Extracting safety information from multi-lingual accident reports using an ontology-based approach

Abstract

This paper describes an approach to extract meaning from multi-lingual free-text safety incident reports. A sample of 5065 safety incident reports from the Swiss Federal Office of Transport were used in the study. Each report was written in either German, French or Italian natural language. An interactive learning approach between a human and computer software was undertaken to identify key terms in the text that are relevant to discovering meaning. A multi-lingual ontology was created to join meaningful semantic patterns and identify specific classes of safety incident on the railway, including injuries occurring whilst passengers were boarding or alighting from vehicles, falling down stairs, struck by closing doors, or struck by objects such as suitcases. A graph database was used to query the text records via the ontology and identify reports of incidents in each class, regardless of the language used in the report. Fluent speakers of each language – German, French and Italian – reviewed the results to confirm true positive results and detect false positives. The performance of the process varied across languages and incident types, however the overall true positive rate was determined by the fluent speakers to be 98.5%.

1.0 Introduction

This paper describes a process to determine how computer-based techniques can be used to accurately categorise large numbers of multi-lingual safety incident reports into pre-determined categories. The analysis was undertaken on a sample of 5065 incident reports provided by the Swiss Federal Office of Transport (FOT) covering events that occurred between the years 2000 and 2016. Each report contains a free-text description of an incident and is written in one of the three main languages spoken in Switzerland, *viz.* German, French and Italian. The number of source records, and the diversity of languages makes human analysis cumbersome and potentially unreliable. This work was undertaken to determine whether it is possible to use computer analysis to extract useful safety information from the textual description of the events; the work was performed by an analyst who had no fluency in any of the source languages.

When incidents occur, each event is responded to individually, the FoT collects statistics on the number of events that have occurred. However, since key information regarding events is provided in text descriptions there is difficulty in providing aggregate data on categories of events. To improve safety management, there is a need to categorise event descriptions by information in the text and report on trends in such categories. The study set out to identify incident reports in any of the source languages that relate to injury occurring as a result of passengers:

query 1: alighting vehicles;
query 2: falling down stairs;
query 3: boarding vehicles;
query 4: being trapped by closing doors;
query 5: being struck by falling bags.

The approach used in this study involved importing the source text into a graph database. The text was then cleansed to separate words from punctuation marks and convert all words to lower-case. Key terms used in the incident records were identified. An ontology was constructed to show the relationships between these terms. Finally, queries were run to find which records contain descriptions relating to each query.

The contribution of this study is the use of a single graph database to connect text data with an ontology; simultaneous analysis of multi-lingual text in a single database; use of common querying to extract data from multi-lingual sources; and explicit use of human-machine iterative interaction to provide information for ontology learning. The application of the technique is also novel: as well as providing the theoretical method; the work in this study provided results that are useful to the real-world management of safety where previously no data were available.

Section 2 of this paper discusses prior work that supports this study, in particular computational analysis of safety incident reports; machine translation; network analysis of natural language; and graph databases and ontology learning. Section 3 describes the method used in this study, with the results presented in Section 4. Section 5 provides a discussion of the results and their implications. A conclusion is given in Section 6.

2.0 Literature Review

2.1 Computational analysis of safety incident reports

The body of work on computational analysis of text, contains examples from the health and medical domain (for example Toyabe, 2012; Chase, Mitrani, Lu and Fulgieri, 2017) aviation domain (for example Saeeda, 2017), to identify information from highway accidents (for example Mannering, Shankar and Bhat, 2016) and to obtain information from social media (for example Proctor, Vis and Vos, 2013).

Popping (2000) developed a categorisation of computational text analysis approaches, describing them as being either *thematic*, *semantic*, or *network*. The predominant approach used by researchers is the thematic approach (Donaldson, Panesar and Darzi, 2014; Church and Hanks, 1990; Dale, Moisl and Somers, 2000). However, Taler et al. (2013) provide a noteworthy attempt to extend the approach to a semantic analysis. Wu and Heydecker (1998) identify several key problems with analysing road accident text leading them to state that there is a requirement for "*human knowledge that is essential for comprehension*" alongside computational analysis to overcome comprehension issues, colloquialisms, ill-formedness etc. They report some successes where these problems have started to be overcome. Kayser and Nouioua (2009) report similar findings.

Despite the wide application of computational analysis, there is not a large body of literature describing natural language processing (NLP) techniques being applied for the analysis of safety incident reports. Hughes et al. (2018) proposed a technique for categorising text-based hazard reports in accordance with a bow-tie diagram. Tanguy et al. (2016) describe the use of NLP techniques for analysing safety incident reports within the aviation industry and its

current constraints, similar approaches are described by Ittoo et al. (2016). The authors explain that their approach requires a lexicon of terms to be established prior to analysis and therefore that the technique "*cannot be used for the identification of emerging threats*". This problem can however theoretically be overcome by mining for patterns within the text data, though the authors reported mixed results for this. Tanguy et al. (2016) discuss active learning as an alternative approach to identify key terms and describe the process as "*a smart usage of the expert's time by submitting to his judgment only the difficult or borderline items. This can only be done through an iterative process with a dose of interaction with the user*".

2.2 Machine translation

Attempts to find a solution to the problem of how to understand text written in different languages can be traced back to Descartes (1629) and his discussion of an "*artificial universal language*". Statistical machine translation (SMT) was posited by Weaver (1955) and its use continued with the widespread accessibility of computers with Hutchins (1978) providing a turning point between the theory and practice of machine translation.

Since its inception, SMT has remained the predominant approach for machine translation. As described by Brown et al. (1988) and Brown et al. (1990), the approach consists of separating the source text and using a glossary, as well as information about the context of the text, to arrange the translated words into the correct sequence to produce the target sentence. However, the approach can suffer from inaccuracy; Brown et al. (1990) report a loss of meaning in 16% of cases between French and English; two languages that were selected on account of their similarity. This statistical approach has subsequently been developed, but SMT is generally regarded as having its shortcomings (Koehn, 2005; Brown et al., 1988; Brown et al., 1990; Och, 2003).

An emerging approach that is being adopted as a more accurate successor to SMT is Neural Machine Translation (NMT) (Wu et al., 2016). This approach consists of a neural network divided into three parts: an encoder network, an attention network and a decoder network. The encoder network transforms the original sentence into a series of vectors. The attention network allows the decoder to "*focus on different regions of the source sentence during the course of decoding*"; the decoder network then uses the list of vectors to produce symbols in turn. Whilst more sophisticated than SMT, NMT has also encountered problems some of which are similar to SMT such as out-of-vocabulary words. On the whole, however, NMT has provide to be more accurate than SMT; for example, Wu et al. (2016) demonstrated a reduction in translation error of 60% compared to Google's phrase-based production system.

2.3 Network analysis of natural language and graph databases

An emerging area of research for understanding natural language is the use of network analysis. Popping (2000) notes that "*network text analysis originated with the observation that after one has encoded semantic links among concepts, one can proceed to construct networks of semantically linked concepts*". Popping goes on to note that the use of the terms differs from that used in conventional network theory but rather follows "*the terminology used in*

the main texts on network text analysis". A network approach was applied by Figueres-Esteban et al. (2016) who describe network text analysis as: "*a method that represents text as a graph: the words or concepts are the nodes, and their relationships are the edges*". In their work, the researchers used visual analytic techniques on railway safety incident reports to uncover patterns in otherwise unstructured text. Other similar approaches for network analysis of text are emerging (for example Sadoddin and Driollet, 2016; Patel and Dharwa, 2017). Figueres-Esteban et al. (2016), explain "*visual analytic methods cannot be used without additional interpretation by a risk specialist*". Further to this point, a comparison was done to demonstrate the difference in workload between visual analytic (VA) assistance and traditional methods and it was concluded that the workload of the risk specialist might be greatly reduced with VA assistance.

Overall the literature show that NLP techniques have been able to demonstrate some successes, although accuracy of the results is a consistent problem, and machine translation techniques are currently incomplete. There appears to be a consensus amongst the literature that human input is necessary as part of the process of extracting meaning from text. There is an emerging area of research in network text analysis and VA that provides a rich avenue for further research.

2.4 Ontology learning

An ontology is a structured representation of the key entities that exist within a domain and the relationships between them; the notion of an ontology was established in ancient Greece where philosophers sought to understand the fundamental nature of what can be considered to exist (Smith & Welty, 2001). In its general sense, the entities within an ontology can be any imaginable concept for managing knowledge within a domain, including abstract concepts such as *emptiness*. For practical use, Dahlgren (1995) establishes the notion of a naïve ontology, which is an ontology that considers only objects and their classifications. Ontologies are used in knowledge management in domains where there are large numbers of concepts with complicated interactions. Particular examples of well-established ontologies can be found in the domains of biology and medicine (Liu et al., 2011; Hoxha et al. 2016).

The science of managing knowledge in ontologies is still emerging; the literature do not provide a consensus on what is the definition of an ontology, nor how relationships between entities should be represented. It is generally agreed that an ontology can contain so-called taxonomic relationships (an "*is a*" relationship) although the literature are confused on the meaning of even this basic term and do not agree whether it indicates classification (for example a tiger *is a type* of animal) or instantiation (for example Algeria *is an instance* of a country). Numerous differing relationship schemas can be identified in the literature (Colace et al, 2014; Ruiz Martínez et al., 2011; Smith et al., 2005; Alvarado et al., 2016).

As a representation of knowledge, a schema needs to be developed to describe the entities and relationships in a domain. It is possible to develop an ontology entirely manually although the process is laborious (Ruiz Martínez et al., 2011; IJntema et al., 2012). When developing an ontology from text sources, a number of automated text

analysis processes exist to help the development of the ontology (these are referred to as semi-automated processes) (Kang et al., 2014; Hoxha et al., 2016). Ruiz-Martínez et al. (2011) note the need for manual intervention in the ontology learning process since "*techniques for learning domain ontologies from free natural language text have important drawbacks".* Research into the techniques for semi-automated ontology learning techniques continues and a number of techniques are proposed, such as latent Semantic Analysis, term subsumption, contrastive analysis, inductive logic programming, logical inference, neural networks, hidden Markov models, conditional random fields, maximum entropy modelling support vector machines (Colace et al, 2014; Ruiz Martínez et al., 2011). Currently there is no consensus on the method to develop an ontology, the only common theme found in the literature is that there is a need for human intervention to support any automated techniques that are applied. Establishing an ontology is an incremental process (Sánchez Ruenes, 2007), and van Gulijk et al. (2016) caution that there is no such thing as a perfect ontology; rather there can be a number of alternative ontologies that serve the same purpose.

3.0 Method

The method applied in this work is based on the five-step approach described by Hughes et al. (2016) for extracting safety-related information from incident reports:

1. Text cleansing, tokenising, and tagging: these initial cleansing processes prepare the text for later processing.
2. Ontology construction: creation of semantic patterns in an ontology.
3. Clustering: creation of groups of records that are semantically similar.
4. Text analysis and method refinement: analysis of the results to provide refinement of cleansing, parsing and clustering.
5. Information extraction: automated extraction of safety-related information.

To apply network analysis techniques and to benefit from the knowledge that is available from human experts, this approach was updated for this analysis to the following five-step procedure:

1. The text descriptions of the incidents were imported into a custom-built NoSQL database.
2. Automated text analysis techniques were used to identify terms in the text that appeared to be significant within the corpus of incident reports; these terms were reviewed by a human analyst.
3. To allow for the network analysis discussed in Section 2.3, an ontology of semantic patterns was created within the database to represent key concepts and actions related to safety of the railway and the relationships between them.
4. The ontology was used to structure and perform queries on the source records; candidate records were identified for each class of incident. Steps 2, 3 and 4 were repeated iteratively until the ontology appeared appropriate for extraction of safety-related information appropriate to each query.
5. The results were reviewed by fluent speakers of each of the languages used in the source records.

The steps of this process are described in the sections below.

3.1 Step 1: import source data into a No-SQL database

The source records were provided by the FOT in a comma-separated values (*.csv*) document. A total of 5065 records were provided; Table 1 shows the data fields for each datum and example content.

Data field	Example data
Year	2014
Event-ID	EAQU5
Imfrastructure manager	SBBI
Railway undertaking 1	SBBP
Railway undertaking 2	CINT
Occurrence carrier	TU1
Transportation mode	Normalspur
Location from	Bern
Location to	Thun
Time	04/01/2014
Date	13:05:00
Description	Beispieltext, der ein Ereignis beschreibt, das auf der Bahn stattgefunden hat. Namen wurden entfernt, um die Informationen privat zu halten.
Measures taken	*(no data provided)*
Investigation result	*(no data provided)*
Passenger fatalities	0
Passenger serious injuries	0
Passenger light injuries	1
Staff fatalities	0
Staff serious injuries	0
Staff light injuries	0
Others fatalities	0
Others serious injuries	0
Others light injuries	0
Unauthorised fatalities	0
Unauthorised serious injuries	0
Unauthorised light injuries	0
Costs of material damage - others	0
Costs of material damage - infrastructure	0
Costs of material damage - rolling stock	0
Classification WHAT	Personenunfall
Classification WHERE	Im Zug
Classification WHY	Übrige Fehlhandlungen (Reisende + Dritte)
Classification WHO	Reisende(r)
Railway vehicle in motion	Ja
Dangerous goods on board	Nein
Dangerous goods released	*(no data provided)*
Report according to RID	*(no data provided)*
Level crossing type	SCHA
Investigation report Nr.	Nein
Cost of environmental damage	0
Train type	Reisezug
Classification WHO primer responsible	Reisende(r)
Level Crossing ID	8901
Line interruption >6 hours	1
Total fatalities	0
Total serious injuries	0
Total light injuries	1
Personeschaden	0
Material damage >=100000	0
Material damage >=200001	0
Material damage total	0

Table 1: Data fields and example data provided for the analysis

These data were imported into a graph database. Graph databases structure data as a graph using *nodes* and *edges*, rather than using the structure required by Structured Query Language (SQL), which has been prevalent for

databases some decades. As such, graph databases belong to a class of databases known as *NoSQL* and have properties that make them suited to analysing large and complex data sets (Rashidy et al, 2017).

Data relating to an individual incident was imported as a single node in the database; therefore 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. It was assumed 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.

The basic unit of text analysis is a word; the process to establish meaning from text is performed by analysing the occurrence, frequency, and co-location of words or groups of words. In this work each sentence was divided into individual words: punctuation marks were separated from words by inserting spaces, for example a space was inserted between a word and a full-stop that followed it, or between a word and a following comma. Each word was then 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. Co-locations of pairs of words are shown by the creation of an edge marked NEXT; the NEXT edge also records data on the frequency of the co-location 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: apply automated text analysis techniques to identify key terms

Key terms that were used in the text were identified by using the *term frequency inverse document frequency* (TFIDF) method of word ranking. It is not possible to locate the source of the TFIDF technique with certainty and applications of the technique vary between texts, however a common formulation is given in Equation (1).

$$TFIDF = \frac{f_T}{N} \ln\left(\frac{R}{R_T}\right) \qquad (1)$$

Where:

f_T:	the frequency of occurrence of term T within a record;
N:	the total number of terms in the corpus;
ln(…):	the natural logarithm;
R:	the number of records in the corpus; and
R_T:	the number of records that contain term T.

The TFIDF method provides a score to each word that increases with the frequency of occurrence of the word in the corpus, but reduces for words that are used in many sentences. In this way the method attempts to identify key terms from the text whilst ignoring commonly occurring words that provide little semantic meaning, so called *stop words*. For example, in English words that are usually considered to be stop words include: *the*, *of*, *a*, *to*. A modification of the TFIDF method was also used to identify pairs of collocated words (bigrams) in the text that appeared to be key terms; the modified process worked the same as the single-word TFIDF process, but considered every pair of words in a sentence to be a single token, and applied the standard TFIDF calculation to every token.

A list of words and bigrams, ranked in accordance with the TFIDF value was presented to the analyst for consideration. The analyst undertaking the work had no knowledge of any of the languages that were used for the text descriptions, instead simple online translation tools to provide an English translation of the term. The analyst worked systematically down the ranked list considering each word and bigram and obtained a translation using online tools. The analyst made a judgement regarding whether the words and bigrams would provide valuable information to identify safety-related events that were relevant to each incident category. A new node, a TERM node, was created in the database for each word that was judged relevant to the purpose. Each TERM node was linked to corresponding WORD nodes.

3.3 Step 3: establish an ontology of terms

Identified terms in the database were linked in an ontology structure. An ONTOLOGY node was created in the database for each concept that was represented by the identified terms. For example, within the source text there are a number of terms that refer to the concept of a *train*; the German term (*Zug*), the French term (*train*) and the Italian term (*treno*). Each of these TERM nodes was linked to the single ONTOLOGY node that represents the concept of *train*. During this process, where the analyst identified terms that have equivalent meaning, either synonyms within a single language, or terms that are equivalent between different languages, these terms were linked to a common ontology node. Categorisation of terms within the ontology was performed based on the analyst's domain knowledge. The purpose of an ontology is to provide a structured method of relating concepts that is consistent with the analyst's understanding of the inter-relationships between the concepts; as such an ontology may have many layers. For ease of analysis to demonstrate the ontology developed for this work was limited to only two layers, Figure 1 shows example entries in the ontology where a lower element in the ontology is a type of the upper element it is connected to.

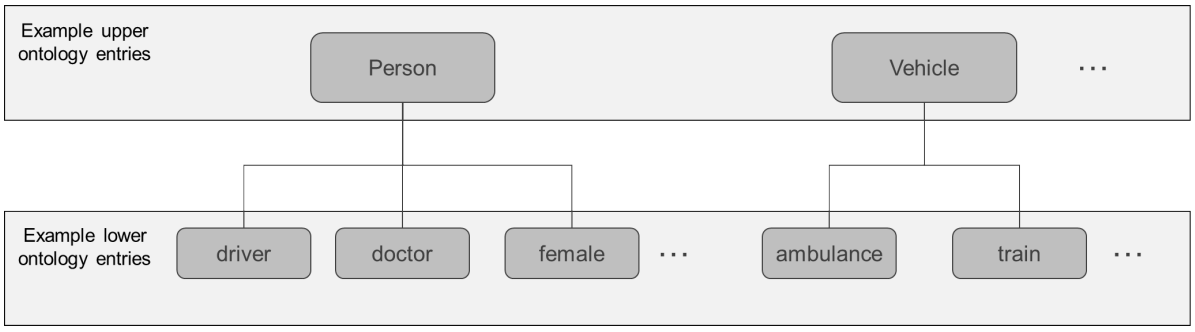


Figure 1: Example entries in the two-layer ontology

The ontology used in this work was created entirely from terms that were found in the source text, as such the ontology was build using a *bottom-up* process (from concepts in the source text up to the ontology and then to the upper concepts). If applied to a sufficient large corpus of source text, it can be imagined that over time the ontology could be expanded to provide a comprehensive description of the railway and its operations.

The overall structure of the database is shown in Figure 2, which illustrates the types of nodes and the relationships between them. It can be seen that RECORD nodes are to linked to SENTENCE nodes by a CONTAINS

relationship, indicating that the record contains the sentence. In turn SENTENCE nodes are linked to the WORD node for each word that is contained in the sentence; again a CONTAINS relationship is used. As described in Section 3.1, co-located WORD nodes are related to each other with a NEXT relationship. Relevant terms for querying are made up of words, therefore there are FORMS relationships between WORD nodes and TERM nodes. The purpose of including TERM nodes is to allow for extendibility of the structure to allow for terms to be comprised of multiple words, for example, in an English language database, the term *level crossing* is made up of two words. In such case it would be necessary to join both WORD nodes to the TERM node. TERM nodes are linked to the ONTOLOGY nodes with relationships that denote the language of the term: for terms in German, the relationship GERMAN_MEANS is used; similar relationships are used for French and Italian terms. Lower level elements in the ONTOLOGY are linked to the upper elements with an IS_TYPE relationship indicating that the lower element is a type of the upper element, for example a *train* is a *type* of vehicle. In general, it is possible to link ontology elements with any type of relationship include relationships that indicate that one element *is a part of* another. It is also possible for ontology elements to be joined in with non-taxonomical relationships, such as to indicate that trains *travel between* stations, or that passengers *must be in possession of* a ticket. In this study only *is a type* of relationships were used between ONTOLOGY nodes.

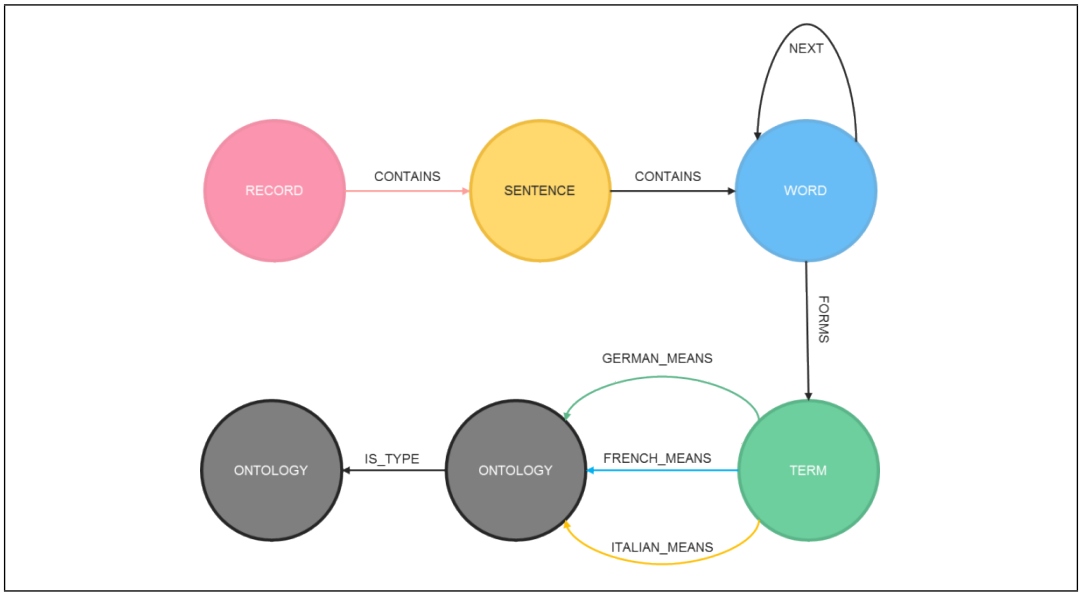


Figure 2: Overall structure of the graph database

3.4 Step 4: perform queries

The database was used to perform queries to identify records that contained safety-related information relevant to the questions set out in Section 1; queries were structured in accordance with concepts that occurred in the ontology. In this way, records could be identified regardless of the language that the original text was written in. For example, to identify records where old people were injured on trains, the ontology items for *old people* and *trains* were used

as the basis for starting the queries. From the relevant ontology items, terms, words, sentences, and eventually records were identified.

The results of the query were reviewed by the analyst to ensure that the records returned by the queries provided results that appeared to be applicable to each query. Key terms were reviewed to determine whether there were any anomalous results being returned; again online translation tools were used to support this work. The review led to WORD and ONTOLOGY nodes and their relationships being updated in the graph database. The process of identifying key terms, ontology creation, and running queries was repeated iteratively until the results of the queries appeared to be correct.

3.5 Step 5: review by fluent speakers of each language

The results from each query were separated into three groups: one for each of the languages the source text was written in. A fluent speaker of each language reviewed the accuracy of a randomly selected sample of the identified records. Accuracy was assessed by the reviewer deciding whether each record in the sample correctly described the event relating to the query. For example, the German speaker reviewed a sample of the German-language records obtained from Query 1 to determine whether each record described an event where a person was inured whilst alighting vehicles.

4.0 Results

The results for each step of the method are shown in the following sections.

4.1 Step 1: import source data into a custom-built database

The source text was imported to a Neo4j graph database (version 3.2.2). A total of 14,039 SENTENCE nodes were created from the 5065 source records. Figure 3 shows an example of two RECORD nodes linked to their corresponding SENTENCE nodes. The RECORD nodes appear in red with edges marked CONTAINS linking the records to the SENTENCE nodes shown in yellow. The text from these example records is shown in Table 2, which also includes translations from the original text into English.

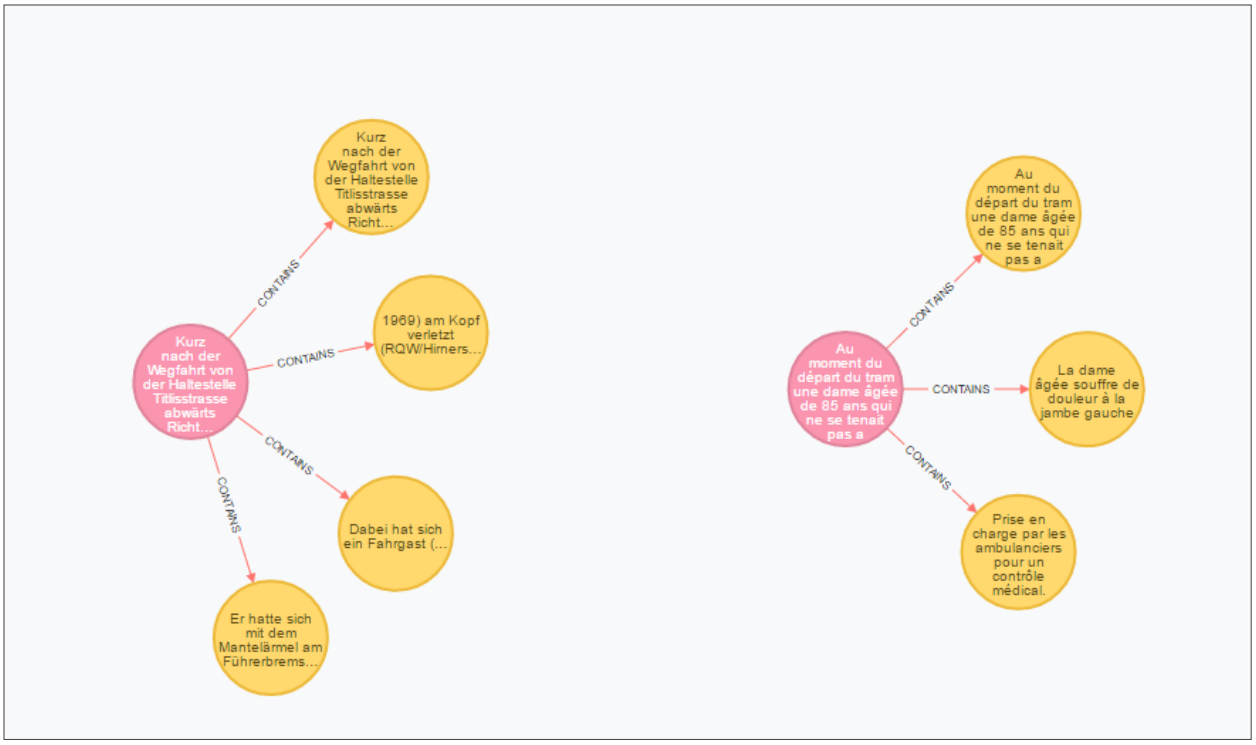


Figure 3: Two example RECORD nodes with their corresponding SENTENCE nodes

Original text	Original language	English translation
Kurz nach der Wegfahrt von der Haltestelle Titlisstrasse abwärts Richtung Römerhof kommt es zu einem Nothalt Dabei hat sich ein Fahrgast (Jg 1969) am Kopf verletzt RQW/Hirnerschütterung).Der Wagenführer erklärte gegenüber der Polizei, dass er einen Manipulationsfehler gemacht hat Er hatte sich mit dem Mantelärmel am Führerbremsventil verheddert und beim Befreiungsversuch unabsichtlich den Notschalter ausgelöst.	German	Shortly after leaving Titlisstrasse station towards Römerhof there was an emergency stop. A passenger (born 1969) injured his head (RQW / concussion). The driver explained to police that he had a made a mistake. His sleeve had become caught on the brake valve and he had inadvertently activated the emergency brake as he attempted to free it.
Au moment du départ du tram une dame âgée de 85 ans qui ne se tenait pas a chute. La dame âgée souffre de douleur à la jambe gauche. Prise en charge par les ambulanciers pour un contrôle médical.	French	As the tram left an 85-year-old lady who was not holding on fell over. The elder lady had pain in her left leg. Paramedics attended and performed a medical examination.

Table 2: Full text of example records with translation to English

A further 16,419 WORD nodes were created from the sentences; Figure 4 shows an example of WORD nodes in blue associated with the sentences that are shown in Figure 3. The figure shows that WORD nodes are linked with an edge labelled CONTAINS to indicate that the word occurs in the sentence. A word may appear in many sentences and therefore a WORD node may be linked to more than one SENTENCE node.

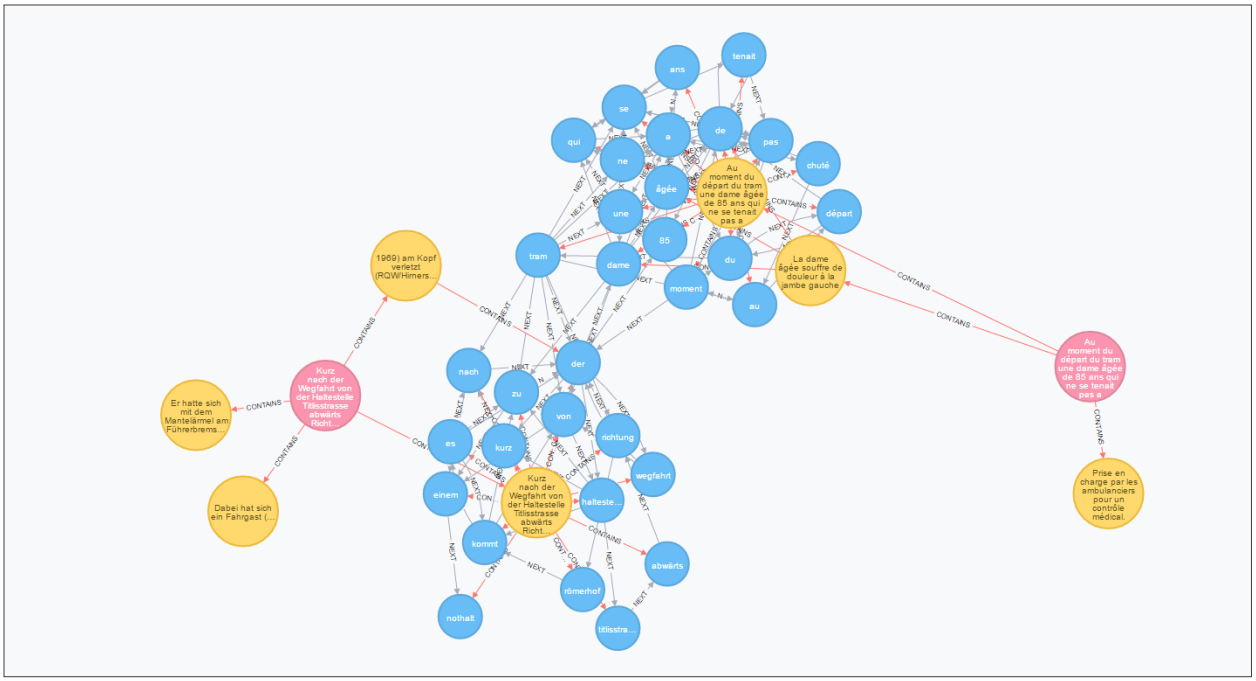
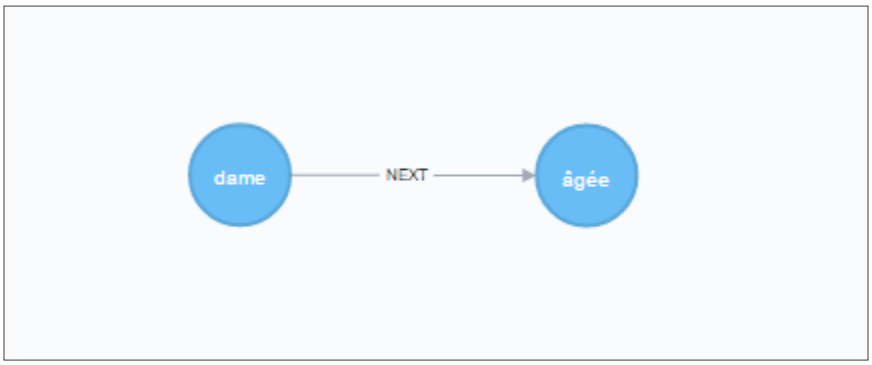


Figure 4 Example WORD nodes linked to two SENTENCE nodes

Figure 5 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 corpus, the word *âgée* occurs 202 times and, as a co-location, *dame âgée* occurs 150 times.



WORD *dame*	NEXT	WORD *âgée*
frequency: 620	co-location occurrence: 150	frequency: 202

Figure 5: Example of the co-location *dame âgée* and the NEXT edge linking the words

4.2 Step 2: apply text analysis techniques to identify key terms

The TFIDF process for single words and the modified TFIDF process for bigrams identified 389 terms from the corpus that appeared relevant to the queries listed in Section 1. Table 3 shows a sample of the identified terms.

Identified term	Language	Equivalent term in English
ältere dame	German	elderly lady
bus	German, French	bus
contrôle médical	French	medical examination
dame	German, French	lady
dame âgée	French	elderly lady
fahrgast	German	passenger
frau	German	woman
stürzte	German	fell
une ambulance	French	an ambulance
zug	German	train

Table 3: Example terms identified

4.3 Step 3: establish an ontology of terms

The two-level ontology identified seven core concepts within the text that appears to be related to the queries. A further 46 concepts were identified that appeared to be subordinate to these core concepts. Table 4 shows the list of core concepts and subordinate concepts that were identified and linked in the ontology.

Core concept	Subordinate concepts
actions	hit, medical, injure, get out, fall, enter, rush
body parts	foot, head
direction	direction, in between, in front, backwards
object	bag, alcohol, drugs, stairs, footboard, customer information system, ticket, door
person	doctor, self, customer, person, driver, passenger, months old, years old, baby, young, old, female, male
places	line, station, pavement, hospital, ground, platform
vehicle	carriage, vehicle, ambulance, tram, train, bus

Table 4 Concepts identified from the text

Figure 6 shows an extract from the ONTOLOGY node in the database for the concept *train* and the edges that link this concept within the ontology to the terms that are used to describe the concept.

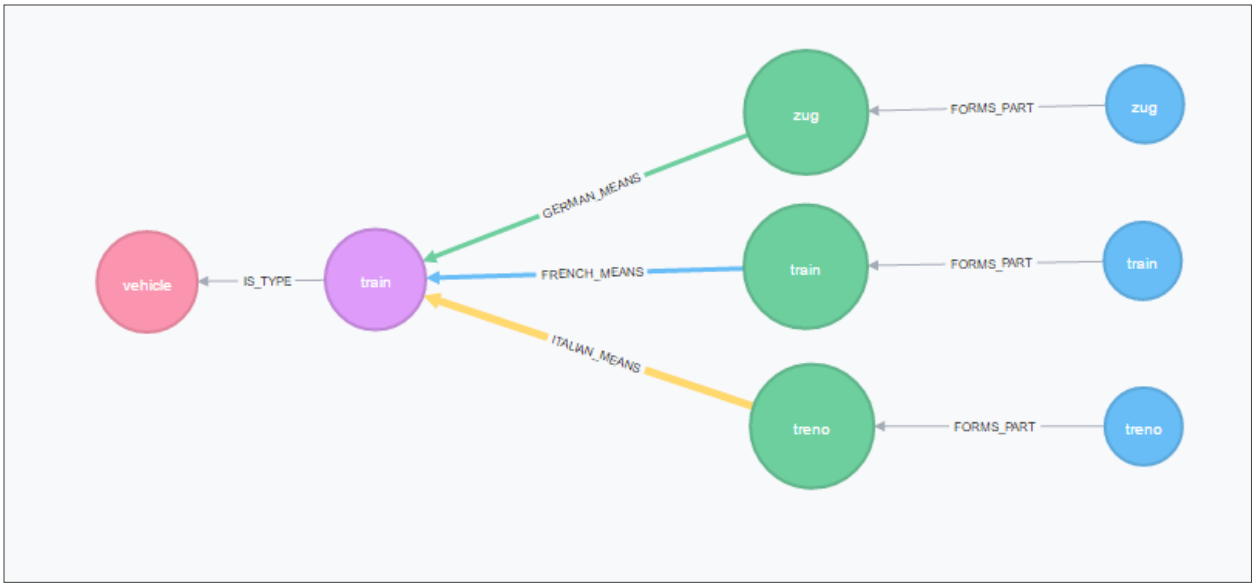


Figure 6: Example ontology relationships for terms that refer to the concept of train

4.4 Step 4: perform queries

Using the approach described in Section 3.4, queries were performed to identify records that related to injuries occurring as a result of passengers:

query 1: alighting vehicles;
query 2: falling down stairs;
query 3: boarding vehicles;
query 4: being trapped by closing doors;
query 5: being struck by falling bags.

For example, to identify records relating to query 2, a query was submitted to the database that:

1. identified the ONTOLOGY nodes describing the action *falling*, and the place *stairs*;
2. identified all TERM nodes linked to these ONTOLOGY nodes, regardless of the type of relationship (indicating agnosticism to whether the term is German, French, or Italian);
3. identified all WORD nodes linked to these TERM nodes; and
4. in turn, identified all SENTENCE nodes linked to these WORD nodes and then all linked RECORD nodes.

To reduce false positive results, a second part of the query then tested each RECORD node and rejected any node that was not linked to *both* the ONTOLOGY nodes relating to *falling* and to *stairs*.

Table 5 shows the number of records that were returned for each query by the language the record was written in.

Query	Language		
	German	French	Italian
1. Alighting	693	296	34
2. Falling down stairs	73	9	1
3. Boarding	1100	349	16
4. Closing doors	1220	464	409
5. Falling bags	3	1	0

Table 5: Number of records returned for each query by language

4.5 Step 5: review by fluent speakers of each language

A sample of the records from each query for each language was extracted for review by a fluent speaker. The review determined whether each record appropriately described the item in the query. Table 6 shows the number of records from the sample that correctly described the incident for each query for each language; the numbers in the table show the number of queries that were correctly identified followed by a slash (/) followed by the number of records in the assessed sample.

Query	Language			
	German	French	Italian	Total
1. Alighting	12 / 12 (100.0%)	12 / 12 (100.0%)	12 / 12 (100.0%)	36 / 36 (100.0%)
2. Falling down stairs	12 / 12 (100.0%)	8 / 9 (88.9%)	1 / 1 (100.0%)	21 / 22 (95.5%)
3. Boarding	12 / 12 (100.0%)	12 / 12 (100.0%)	12 / 12 (100.0%)	36 / 36 (100.0%)
4. Closing doors	12 / 12 (100.0%)	12 / 12 (100.0%)	12 / 12 (100.0%)	36 / 36 (100.0%)
5. Falling bags	3 / 3 (100.0%)	0 / 1 (0.0%)	0 / 0 (-)	3 / 4 (75.0%)
Total	51 / 51 (100.0%)	44 / 46 (95.7%)	37 / 37 (100.0%)	132 / 134 (98.5%)

Table 6: Number of correct records / number of records in sample for each query

Overall these results show 98.5% accuracy in detecting records that correctly correspond to the safety-related incidents described by each queries. The performance is uneven across the queries: the sample records were completely correct for Query 1 (alighting), Query 3 (boarding) and Query 4 (closing doors). However, accuracy was as low as 75% in correctly identifying records belonging to Query 5 (falling bags). Examination of the incorrectly identified record showed that each term in the query was present in the record (such as falling and stairs both being present) but the two not being linked: in this case a fall occurred and then some time later, stairs are mentioned in the report but they are not involved in the fall.

5.0 Discussion

This work involved a number of steps being integrated as part of a larger process. Whilst literature are available describing prior research for each of the steps, we have not found an example of the full process being completed, as such the work described in this paper is the first known work of its kind for safety incident data. As an exploratory

process, the overall performance of 98.5% is a robust result that suggests the method used has potential to be applied as a reliable tool for extracting information from multilingual text sources. This result is especially encouraging since the analyst undertaking the work did not have any knowledge of the languages the source text was written in. The inconsistent results across queries show that further research in this area should be undertaken.

To date, the analysis has been undertaken using only a thematic or *bag of words* approach to identify records that correspond to each query: *i.e.* records are identified if they contain key terms, regardless of how the terms occur in the source text. For example, the terms *passenger*, *struck*, and *door* all occurring within a sentence may indicate that a passenger was struck by a closing door, or may describe an angry passenger who struck a door that was already closed. The *bag of words* approach is also vulnerable to incorrectly resolving homonyms. For example, in English the word *train* is both a noun (the vehicle) and a verb (to teach). The approach used in this study does not give consideration to the meaning of words and may mistake the statement "*we train all staff in injury management*" to indicate that a member of staff was injured on a train. An improvement to the method would be to consider *semantic* text analysis approaches, for example text analysis that considers not only whether key terms occur in a records, but the relationship between the key terms.

The size of the source data, only 5065 records, may be a limiting factor in this research. A larger corpus of records would allow more terms to be identified during the ontology establishment (Section 3.3). An expanded ontology would, in turn, be useful to perform more precise queries. Another limitation of this study is that the review by fluent speakers (see Section 3.5) was able to identify only positive matches and false positives. With large numbers of source records it is infeasible to identify false negative results (*i.e.* records that should have been categorised as corresponding to one of the queries, but were not identified as such by the query). A fuller application of the process might be to attempt to identify categories to which every record in the source data could be assigned. With such an implementation, every record in the source data set would be categorised to at least one category. The sampling process would therefore provide a fuller picture of the overall accuracy by identifying either: records that were categorised to an incorrect category; or records that were not categorised to any category.

A limitation of this study is the relatively small size of the data set used for validation which resulted from the difficulty in obtaining access to staff who had sufficient knowledge of both railway operations and the languages in which the records were written. It would be possible, in principle, to use automated translation tools (such as the online tool used in this study) to convert all records to a common language (in this study English was used) to allow validation to occur. Such an approach would, however, require a confidence in the accuracy of the translation tool. Confidence could be obtained by testing the tool's translation on a large number of sample records to ensure that the translator does not change the essential meaning of the record. Such an approach would, again, require access to staff with suitable knowledge for a considerable time. Such a commitment in effort may be worthwhile in cases where there are a large number of records to be translated and sampling is required to ensure the accuracy of the categorisation. Another approach to obtain confidence in a translation tool would be to train a tool specifically for the types of records that are submitted. Again, such an approach would require access to suitably qualified staff during the training period.

An important consideration in this study is that the ontology was driven by the text itself: only terms that appeared to be important from application of the TFIDF queries were included in the ontology. In theory, the bottom-up ontology could be complemented be a *top-down* approach, whereby an analyst with suitable knowledge of railway safety would create an ontology of abstract concepts and work down to specific terms that may be sought in the source text.

The work undertaken in this study created a rudimentary dictionary of railway-safety-specific terms in the three languages used for reporting events. The work also created a simple ontology of entities and their taxonomical relationships. These tools can provide knowledge that could be used in subsequent studies. For example, the ontology of railway entities could be used for classification of records in another language, since the nature of the railway and the inter-relationships of entities remains the same regardless of the language that is used to describe them. Similarly, it is possible that further work may be able draw on dictionaries or ontologies previously created by others. Multi-lingual dictionaries are commonplace and we expect that we could readily acquire a dictionary that would reduce the work required to identify railway terms. In so doing there would be a choice as to whether the entire dictionary would be imported into the database, or whether only railway-safety-specific terms were included. Care would have to be taken to ensure that railway-specific usage was correctly included in the database so that nuance is not lost; for example on the railway, the English term *frog* describes a particular part of the rail although this may not be described in a common dictionary. Finally, projects such as ConceptNet (2019) seek to develop general purpose, open-source ontologies that can be used to understand the relationships between real-world entities. Again, where sufficient detail has been added by people with appropriate knowledge of the domain (in this case railway safety), these tools may provide a means to expedite similar work in the future.

6.0 Conclusion and further work

Overall the technique described in this paper provides a robust technique to identify safety-related information from multi-lingual text sources. The overall accuracy of 98.5% suggests the approach might be applied as a reliable tool. Having proven the potential of the work with a sample of 5065 records, we believe the work would benefit from a larger and more varied corpus of records from which we hope to be able to extract reliable results for all of the queries presented in Section 1. A further study may also seek to categorise all records in the database based only on the themes being presented in the records, *i.e.* categorisation without having pre-defined queries. Such a bottom-up exploration of the text would allow the accuracy of the categorisation to be more fully tested. Furthermore, a bottom-up approach would allow new knowledge to be obtained from the incident reports; themes in safety incidents would be identified regardless of whether there is a pre-existing query designed to identify the theme.

Many organisations have large sources of textual data, for example hazard reports, accident investigation reports or safety audit reports. Being able to extract information from these data sources can provide a powerful means to improve safety. Furthermore many jurisdictions around the world are multi-lingual; within such jurisdictions an inability to bring together safety learnings from text in different languages would present a serious impediment to continuous safety improvements. We present this work as a potential method to allow various text-based data sources to be integrated in a way that supports organisations' overall safety management programmes.

7.0 Acknowledgements

This work was funded through the strategic partnership with the Rail Safety and Standards Board (RSSB). We are grateful to the Swiss Federal Office of Transport (FOT) for making available the source data used in this study.

8.0 References

Alvarado, A. B. R., Arevalo, I. L., & Leal, E. T. (2016). The acquisition of axioms for ontology learning using named entities. *IEEE Latin America Transactions*, *14*(5), 2498-2503.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., ... & Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2), 79-85.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., & Roossin, P. (1988, August). A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics-Volume 1* (pp. 71-76). Association for Computational Linguistics.

Chase, H. S., Mitrani, L. R., Lu, G. G., & Fulgieri, D. J. (2017). Early recognition of multiple sclerosis using natural language processing of the electronic health record. *BMC medical informatics and decision making*, *17*(1), 24.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

Colace, F., De Santo, M., Greco, L., Amato, F., Moscato, V., & Picariello, A. (2014). Terminological ontology learning and population using latent dirichlet allocation. *Journal of Visual Languages & Computing*, *25*(6), 818-826.

ConceptNet. (2019). ConceptNet, an open, multilingual knowledge graph. Retrieved from http://conceptnet.io/; 16 January 2019.

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

Dale, R., Moisl, H., & Somers, H. (Eds.). (2000). *Handbook of natural language processing*. CRC Press.

Donaldson, L. J., Panesar, S. S., & Darzi, A. (2014). Patient-safety-related hospital deaths in England: thematic analysis of incidents reported to a national database, 2010–2012. *PLoS medicine*, *11*(6), e1001667.

Rashidy, R. A. H. E., Hughes, P., Figueres-Esteban, M., Harrison, C., & Van Gulijk, C. (2017). A Big Data modeling approach with graph databases for SPAD risk. *Safety Science*.

Figueres-Esteban, M., Hughes, P., & Van Gulijk, C. (2016). Visual analytics for text-based railway incident reports. *Safety science*, *89*, 72-76.

Hoxha, J., Jiang, G., & Weng, C. (2016). Automated learning of domain taxonomies from text using background knowledge. *Journal of biomedical informatics*, *63*, 295-306.

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

Hughes, P., Shipp, D., Figueres-Esteban, M., & van Gulijk, C. (2018). From free-text to structured safety management: Introduction of a semi-automated classification method of railway hazard reports to elements on a bow-tie diagram. *Safety Science*.

Hutchins, W. J. (1978). Machine translation and machine-aided translation. *Journal of Documentation*, *34*(2), 119-159.

IJntema, W., Sangers, J., Hogenboom, F., & Frasincar, F. (2012). A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics: Science, Services and Agents on the World Wide Web*, *15*, 37-50.

Ittoo, A., Nguyen, L. M., & van den Bosch, A. (2016). Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry*, *78*, 96-107.

Kang, Y. B., Haghighi, P. D., & Burstein, F. (2014). CFinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications*, *41*(9), 4494-4504.

Kayser, D., & Nouioua, F. (2009). From the textual description of an accident to its causes. *Artificial Intelligence*, *173*(12-13), 1154-1193.

Koehn, P. (2005, September). Europarl: A parallel corpus for statistical machine translation. In *MT summit* (Vol. 5, pp. 79-86).

Liu, K., Hogan, W. R., & Crowley, R. S. (2011). Natural language processing methods and systems for biomedical ontology learning. *Journal of biomedical informatics*, *44*(1), 163-179.

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.

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic methods in accident research*, *11*, 1-16.

Och, F. J. (2003, July). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 160-167). Association for Computational Linguistics.

Patel, A. A., & Dharwa, J. (2017). Graph Data: The Next Frontier in Big Data Modeling for Various Domains. *Indian Journal of Science and Technology*, *8*(1).

Popping, R. (2000). *Computer-assisted text analysis*. Sage.

Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, *16*(3), 197-214.

René Descartes to Marin Mersenne, Tuesday, 20 November 1629 [descreCU0030010a1c]. (n.d.). Electronic Enlightenment Document Collection. doi:10.13051/ee:doc/descrecu0030010a1c

Ruiz-Martínez, J. M., Valencia-García, R., Fernández-Breis, J. T., García-Sánchez, F., & Martínez-Béjar, R. (2011). Ontology learning from biomedical natural language documents using UMLS. *Expert Systems with Applications*, *38*(10), 12365-12378.

Sadoddin, R., & Driollet, O. (2016, March). Mining and Visualizing Associations of Concepts on a Large-Scale Unstructured Data. In *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on* (pp. 216-224). IEEE.

Saeeda, L. (2017, May). Iterative Approach for Information Extraction and Ontology Learning from Textual Aviation Safety Reports. In *European Semantic Web Conference* (pp. 236-245). Springer, Cham.

Sánchez Ruenes, D. (2007). Domain Ontology learning from the Web.

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., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., et al. (2005). Relations in biomedical ontologies. Genome Biology, 6(5), R46.

Taler, V., Johns, B. T., Young, K., Sheppard, C., & Jones, M. N. (2013). A computational analysis of semantic structure in bilingual verbal fluency performance. *Journal of Memory and Language*, *69*(4), 607-618.

Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., & Raynal, C. (2016). Natural language processing for aviation safety reports: from classification to interactive analysis. *Computers in Industry*, *78*, 80-95.

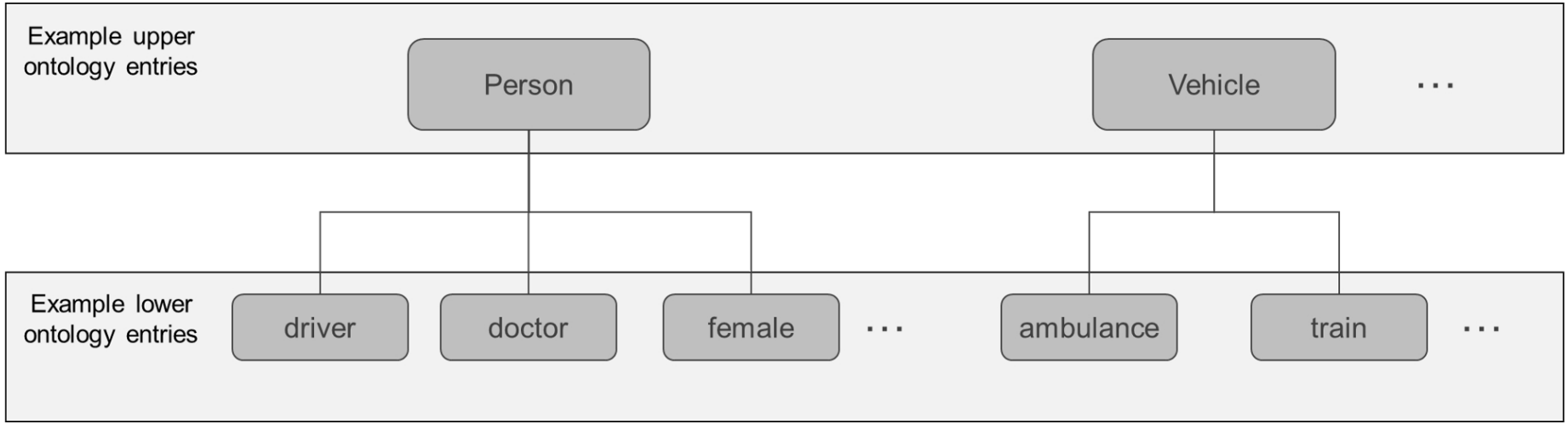
Toyabe, S. I. (2012). Detecting inpatient falls by using natural language processing of electronic medical records. *BMC health services research*, *12*(1), 448.

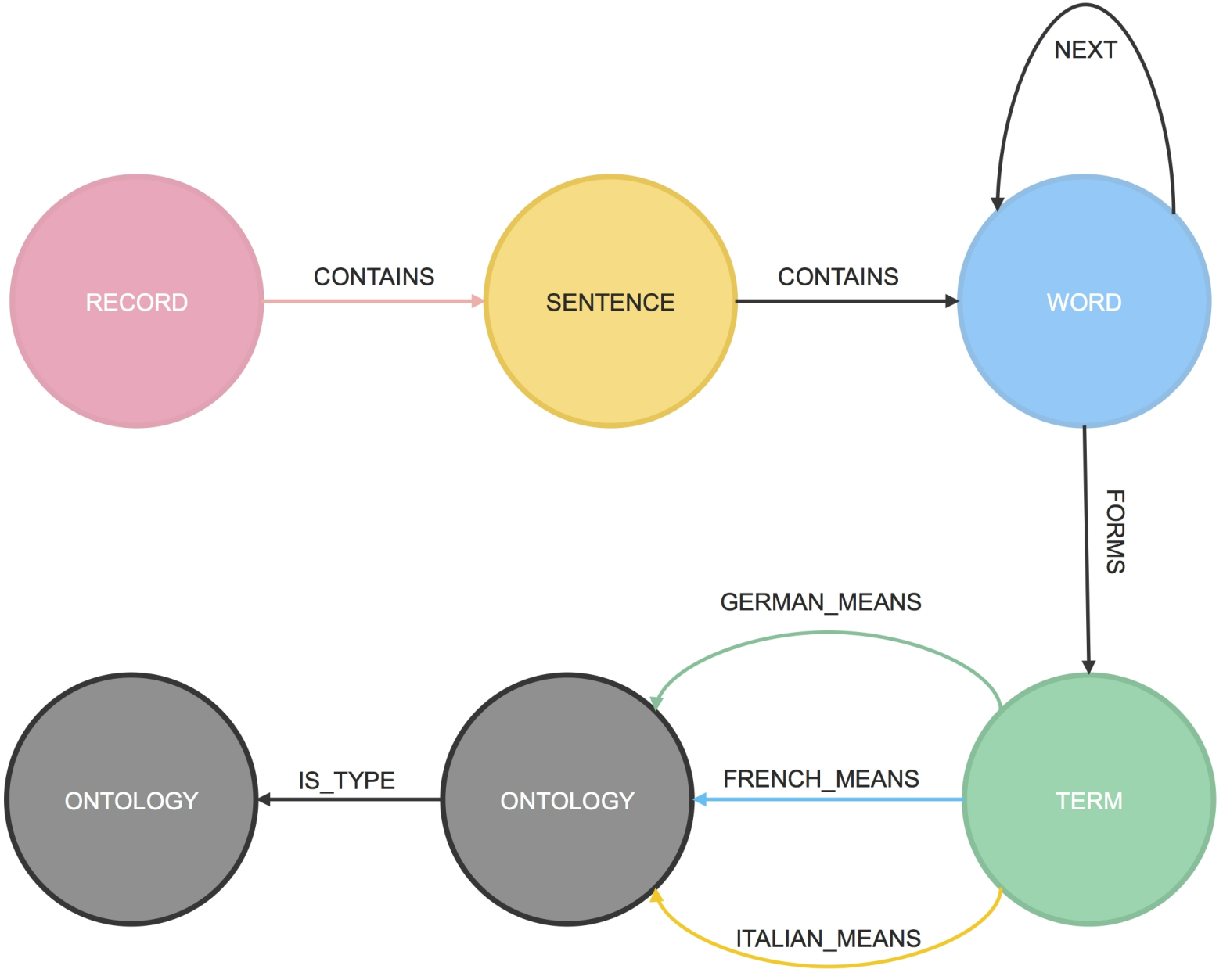
Van Gulijk, C., Hughes, P., & Figueres-Esteban, M. (2016). The potential of ontology for safety and risk analysis. In *Proceedings of ESREL 2016*. CRC Press. Chicago

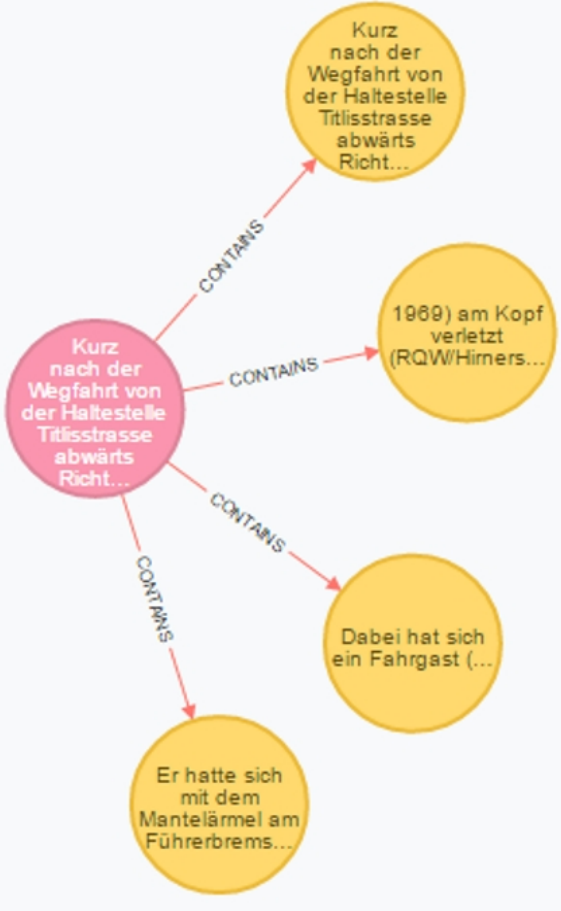
Weaver, W. (1955). Translation. *Machine translation of languages*, *14*, 15-23.

Wu, J., & Heydecker, B. G. (1998). Natural language understanding in road accident data analysis. *Advances in Engineering Software*, *29*(7), 599-610.

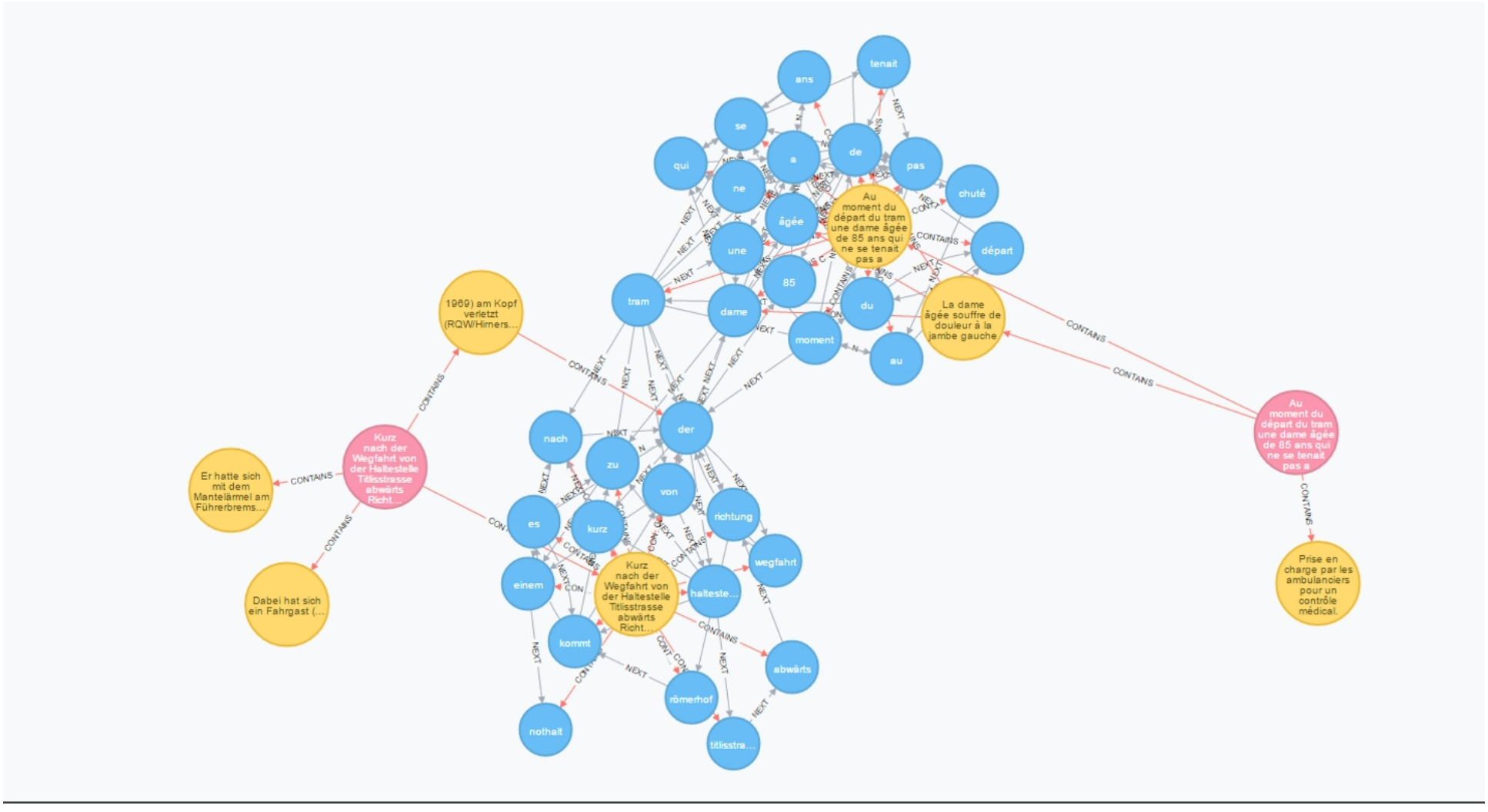
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

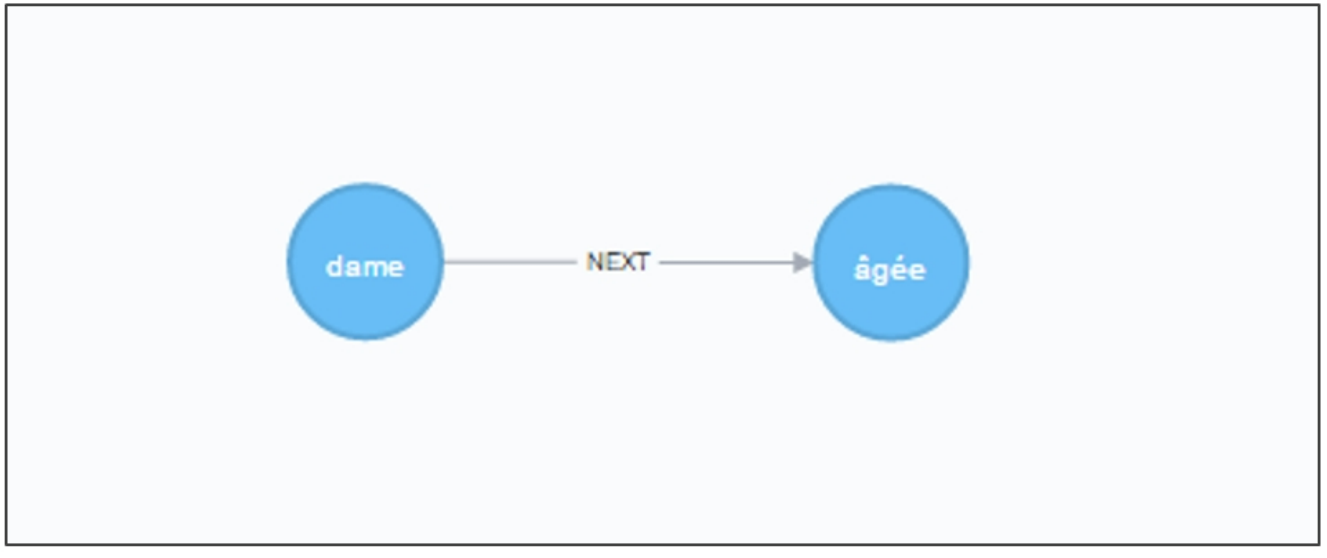

Example upper ontology entries
Person
Vehicle
...
Example lower ontology entries
driver
doctor
female
...
ambulance
train
...


NEXT
RECORD
CONTAINS
SENTENCE
CONTAINS
WORD
FORMS
GERMAN_MEANS
ONTOLOGY
IS_TYPE
ONTOLOGY
FRENCH_MEANS
TERM
ITALIAN_MEANS

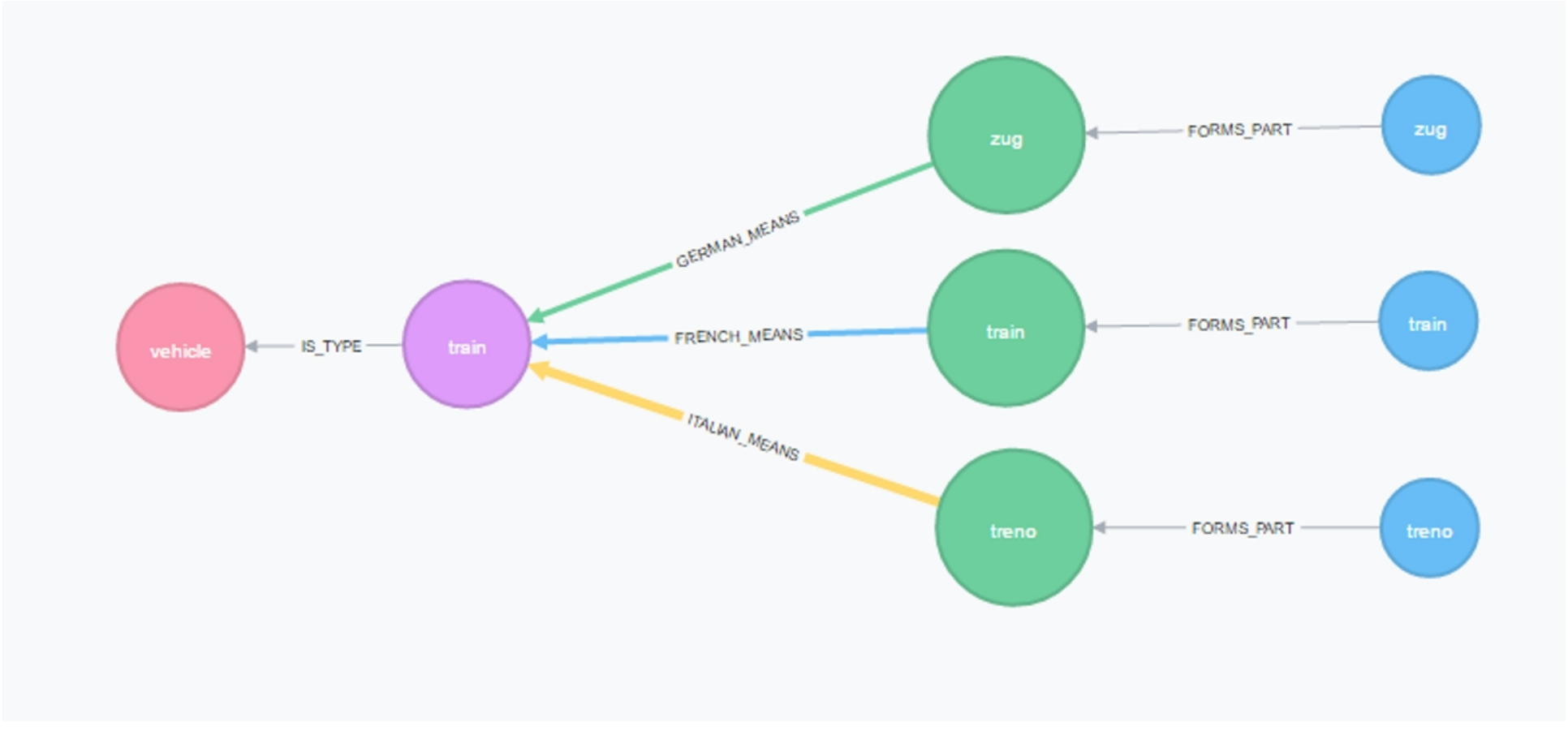

Kurz nach der Wegfahrt von der Haltestelle Titlisstrasse abwärts Richt...
CONTAINS
1969) am Kopf verletzt (RQW/Hirners...
CONTAINS
Kurz nach der Wegfahrt von der Haltestelle Titlisstrasse abwärts Richt...
CONTAINS
Dabei hat sich ein Fahrgast (...
CONTAINS
Er hatte sich mit dem Mantelärmel am Führerbrems...


Au moment du départ du tram une dame âgée de 85 ans qui ne se tenait pas a
CONTAINS
La dame âgée souffre de douleur à la jambe gauche
CONTAINS
Au moment du départ du tram une dame âgée de 85 ans qui ne se tenait pas a
CONTAINS
Prise en charge par les ambulanciers pour un contrôle médical.


Au moment du départ du tram une dame âgée de 85 ans qui ne se tenait pas a
La dame âgée souffre de douleur à la jambe gauche
Kurz nach der Wegfahrt von der Haltestelle Titlisstrasse abwärts Richt...
Kurz nach der Wegfahrt von der Haltestelle Titlisstrasse abwärts Richt...
1969) am Kopf verletzt (RQW/Hirners...
Er hatte sich mit dem Mantelärmel am Führerbrems...
Dabei hat sich ein Fahrgast (...
Prise en charge par les ambulanciers pour un contrôle médical.
tenait
ans
se
qui
a
de
pas
chuté
ne
âgée
une
85
départ
tram
dame
du
moment
au
der
nach
zu
von
es
kurz
richtung
wegfahrt
einem
halteste...
kommt
abwärts
römerhof
titlisstra...
nothalt
CONTAINS
NEXT



WORD *dame*	NEXT	WORD *âgée*
frequency: 620	colocation occurrence: 150	frequency: 202


zug
FORMS_PART
zug
GERMAN_MEANS
vehicle
IS_TYPE
train
FRENCH_MEANS
train
FORMS_PART
train
ITALIAN_MEANS
treno
FORMS_PART
treno