data is provided as free-text rather than structured machine-coded data. Combining data from such a large data source with data from other, potentially very large, data sources can be problematic. This data data management challenge is commonly referred to as *big data*. There is an emerging body of work describing the challenges of big data and techniques that can be used to extract useful information from the data. To obtain information that supports safety management, Van Gulijk et al. (2016) introduce the concept of Big Data Risk Analysis (BDRA), and describe the four *enablers* of BDRA:

- data and data-management,
- visualisation interface,
- analytics and software, and
- ontology and knowledge representation.

Figure 1 presents a schematic overview of the interaction of these enabling components. Each enabler is described below.

Data and data management is the initiating reason for BDRA. Modern organisations and their safety management systems collect large amounts of data with the intention of using these data to improve safety and safety management. Data may be collected from manual processes, such as workers completing forms as part of a safety-critical work process; from automatic systems, such as supervisory control and data acquisition (SCADA) systems; or from external sources, such as information on the internet regarding weather or traffic conditions. Collecting data on hazards, incidents and accidents is at the heart of modern safety management systems. The FOT has a database of thousands of reports describing all detected accidents on the transport network, including minor accidents, such as where a passenger fell over and sustained only very minor injuries. The data in the accident database is a valuable source of information that can be used to understand the causes of accidents and any underlying trends, and to determine actions that may minimise the likelihood of recurrence.

A *visualisation interface* is an essential component for understanding large data sources. Humans have a capacity for visualising concepts and their relationships, which is valuable for understanding big data sources which can contain thousands, or even hundreds of thousands of inter-related concepts. For example the safety management of an operating railway requires an understanding of concepts such as all types of rail vehicles including locomotives, carriages, wagons, and maintenance machines; all railway locations including track locations, stations, sidings, maintenance facilities; all organisations who operate within the railway including law enforcement organisations; users of the railway; routes and timetables *etc*. It is not feasible to expect staff who operate the safety management system to be able to visualise all these concepts and the multitudinous interactions between them without external support. Visual analytic tools facilitate the understanding of these concepts and include applications to help understand different categories of safety risks and the geospatial distribution of incidents and accidents. Such tools may even include visual causation models, such as bow-tie diagrams, that describe the chains of events that can lead to an accident and the risk controls that are in place to reduce the risk.

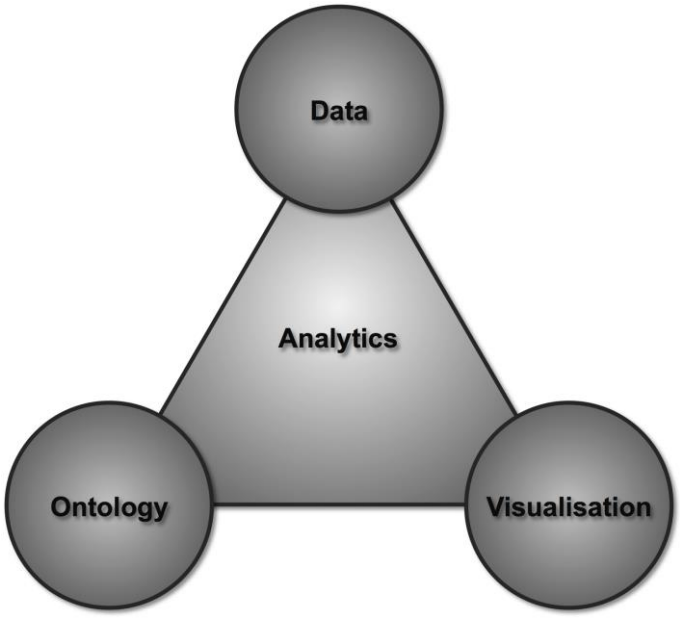

Figure 1, the four enablers of BDRA

Analytics and software are the backbone of BDRA: modern data management systems are based on software and BDRA requires large software analytical capability such as that available on modern cluster computers or provided by cloud computing services. Several software services are required for BDRA in order to:

- store data in short-term stores and long-term archives;
- organise data into meaningful data units and transport it to the locations required for processing;
- pre-process data ready for analysis and format it as required;

- analyse the data to produce results that assist safety-related decision-making;
- collate results of the analytics and aggregate them as necessary;
- present the results – in a visual form – to analysts to allow understanding of underlying hazards, risks, controls, accidents and consequences;
- iterate through analysis depending on the findings of earlier analyses;
- store results of analyses to allow subsequent analysis to expand on the results of earlier work;
- oversee and coordinate all of the above processes and distribute computer resources in an optimal way.

The software tools to support these tasks may include traditional tools such as simple graphing tools and SQL databases, as well as technologies that have emerged in the past decade such as interactive visualisation tools, graph databases and massively parallel, distributed analytical tools.

Ontology is a structured method to classify entities within a domain and the interactions between them. An entity is any item that can have properties and interact with other entities. A simple example of an entity in a railway safety context are *trains* which have properties of rolling stock class, furthermore individual instances of trains have train numbers as properties. Trains interact with stations, track and passengers. Other examples of entities include tickets (which have properties such as valid dates and routes) and railway staff (who have properties such the job function and interact with people and organisations). Entities within an ontology may also be abstract, such as dates and times. Regardless of whether an entity is abstract or has a physical form, a defining feature of all entities is that they can be referred to by themselves; this differentiates them from *relationships* that require entities in order to be meaningful. An example relationship is that a passenger can *board* a train. Both the passenger and the train are entities that can exist by themselves. However the boarding relationship requires these entities to be present in order to be meaningful: boarding cannot occur by itself. The ontology provides a structure that allows data to be joined to the key entities that are relevant to safety management. For example as an instance of data, an accident report may be joined in an ontology to the station where it occurred; to the members of staff who responded; and to the date and time where it occurred. If the ontology contains an incident causation model, such as a bow-tie diagram, the accident report can also be linked to the particular risk control breakdowns that led to the accident and the outcomes that occurred. Individual stations in the ontology may be linked to the general concept of a station; individuals may be linked to the organisations for whom they work, and so on. In this way it is possible to structure queries of the data to request accident reports that occurred at particular stations, or at stations in general; or to identify staff responders from particular organisations. The ontology may also be linked to other data in the safety management system such as audit reports, or maintenance logs. In this way it is possible to identify accidents that occurred at stations where audits have not recently been performed; or accidents at locations where particular maintenance activities have occurred. Where the ontology contains dates or chains of causation, these entities can also be queried to extract precise and meaningful information to support safety management.

2.2 *Ontologies for data management and understanding*

Smith & Welty (2001) describe the traditional understanding of the word *ontology*, as being an ancient Greek concept that addresses the fundamental nature of reality. Aristotle established ten categories of reality *viz.*: action; habit; location; passion; position; quality; quantity; relation; substance or essence; and time (Ritter & Kohonen 1989). This ontology was established to address underlying questions in ancient Greek philosophy such as *what is real?* and *what can be said to exist?* A basic method for establishing an ontology can be considered in two stages: firstly observation and conceptualisation of the real-world domain; and secondly explicit formalisation of the of the identified concepts (Evermann & Fang 2010). Such a formalisation of the world into well-defined concepts that can be reasoned about in a meaningful way provides a framework that is well-suited to computational analysis of data (Searle 2006, Smith 1998). An open question in ontology research is the degree to which an ontology needs to be complete in order to support understanding and decision-making. The concept of a *naïve ontology* is discussed by Dahlgren (1995) which is an ontology that considers only objects and their classifications. For example a naïve ontology would consider a *passenger train* as being a type of *train*, which in turn is a type of *vehicle*. However such an ontology may not consider abstract concepts such as the concept of *lateness* in relation to a train running to a timetable. This *naïve* approach underpins the resource description framework established by the World Wide Web Consortium as well as the approach taken by Noy & McGuinness (2001). Dahlgren (1995) asserts that the approach is sufficient to perform almost 80% of common-sense reasoning. Brewster and O'Hara (2007) argue that ontologies are particularly useful in well-defined domains such as individual organisations. Noy & McGuinness (2001) assert that the ontological ap-

proach to data management provides a valuable method to reuse domain knowledge. For example database queries can be stored within the ontology so that information found by one analyst can be found again later by other analysts. In this way the ontology adds to the amount of data that needs to be stored, but reduces the need for repetition of work.

2.3 *Ontologies for natural language processing*

Popping (2000) classifies natural language processing (NLP) techniques in three categories, which are in order of sophistication: *thematic*, *semantic*, and *network*. Thematic analysis considers the relative occurrence of words within the source text and can be used as a broad categorisation method to identify texts (such as accident records) that contain similar words and therefore may relate to the same broad themes. Semantic analysis expands the thematic approach by consider the function of words within a sentence (their *part of speech*, such as whether a word is a noun, a verb, or an adjective) to identify subject-verb-object triples. These triples provide the underpinning for complex formal ontology systems that develop from a mereology of objects or actions (Bateman et al. 2010; Bierwisch & Schreuder 1992; Miller & Fellbaum 1991; Ritter & Kohonen 1989).

Network analysis of text establishes the source text as a graph consisting of nodes (which usually represent single words) and edges (which describe relationships between the nodes such as co-occurrence of words within a single sentence). As such the network approach establishes a form of ontology of text within a document, although such an ontology is based on abstract concepts (using words as labels for objects) and therefore fundamentally differs from naïve ontologies which consider the nodes in the ontology as representations of the objects themselves. Miller and Fellbaum (1991) note that a disadvantage with network analysis can be the need for additional software tools such as graphing and network analysis tools, which can complicate the analysis if there is a need to transfer data between separate software products. The introduction of graphical text analysis tools introduces a new domain of research, for example as discussed by Figueres-Esteban et al. (2015).

For general text analysis, Miller & Fellbaum (1991) propose an ontology of 26 basic concepts, these concepts can be used to form a basis for domain-specific ontologies; they are:

- act, action, activity;
- animal, fauna;
- artefact;
- attribute, property;
- body, corpus;
- cognition, ideation;
- communication;
- event, happening;
- feeling, emotion;
- food;
- group, collection;
- location, place;
- motive;
- natural object;
- natural phenomenon;
- person, human being;
- plant, flora;
- possession, property;
- process;
- quantity, amount;
- relation;
- shape;
- society;
- state, condition;
- substance;
- time.

Finally, Van Gulijk et al. (2016) conclude their discussion of the use of ontologies in computer science by providing the following three counsels. Firstly there is no such thing as a perfect ontology; rather there can be a number of alternative ontologies that serve the same purpose. Secondly, effective ontologies are developed iteratively, perhaps as users interact with an ontology and seek more detail from it. Thirdly, ontologies that relate to physical objects are the easiest to create; ontologies that relate to abstract concepts can provide conceptual difficulties for both the ontology builder and the analyst using the ontology.

3 METHOD

Extraction of information from the text was performed in a process that consisted of four mains steps. The first step being the preparation of the data to allow for efficient completion of the later steps. The second step was the identification of key terms in the text and the construction of an ontology to make explicit the relationships between the terms. The third step involved the execution of queries to identify records that correspond with each of the categories of incident; this is an automated step performed by software. The final step was a review of the returned results and consequent refinement of the ontology and queries until an acceptable result was achieved. Each step is described in detail below.

3.1 *Data preparation*

The accident reports provided by the FOT were imported into a graph database. Graph databases

structure data as nodes and edges, rather than using the structure required by Structured Query Language (SQL), which has been a prevalent structure for databases some decades. As such, graph databases belong to a class of databases known as *NoSQL*.

Data relating to an individual incident was imported as a single node in the database: 5065 record nodes were created in the database. An automatic process was used to analyse the text in the source records and create a new node for each sentence in the text. During this process a simplifying assumption was made that a full-stop followed by a space (.) would always mark the end of a sentence. The sentence nodes were linked to the node that contained the data from the record.

In alphabetic languages, such as the European languages used in Switzerland, the basic unit of text analysis is a word. Fundamentally the process to establish meaning from text is performed by analysing the occurrence, frequency, and colocation of words or groups of words. In this work each sentence was broken into individual words: punctuation marks were separated from words by inserting spaces. Each word was converted to lower-case text and added as a word node in the graph. During this process the frequency of occurrence of each word was stored in the word node. Colocations of pairs of words were shown by the creation of an edge marked *next*; the edge also recorded data on the frequency of the colocation of the pair. This process of creating word nodes and *next* relationships is the same as the process applied by Lyon (2015).

3.2 *Step 2: Ontology learning*

The source data were analysed to identify terms (words and bigrams) for inclusion in an ontology of items. Candidate terms were identified by calculating the TFIDF score for each word in the source text. Subsequently, each bigram was considered to be a single token and a TFIDF score was calculated for the bigram. All terms from all the source records were compiled in a table of descending TFIDF score and presented to an analyst for consideration in the ontology.

The analyst reviewed the list of candidate terms, starting with those with the largest TFIDF score, and selected the terms that appeared to be relevant to each category of incident. Since the TFIDF ranking is intended to list terms in order of relevance, the analyst worked through the list until reaching a point where it was determined the that the terms were generally irrelevant and no further terms would be considered. The analyst in this trial spoke only English and had no fluency in any of the source languages (German, French, nor Italian) and used simple on-line translation tools to understand the candidate terms. Terms were selected based solely on the analyst's understanding of railway operations and safety. After identifying terms, the analyst created an ontology that joined matching or similar concepts. The ontology allowed equivalent terms in different languages to be joined to the same node. For example, in records written in German, the words *ambulanz* and *krankenwagen* were both used to refer to an ambulance and were linked to the *ambulance* ontology node. Similarly the French term *ambulance*, and the Italian term *ambulanza* were linked to the same node. A simplification was made to link plural terms to singular ontology concepts as a simple form of lemmatisation of the text.

The ontology learning process was limited to the creation of only a naïve ontology: only a single type of relationship was defined indicating that each ontological element can be *a type of* another element; for example *a woman is a type of person*. The ontology did not contain relationships such as *a door is a part of a train*; nor *a train can arrive at a station*.

3.3 *Step 3: Execution of queries based on the ontology*

Queries were performed on the data in the graph database to identify records related to each category of incident. The queries were started at the ontology nodes that defined each category of incident and traced via edges in the graph to identify records that contained terms relating to the incident category.

3.4 *Step 4: Iteration and reporting*

As noted above, the process of ontology creation is iterative. After execution of each query, the analyst reviewed the results to determine whether the records correctly corresponded to the category. Since the analyst had no fluency in the languages used to write the records, the TFIDF ranking process was reapplied to only the records returned as a result of each query. The analyst reviewed the terms occurring in the subset of records, using simple translation tools again, to determine whether terms were occurring that did not appear to relate to each query.

4 RESULTS

This section presents the results of each stage of the analysis.

4.1 *Data preparation*

The 5065 records were loaded into the database and a total of 16,419 unique word nodes were created. Relationships were created to show collocations of words. Figure 2 shows an example of the pair of

words *dame* and *âgée* with the NEXT edge joining them. The figure shows that the word *dame* occurs 620 times in the source text, the word *âgée* occurs 202 times and, as a colocation, *dame âgée* occurs 150 times.

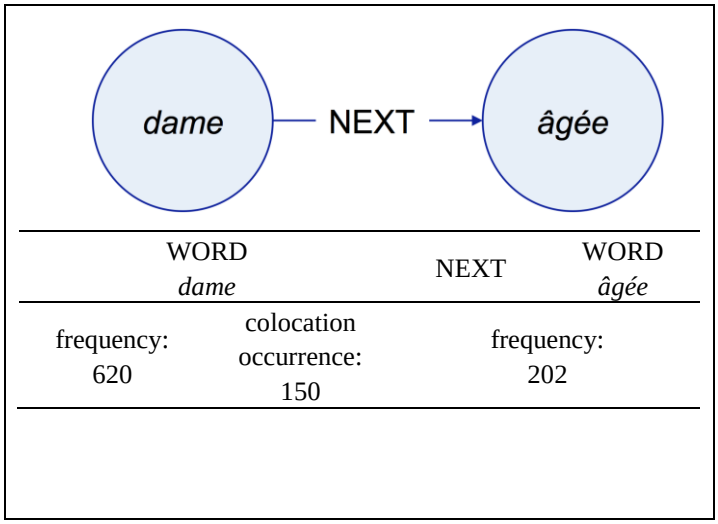

Figure 2: Example of the colocation *dame âgée* and the *next* edge linking the words

4.2 *Step 2: Ontology learning*

The process of TFIDF ranking returned 82,726 candidate terms, being 16,419 single words terms (one for each word node in the database) and 66,307 bigrams. From this list the analyst identified 389 terms that appeared to be related to the specified categories of incident. It is notable that, in some cases, the TFIDF calculation produced similar scores for terms with similar meaning in different languages. For example of the 82,726 identified terms, the term *ältere dame* (German for *elderly lady*) was ranked 314th in the list; the term *dame âgée* (French for *elderly lady*) was ranked 316th on the list.

The analyst constructed an ontology based on the identified terms. For ease of analysis the ontology was limited to only objects and actions and structured in two layers. Table 1 shows example the entities included in the ontology.

4.3 *Step 3: Execution of queries based on the ontology*

Queries were designed and executed for each category of incident based on the terms identified during Step 2. The starting point for each query was the ontological entries that define the key entities being sought in the query. For example to identify records related to injuries caused to passengers by closing doors, the query identified the ontology nodes relating to *passengers*, *closing doors*, and *injury*. The query then identified the terms relating to those concepts, followed by the records that contained those terms. Figure 3 shows an example of records that relate to the ontological concepts of an elderly person and stairs. The query can be thought of executing from the top of the diagram down: firstly the ontology elements for *stairs* and for a *person* – and in particular an old person – are identified. These ontology elements are traced in the query to instances of words that define them, for example the words *Treppe* (being a German word for stairs) and *scale* (Italian for stairs). In turn, these words are traced to instances of accident reports where the words occur. It can be seen that the plural term for men in Italian (*uomini*) has been linked to the singular term in the ontology.

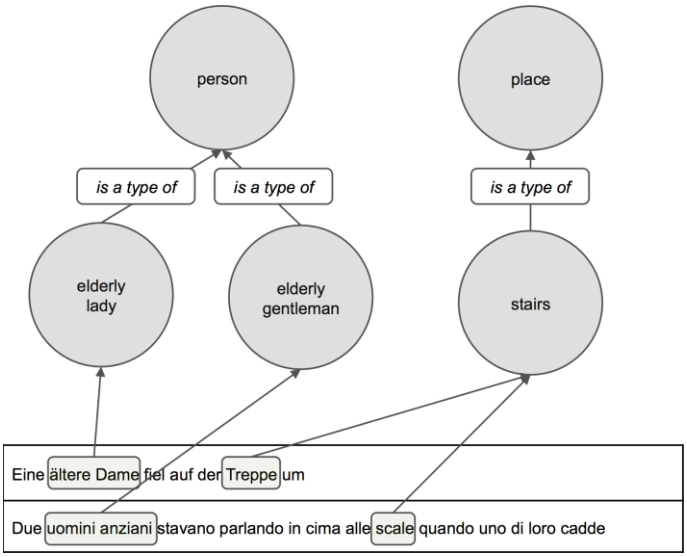

Figure 3: schematic diagram of an example query

4.4 *Step 4: Iteration and reporting*

The results of the queries were reviewed by the analyst for correctness and the process in Steps 2 and 3 was iterated to refine the terms, ontology and queries to create results that appeared to better align with each category of incident. After iteration of the process, the results of the queries were presented to fluent speakers of each language for review. The reviewers were independent of the process and assessed each record entirely on whether it appeared to correctly describe an event belonging to each category. As such the reviewers' assessed only the rate of true positive matches compared with the overall number of records found to match for each category. The overall results from all reviews indicated that the number of true positives was 98.5% of all positive results returned by the queries.

5 DISCUSSION

The overall finding is that this preliminary study has demonstrated the potential for the technique to be a powerful tool for identifying specific instances of safety-related events from multi-lingual accident re-

Ontology upper layer	Ontology lower layer
person	*doctor, self, customer, person, driver, passenger, months old, years old, baby, female, old, elderly, young, male, man, lady*
places	*line, station, pavement, hospital, ground, platform, stairs*
actions	*hit, medical, injure, get out, enter, fall, rush*
body parts	*foot, head, arm, leg*
vehicle	*carriage, vehicle, ambulance, tram, train, bus*
direction	*backwards, direction, in between, in front*
items	*bag, alcohol, drugs, stairs, footboard, customer information system, ticket, door*

Table 1: example ontology entries

ports. The result of 98.5% true positives in all results returned is very strong, especially considering that the analyst did not have fluency in any of the source languages and used only simple translation tools. At this stage, however, it is not clear how many false negatives results occurred as a result of the queries (i.e. records that described one of the categories of incident but were not identified by the queries). Further work would be required to determine the overall accuracy of the process.

The trials of the technique to date have been limited to only a few categories of incident that were specified before the process of ontology creation. Since the process is based on the occurrence of terms in the text, it appears possible that the process could be started by examining the text to determine what categories of incident are being describe, i.e. a *bottom-up* exploration of the text to identify categories rather than starting the analysis with specific categories of incident (a *top-down* analysis). Such a bottom-up approach may be valuable to identify unexpected trends in the data that could not be presupposed; for example unexpected categories of incident caused by emerging technology.

The ontology developed during the process is based on the terms that occur in the text and, as such, it appears that the technique should be applicable to other sources of text that can support safety management such as audit reports, inspection reports; or even to general sources of textual data.

Another limitation of the technique is that it is based on the occurrence of simple concepts being described in the text, but does not consider compounded concepts such as negation. For example when stairs are mentioned in the text, it is presumed that stairs are relevant to the incident, however a record may refer to stairs even though they are not relevant to an incident, (e.g. *an old man, whom I had previously seen on the stairs, fell whilst entering the train*). To address this issue the ontology would need to be updated to include complex ontological concepts. Further work is being carried out to align this study with our previous work (Hughes et al., 2016b) to address these issues in the technique.

6 REFERENCES

Andriulo, S. and M. G. Gnoni (2014). Measuring the effectiveness of a near-miss management system: An application in an automotive firm supplier. *Reliability Engineering & System Safety* 132(0): 154-162.

Bateman, J. a. et al., 2010. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 17: 1027–1071.

Bierwisch, M. & Schreuder, R., 1992. From concepts to lexical items. *Cognition*, 42: 23–60.

Brewster, C. & O'Hara, K., 2007. Knowledge representation with ontologies: Present challenges-Future possibilities. *Int. J. Human Computer Studies* 65: 563–568.

Dahlgren, K., 1995. A linguistic ontology. *Int J Human-Computer Studies*, 43: 809–818.

Evermann, J. & Fang, J., 2010. Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems* 35: 391–403.

Figueres-Esteban, M., Hughes, P. & Van Gulijk, C., 2015. Visualising Close Call in railways: a step towards Big Data Risk Analysis. In Fifth International Rail Human Factors Conference: 725–734. London: RSSB.

Gnoni, M. G., Andriulo, S., Maggio, G., & Nardone, P. (2013). Lean occupational safety: An application for a Near-miss Management System design. *Safety science*, 53, 96-104.

Hughes, P., Figueres-Esteban, M., Van Gulijk, C. (2016a) Learning from text-based close call data. *Safety and Reliability: SaRS Journal.* ISSN 0961-7353

Hughes, P., Figueres-Esteban, M., Van Gulijk, C. (2016b) *From negative statements to positive safety*. In: Risk, Reliability and Safety: Innovating Theory and Practice: Proceedings of ESREL 2016. CRC Press. ISBN 9781138029972

Lyon, W. (2015). *Natural Language Processing With Neo4j - Mining Paradigmatic Word Associations*. Retrieved from http://www.lyonwj.com/2015/06/16/nlp-with-neo4j/. Retrieved 05 May 2017.

Macrae, C. (2014). Close Calls: *Managing Risk and Resilience in Airline Flight Safety*, Palgrave Macmillan.

Miller, G. a & Fellbaum, C., 1991. Semantic networks of English. *Cognition*, 41: 197–229.

Noy, N. & McGuinness, D., 2001. Ontology development 101: A guide to creating your first ontology. *Development*, 32: 1–25.

Popping, R., 2000. *Computer-Assisted Text Analysis*. London, SAGE.

Ritter, H. & Kohonen, T., 1989. Self-organizing semantic maps. *Biological Cybernetics* 61: 241–254.

Ritter, H. & Kohonen, T., 1989. Self-organizing semantic maps. *Biological Cybernetics* 61: 241–254.

Searle, J.R., 2006. Social Ontology: Some Basic Principles. Papers. *Anthropological Theory* 6: 12 -27.

Smith, B. & Welty, C., 2001. Ontology: Towards a New Synthesis. In Proceedings of the international conference on Formal Ontology in Information Systems: 3–9.

Smith, B., 1998. Basic concepts of formal ontology. In *Formal Ontology in Information Systems*: 19–28. Amsterdam: IOS press.

Van Gulijk, C., Hughes, P., & Figueres-Esteban, M. (2016). The potential of ontology for safety and risk analysis. In *Proceedings of ESREL 2016*. CRC Press. Chicago