NWI-IMK013

Keeping up with expressing the sense of structured natural language semantics

Master thesis Information Sciences

Author:
T. (Thomas) Nobel

Student number:
S1013787

Supervisor
dr. S.J.B.A. (Stijn) Hoppenbrouwers

Assessor
dr. ir. E. (Eelco) Herder

Sunday 10th November, 2019

Abstract

Domain-specific languages are not a static entity. On the contrary, they are continuously changing and adapting to their respective environments. Cybernetics can be used to explain why domain-specific language evolution exists. We observe that objects, entities, and terms in different applications such as Fact-Based Modelling (FBM) and lexical databases (such as Wordnet) are fundamentally (logically) equivalent. To assert this claim, we will use a diverse set of theories including first- and second-order cybernetics, Wittgenstein, and Kripke semantics to argue why we believe that in essence the challenges of defining domain-specific languages are strongly related with the challenges of formal information modelling. In this thesis, we aim for soundness and reliability while defining the domain-specific language. The goal of combining seemingly different theories is to propose a new framework and basis for further (applied) research in correctly defining, mapping, and comparing domain-specific languages. Our explanation can be used to improve the newly developed tool called A-Lex, which aims to make building domain-specific lexicons and business data models more approachable for non-IT experts by the use of self-service modelling. A-Lex gives private persons, academics, public authorities and organisations the opportunity to make public collections which can be reused, shared, and imported freely by others (a form of crowdsourcing).

You shall know a word by the company it keeps – John Rupert Firth

Acknowledgements

A master thesis, although written by one person, is never without the support and help of others. Therefore, I would like to give some special thanks to the following persons.

- My supervisor dr. Stijn Hoppenbrouwers for his support, advice, guidance, and not to forget his contagious enthusiasm. I would never have been able to write and complete this thesis without his support.
- My assessor dr. ir. Eelco Herder for his quick and enthusiastic response of becoming a second reader.
- dr. Patrick van Bommel for his lectures on (foundations of) information modelling and his advice concerning the use of the work of John Sowa.
- prof. dr. ir. Arjen de Vries and dr. ir. Faegheh Hasibi for their lectures on information retrieval, advice on using 2dsearch, and advice on useful literature.
- My philosophical housemate Menze van Veen for our talks on language, logic and philosophy.
- My fellow student Floris Roelofs for his grammar check and feedback.

And of course, I also like to thank my family, friends, and not to forget my girlfriend for their moral support during my master's program and especially during the thesis process.

My thesis supervisor gave me the unique opportunity as a master student to help him write an article. We have submitted the article together and has subsequently been presented at the FBM 2019 conference in Rhodes, Greece (see also (Nobel et al., 2020) in the bibliography). In fact, I have been given the honour of becoming the first author of the article. It has been peer-reviewed and accepted into the Lecture Notes in Computer Science (LNCS) series by Springer.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement

Solutions to the information need for domain-specific problems exist in many different forms. Figure 1.1 shows an image of an early 20th-century ship's telegraph. This device is, as strange as it may seem, an information system on its own. Even though this was a mechanical device, the technological implementation such as mechanical, electrical or digital is not relevant concerning the solutions of a communication and/or information need. When the telegraph was designed, a limited set of terms and their meanings regarding the speed and direction were defined, so that both ends of the telegraph exactly knew what it meant when the state of the telegraph was changed and what the consequences were of these changes (S. J. B. A. Hoppenbrouwers, 2003).

A limited well-specified set of concepts can be useful in some cases such as the case with the ship's telegraph[1] and with many other day-to-day applications such as traffic lights, ICMP messages (such as ping, destination unreachable and redirect), bicycle/car gear, etcetera. The ship's telegraph enabled clear and well-formed communication between the bridge and the engine room as was required for technology-related reasons. Running back and forth between the bridge and engine room was not practical and changing the speed/direction in the engine room without the possibility of steering and being able to view the situation around the ship was also not deemed as the best course of action.

[1] often a ship also had a speaking-tube next to the ship's telegraph for non-standard communication with the engine room and acknowledgements

Figure 1.1: A ship's telegraph

(adapted from Lannan Gallery, 2019)

Proper communication is crucial; as the famous organisational theorist Russel Ackoff defined it "We fail more often because we solve the wrong problem than because we get the wrong solution to the right problem" (King, 1993). We can (partially) circumvent the issue stated by Ackoff by letting the client actively participate in building the end-product. Preventing and reducing this problem by participation is what the newly developed program called ABN AMRO Lexicon (A-Lex) of the ABN AMRO bank hopes to achieve. A-Lex gives the user/client the possibility to define their own domain-specific language or business data model. The user defines the semantics and A-Lex provides the syntax. We assume that the domain experts can adequately define and build the lexicon and that they know it better and in more detail than an IT expert. Therefore, they are quite trivially more suitable for the task of defining the semantics (Pleijsant, 2019).

Many companies nowadays do see not only material things as assets but also immaterial things such as data and in a more applied sense 'information' as assets (Khatri & Brown, 2010). Implicitly this also includes modelling the domain-specific language. A shared understanding of knowledge and terms is vital for any modern organisation or institution. Even though this might seem trivial, no single model, ontology or glossary exists which everybody agrees upon in all circumstances/domains (Jayawardana & Jayawardana, 2017). Therefore, we aim to let the users define the semantics of their domain-specific language using crowdsourcing. This will make building information and data models not only more accessible but also help to smooth communication by language standardisation through correctly defined terms.

However, if we assume that the domain experts properly know the semantics of its domain-specific language, then this gives rise to another problem, namely, how do they communicate this towards others who may or may not be knowledgeable of the same domain. We require a way to write down the exact meaning of terms without making it ambiguous or contradictory. Evolution of domain-specific languages can mean that only slight nuances are changed at a time and that therefore, updating the definition of a term might not seem to be much work. However, terms are, in a sense, intermingled. With this, we mean that their exact definitions depend on the exact definitions of other terms because, in order to properly define terms, we require other terms.

Information (system) modelling used to be the domain of highly specialised IT experts and data analysts who were trying to abstract the application-specific domain of their customers/clients into some mathematical information model. With the case of the ship's telegraph, this is a doable assignment. However, understanding both the domain and its specific language (for example, a high-tech company) can be quite complicated. Even for people who work in a specific domain, it can take years before they are able to propose and implement solutions concerning the information need. Understanding of the domain is, however, a vital part of information modelling and domain-specific language alignment. In the context of this thesis, we will view defining domain-specific languages as a part of information modelling.

Thus, we have a twofold problem. The first one being that information modelling cannot reasonably be expected to be done by people apart from a small group of IT specialists. The second one being that apart from the small group of people in a specific domain one cannot reasonably expect that they are aware of every (changing) detail of that domain (including from IT specialists). This holds true for building an arbitrary information system for a domain and thus also for correctly modelling a domain-specific language. In this thesis, we aim to solve both previously stated problems. We will do this by proposing a way for people involved in the same domain to unambiguously and correctly model their own domain-specific lexical elements (terms). If on top of the previously stated problems, we add the business need for business language alignment, then the challenge and necessity grow evident. However, because we need to be precise and aim to propose a solution which can explain, monitor, and define the nuances of domain-specific language evolution, we cannot circumvent (formal) theories concerning cybernetics, crowdsourcing, language, logic, and philosophy.

To achieve these goals, we want to explain why domain-specific languages are not static (as is the case with the ship's telegraph) but instead continuously changing using cybernetics. This also implies that precisely modelling terms of a domain-specific language is a continuous process. In order to use already established formalisations from information modelling, we aim to propose a view that fact-based modelling is, in essence, closely related to defining structured natural language. We will formulate these arguments based on existing theories. Also, we will discuss how the community can be used to define the domain-specific language using crowdsourcing. We will therefore also discuss the semantic web and the web 2.0.

Classical approaches of (statically) defining terms are not sufficient anymore. According to the Oxford English Dictionary, a 'definition' is defined as: "A statement of the exact meaning of a word, especially in a dictionary." (Definition., n.d.). This is, however, a complete opposite explanation of what Wittgenstein claims (see section 2.9.3) and what seems to be the case in practice. How can we define a word exactly when its meaning can significantly differ in the context and domain/environment in which it is used? Even more so, how can this be the case if even exact/correct[2] communication of the meaning of a term is very challenging? A dictionary seems to state a generic and static definition of a word. But a word is just a string of symbols from some alphabet and in a lot of instances the meaning of a word although made from the same sequence of symbols differs in various contexts. We will try to answer these questions in this thesis.

In short, we observe that many applications are using the so-called Subject - Verb - Object (SVO) structure as elementary sentences to define relations. Yoda from Star wars also does this, but he reverts the order and therefore even though logically correct, it may sound strange because it does not conform the perpetuated syntax of structured natural language used by most people (for example English speaking people).

1.2 Goals

This master thesis will aim to define how A-Lex, Wordnet, and FBM (Fact-Based Modelling) are fundamentally related or different from each other. We will do this by applying theories and concepts from linguistics, cybernetics, FBM[3], logic, and philosophy. In addition, we will look at the idea of crowdsourcing and self-service modelling. This is not a completely new concept, but the literature on this particular issue is quite limited in combination with information system modelling and domain-specific language alignment.

[2]Within the scope of this thesis, we will define correctness as something being both sound and reliable.

[3]To avoid name confusion we will use concepts from both FBM and ORM (Object-Role Modelling) but only use the term FBM.

1.3 Research questions

We require research questions whose purpose it is to make the aim of this thesis more precise. Therefore, we have formulated the following three research questions:

1. How does A-Lex relate to fields like lexical databases (Wordnet), the semantic web, and Fact-Based Modelling (FBM)?
2. How can self-service, community-based modelling be related to theory, integrated with the answer of question 1?
3. How can we sketch an integrating formalisation of theory, to bridge the gap between the use of (structured) natural language in A-Lex, FBM, and logic?

In order to achieve this, we will look into several theories including (structured) natural language, formal modelling, cybernetics, Wittgenstein's tractatus logico-philosophicus, and Kripke Semantics.

1.4 Hypothesis

As has been discussed in (Nobel et al., 2020), the levels of description (lexical, relational, and constraint), structure, and formalisation of different approaches with respect to 'dealing with data' can be quite different. With an approach we mean a technology or tool dealing with data, for example, a tool to model a domain-specific language (such as the previously mentioned A-Lex) or a technique used to build a database model (such as the previously mentioned FBM). Even though their purposes/applications might seem to differ substantially as well, we believe that their underlying foundations are not so different. Therefore, in this thesis, we hypothesise that many similarities exist when we abstract from the details and that they are quite similar on a fundamental level. We aim to come up with a theory to help us underpin these similarities.

1.5 Method

In this thesis, we will use literature, argumentation, derivations, and deductions to come up with answers to the research questions. We will not use a data set for qualitative or quantitative research, nor will we perform any form of empirical research. Therefore no extensive method chapter will be written as it would be redundant.

We will not do much practical research besides informally exploring A-Lex and analysing it. The research will most prominently be a literature study, and a fundamentally-oriented comparison study between A-Lex, Semantic web, FBM, and Wordnet extended with crowdsourcing and self-service modelling.

1.6 Outline

In chapter 2, we will discuss a diverse set of theories. We will discuss some concepts and definitions in natural language in section 2.1. In section 2.2, we will discuss the semantic web and its underlying standard RDF. In section 2.3, 2.4, 2.5 we will discuss A-Lex, Wordnet, and FBM respectively. In section 2.6 we will discuss first- and second-order cybernetics and its link to domain-specific language evolution. In section 2.7, we discuss the basics of crowdsourcing and self-service modelling. In section 2.8 we will discuss the different viewpoints of FBM, Wordnet, and A-Lex. Section 2.9 explains the ideas of Wittgenstein from his book the Tractatus Logico-Philosophicus in the context of the discussed theories and concepts of this thesis. Section 2.10 discusses the work of John Sowa regarding conceptual graphs. Section 2.11 discusses the logical system of Kripke semantics. Section 2.12 discusses formal languages, Turing machines, and computability theory.

In chapter 3 we will discuss the results by combining the sections of chapter 2. This is divided into three sections. Section 3.1 tells us about the possibility to connect (elementary) propositions with logical connectives. Second section 3.2 uses the theories of Wittgenstein to combine seemingly different approaches and limitations. Section 3.3 concludes the story regarding domain-specific language evolution and its gluing function in this thesis. Section 3.4 discusses and combines the other sections of chapter 3, adds the power of web 2.0, and concludes the results.

In chapter 4 we will summarise and conclude the findings of chapter 3. We will end this thesis with a discussion and suggestions for future research in chapter 5.

Chapter 2

Theoretical background

In this chapter, we will discuss the relevant concepts, definitions, theories, and techniques used throughout the rest of the thesis.

2.1 Concepts and definitions of natural language

In this section, we will define the relevant concepts and definitions regarding relations in (structured) natural language used throughout the rest of this thesis. Some of the discussed concepts and definitions are rather fundamental but required to be correctly defined and understood in order to bridge the gap between the different disciplines.

2.1.1 Semantic relationships

The semantic relationships between terms form the backbone of developing a language. When modelling a language as a graph $G = (V, E)$, we can define the vertices V as the terms. The edges E are the different semantic relationships between terms. With the help of these terms, it is possible to map or model a (domain-specific) language. Both Wordnet and A-Lex can relatively easily be visualised as a graph. A graph in these instances models the related lexical elements of a language apart from its syntax/grammar. For example, in figure 2.1 an illustration of a graph is made concerning the synonymous (see subsection 2.1.2 and section 2.4) relationships between different terms in Wordnet.

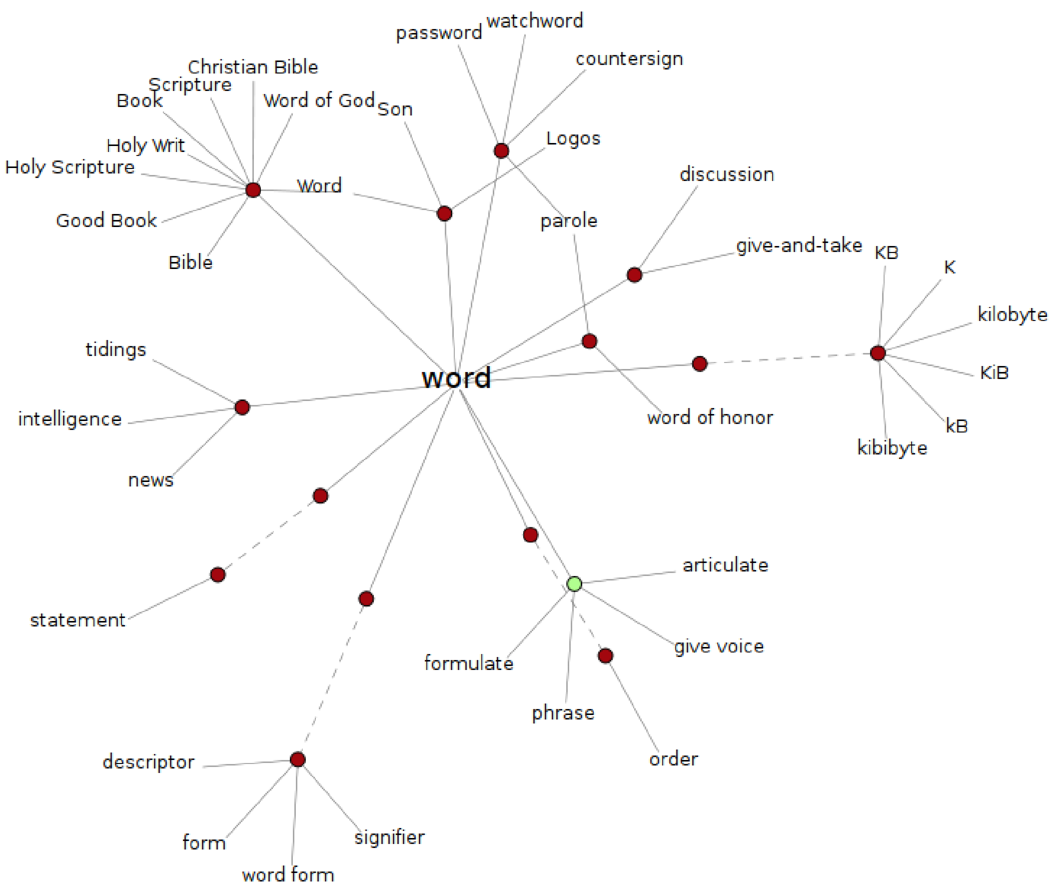

Figure 2.1: Wordnet graph representation

(adapted from Code, n.d.)

This, however, is not quite enough to be able to model the lexicals of an entire (domain-specific) language. We should be able to model all kinds of semantic relationships and, when needed, add new ones. Then, if required by the domain, relation-specific properties such as a rational number denoting the similarity between two terms can be added. We shall first start by explaining the most important and commonly used semantic relations.

2.1.2 Semantic similarity

When using semantic relationships, a similarity graph can be made as has been shown in figure 2.2. This graph can be used to illustrate to what extent different terms are similar to one another. In other words, it denotes the likeness of the meaning of terms compared to one another.

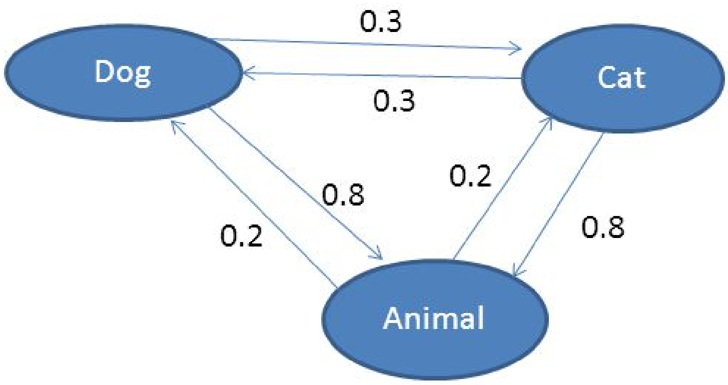

Figure 2.2: Wordnet similarity representation

(adapted from Sutton, 2015)

Many different ways to calculate semantic similarity are in existence. In most cases, nouns are used for similarity calculations because language semantics are largely captured using these. Using a path-based approach is quite common. The resulting score of these distance-based functions are most importantly influenced by (Meng et al., 2013):

1. Path length between concepts (distance between synsets in the graph). This can be calculated using a shortest path algorithm where every synset is viewed as a single node in the graph.

2. Position of the concept in the taxonomy which is related to the depth.

Most other approaches used to calculate semantic similarity are Information Content-based, Feature-based, and Hybrid approaches. Semantic similarity is of immense value for many applications, including artificial intelligence, natural language processing, information extraction, information retrieval, word sense disambiguation, text segmentation, question answering, and recommender systems. Wordnet is used widely in the previously mentioned applications (Meng et al., 2013).

Synonyms

A synonym means that two words mean exactly or approximately the same. According to Leibniz, the substitution of a word with a synonym of that word may not change the truth value of that sentence. However, this implies that the chance that a word has a (or even multiple) synonyms is quite rare. Therefore, another version would be more appropriate by making a synonym relative to a context and thus settle for semantic similarity. This consequently means that if we substitute word A with word B in context C and the truth value does not alter, then we can speak of a synonymous relationship between word A and B. This relationship is self-evidently symmetric because if B is similar to A, then B is equally similar to A because we can substitute B back again for A and still not change the truth value of C (G. A. Miller et al., 1990).

Antonyms

Antonyms are also called opposites. They define the opposite of two words which are linked by an inherently incompatible binary and symmetric relation. For example, the two words 'high' and 'low' are an antonym of each other. It should be noted that it is explicitly about opposite word meanings and not about things that exist in reality. The following three different categories of antonyms exist.

Someone can be a parent and a child at the same time while normally 'parent' and 'child' would be an antonym. In this case, we would define it as a relational antonym because the antonym only makes sense in the context of the relationship between the parent and the child.

Another (second) form of antonyms is gradable antonyms, which means that the two opposite antonyms are on a continuous spectrum, for example, the two words 'heavy' and 'light'.

The final class of antonyms are complementary antonyms. The complementary antonyms are not on a continuous spectrum, but they are still each other's opposite. For example, the two words 'come' and 'go'. Even though they are each other opposite, there is no spectrum between the two on which other words lie.

Homonyms

A term is a homonym if and only if another word exists for which holds that it is pronounced and spelt the same but for which a different meaning exists. For example, the word: 'bat' can mean a nocturnal flying animal or the club used for playing baseball.

In this case, the two words are homographs because both words are written in the same manner. They are however no homophones because this would mean that they do not have a different spelling but the same pronunciation (for example, piece and peace). Homophones cause many real-world spelling errors and homonyms in a general sense, generate many problems in information retrieval, question answering and (automatic) translation between languages. Word-sense disambiguation remains a significant open problem in computational linguistics. Both homographs and homophones are forms of homonyms. We will however, limit ourselves to homographs (Chapter C1 (Jurafsky & Martin, 2008)).

Hyponym and hypernym

A word is a hypernym if and only if another word exists for which holds that the former word is a superset of the latter (which is called the hyponym). For example, a fish is a hypernym for shark (the hyponym in this hypernymy relation), but a shark is a hypernym for the great white shark (which in this hypernymy relation is the hyponym).

The hypernymy relation is used for many different purposes, such as identifying similar words in information retrieval, deriving inferences in question-answering machines and to avoid word repetition in natural language generation systems. When defining hypernymy relations between words, a taxonomy will emerge (vor der Brück, 2010).

Meronymy and holonymy

A meronymy defines a constituent part of a member relationship between two words. It denotes that x is a meronymy of y if and only if x is a member of y or x is a part of y. A meronymy can be seen as a partial order because even though x is a meronymy of y, it is an asymmetric relationship because it does not mean that y is also related to x.

For example, a toe is a meronymy of a foot because a toe is a part of a foot and a student is a meronymy for a class because he or she is a member of a class.

A holonymy is the opposite of a meronymy. It denotes that x is a holonymy of y if and only if ys are parts of x or ys are members of x. Just as with the meronymy, this relation is a partial order.

For example, an apple tree is a holonymy for apple because apples are parts of the apple tree, and a class is a holonymy for students because students are part of a class.

Concept and object relation

The concept and object relation describes how type instances are related to more abstract object types. For example, an ‘apple’ can be a type instance of the object type (or concept/term) ‘fruit’, and ‘Google’ can be regarded as a type instance of the object type ‘search engines’. Multiple type instances together are a set and are related because they are all linked with a concept-object relation to an object type.

Ontology

An ontology has in philosophy the meaning of ‘subject of existence’ and is often confused with epistemology, which means: “knowledge and knowing”. Many different definitions and applications of ontology exist. In this thesis, however, we will use the following definition of an ontology: "A formal and explicit specification of a shared conceptualisation". So, in other words, it is a network or a (conceptual) graph of relations and concepts described formally for one or more agents (Jayawardana & Jayawardana, 2017).

2.1.3 Is-a relationship

The Is-a relationship is useful in many different applications such as knowledge representation (including A-Lex and Wordnet), object-oriented programming, and many others. It is used to denote the type of relations between two objects or types. For example, we can define that A is a subset of B or that A is a generalisation of B. The type of relations depends on the exact definition of the Is-a relationship. In most lexical applications these relations are based on the following semantic relations (which we have all discussed in the previous subsections).

1. Synonyms
2. Homonyms
3. Antonyms
4. Hypernyms and hyponyms
5. Holonyms and meronyms
6. Concept and object relations (type-token)

2.1.4 Structured natural language

Natural languages are as opposed to formal languages, which are deliberately depicted, an undeliberate product of people and time. They slowly evolve and are not a direct result of human design. Traditionally the description of the structure of language was divided into three parts, namely: syntax, semantics, and pragmatics (De Beule & Stadler, 2014). A formal language is a language which has strict syntax rules (it can be mathematically described) and is designed by humans for a specific application such as programming, mathematics or chemistry.

Natural languages have some properties which make them interesting yet in some cases problematic. These properties are the core dealing of A-Lex because often these properties depend on the context/domain in which the language is used. Natural language is full of ambiguities partly because it is filled with idioms and metaphors. In addition, it is also often filled with redundancies, which makes the language verbose, less concise, and less dense compared with formal languages. These ambiguities arise not because of unclear syntax (a proper natural language has just as a formal language, a clear set of syntax rules, and should be well-formed) but arise due to the semantical and pragmatical part of natural languages.

If we look at a mathematical statement such as $x + y = z$ then even though the input of the formula may differ, the semantics of the statement remains the same, the value of z is the sum of x and y[1] (B. Miller et al., 2014). If we look at an equivalent statement in natural language, then the statement "x plus y is equal to z" would be an option. This is however, quite ambiguous for it could imply that xy is equal to z (using concatenation). However, it could also mean that x added to y is equal to z (using addition). The underlying problem is that it is not clear what in our context the 'plus' operation exactly means[2].

Semiotic ladder

Natural languages have, as opposed to formal languages (such as first-order predicate logic), a complex and irregular grammar. Both use operators and connectors, but in a very relaxed manner, this means that there are multiple variants of connecting parts of a sentence. Natural languages are more expressive than formal languages but also far more complex. Transferring natural languages into formal languages is a difficult concept. For example, dealing with irony in a sentence or homonyms (words which are spelt and pronounced the same, but have different meanings).

The field of semiotics is dedicated to the study of signs and is applicable/used in numerous other fields of science. It is about the process of creating meaning. In recent years this field of study has gained more attention due to the growth of multimedia and information retrieval. A sign (which can be a word, image, multimedia, etcetera) identified by its constituent parts is called the signifier. The signified part is the meaning conveyed by the signifier (Hébert, 2018).

The semiotic ladder, as can be seen in figure 2.3, aims to define the relationships between information, communication, and meaning in natural language (Ferreira et al., 2007).

[1]This is not entirely true. We must first have formally defined the semantics of '+' and '='. We will talk about this later using Wittgenstein's theory. For now, it is safe to assume that this is true.

[2]In programming, operation overloading is often used and is regarded as a form of 'syntactic sugar'. This requires rewriting the semantics of operators.

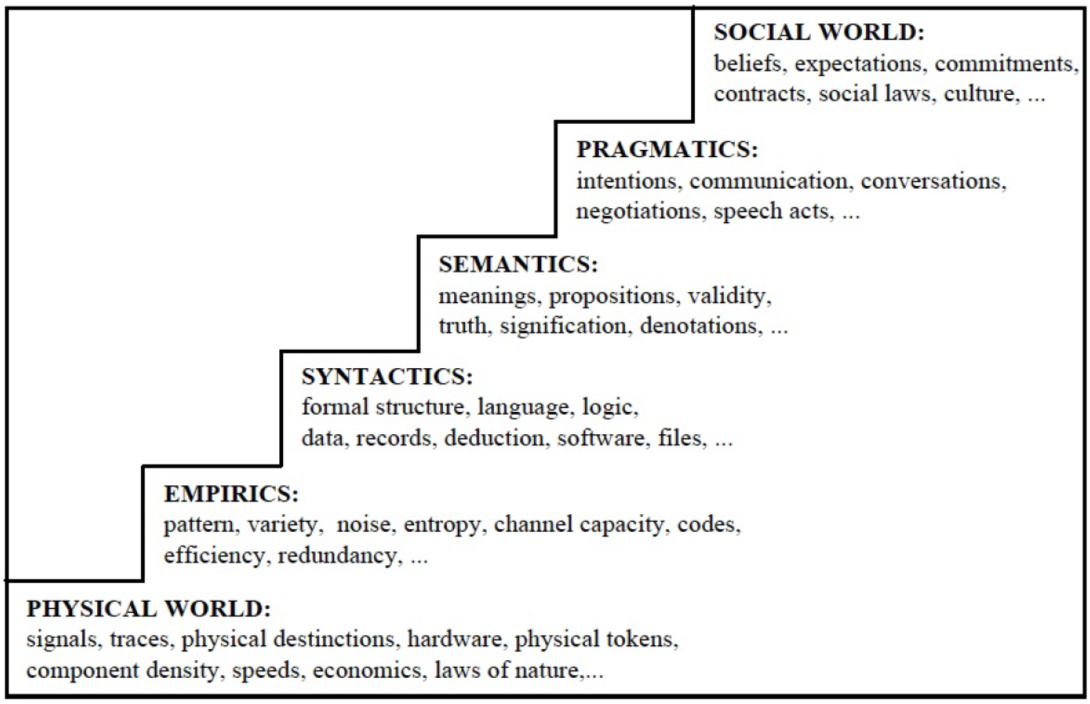

Figure 2.3: The semiotic ladder

(adapted from S. Hoppenbrouwers, 2018 p. 9)

The FRISCO approach

The FRISCO (FRamework of Information System COncepts) tetrahedron aims to close the gap between reality and the modelling concepts based on the principles of semiotics. In figure 2.4 we can see an image of the FRISCO tetrahedron. The theoretical concepts it combines are meant to close the gap between syntax, semantics, and pragmatics in understanding language, (meta-)communication, and information in organisations (Hesse & A. Verrijn-Stuart, 2000) (S. J. B. A. Hoppenbrouwers, 2003). It posits the link between the conception, representation, domain, and actor, with the actor being both interpreter and representer. The 'actor' aspect is crucial; it states the subjectiveness of the language used for communicating about the UoD (Universe of Discourse); its origin and existence in the minds of (groups of) individual humans. It implies the viewpoint that social agents involved in communication are also active concerning the relationship between information and actions. This line of thought implies the existence of 'Environments of Discourse' (EoD) in which agents, artefacts, vocabularies, media, and various other entities impact and perform (meta-)communication (S. J. B. A. Hoppenbrouwers, 2003). Every EoD includes what we can refer to as the 'Community of Discourse' which is basically the UoD/EoD language community. When two groups communicate, a new EoD emerges with new conventions and agreements, its UoD.

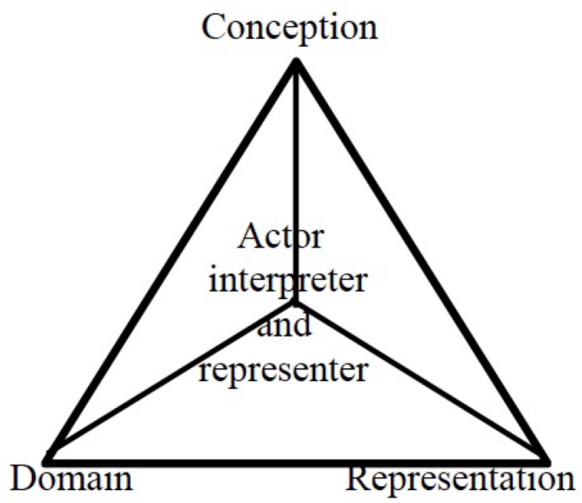

Figure 2.4: The FRISCO tetrahedron

(adapted from Insfran, 2003)

The CoD usually engages in linguistic meta-communication (talking about and therefore verifying and fine-tuning 'what people mean with what words') in an ad-hoc fashion, almost implicitly as part of general conversation. However, in some cases (data management as a primary example), it pays off to make such meta-communication explicit and even to organise it (S. J. B. A. Hoppenbrouwers, 2003). Tools like A-Lex help to do this.

2.2 Semantic web

The semantic web aims for standardisation of web data in such a way that finding, sharing, and reusage of data become easier. With the use of machine-readable and human-readable information markup languages such as XML (Extensible Markup Language), it provides an easy way for the exchange of web data. By representing web page data in a machine-readable way means that machines can search, aggregate, and combine data without human interaction. The World Wide Web Consortium (W3C) has created the semantic web as an extension to the World Wide Web of Documents. It is known as the 'Web of Linked Data' to help create and store data on the web and create rules for handling data. It uses formats and techniques like RDF, SPARQL, OWL, and SKOS. We focus on the RDF part of the Semantic Web here, since we are concerned with the basic word meaning component of the Semantic Web and not with the extensive use of hierarchical relations and axioms/rules that constitute, for example, complex ontologies, 'OWL Style' (Nobel et al., 2020) (Jambhulkar & Karale, 2016).

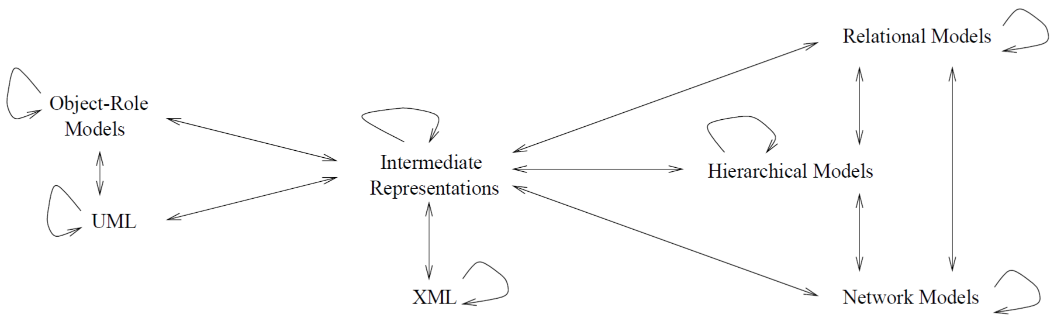

Figure 2.5: Transformation of models

(adapted from van Bommel, 2019 p. 7)

2.2.1 RDF

The Resource Description Framework (RDF) is an application of XML and is meant as an infrastructure to encode, exchange, and reuse metadata in a structured way. It provides the means for representing the data in both a human and a machine-readable manner. The structural constraints of RDF ensure consistent encoding, standardisation, and the interchangeability of metadata (E. Miller, 1998).

The World Wide Web Consortium has guided the development of RDF. The creation of RDF was a collaborative design effort. Its initial intent was a metadata model, but eventually, it evolved as a format for the exchange of data (E. Miller, 1998).

The World Wide Web contains a massive amount of information, and in order to effectively use the metadata conventions regarding the syntax, semantics, and structure are required. RDF enables the creation and use of metadata elements (see figure 2.6), including definitions. Humans are generally quite good at extracting semantic meaning from different syntactic constructs, whereas machines are quite inept for this task. RDF uses a triadic model of resources, property types, and corresponding values to associate resources with specific properties. For example, a resource could be "sprint race" where the corresponding property type would be "won". To give meaning to the property type, the corresponding value could be the atomic value “John Johnson”. Thus we will get the human-readable sentence: “the sprint race is won by John Johnson”. Note that actual definitions (or ‘descriptions’, in RDF and Dublin Core terminology) are created utilising items identified and accessed through URIs (Uniform Resource Identifiers) (Nobel et al., 2020) (E. Miller, 1998).

RDF, in combination with URIs, thus provides the means to publish both machine and human processable vocabularies. In this context, vocabularies are sets of properties or a group of metadata elements defined by communities. This form of domain-specific semantics gives rise to semantic modularity because it creates an infrastructure which enables distributed attribute registries which may be reused, extended, and refined to fit domain-specific requirements (Nobel et al., 2020) (E. Miller, 1998).

RDF also provides them with the means to publish both machine and human processable vocabularies. In this context, vocabularies are sets of properties or a group of metadata elements defined by communities. This form of domain-specific semantics gives rise to semantic modularity because it creates an infrastructure which enables distributed attribute registries which may be reused, extended, and refined to fit domain-specific requirements (Nobel et al., 2020) (E. Miller, 1998).

We have briefly mentioned the SVO structure in the introduction 1. RDF uses an equivalent structure called Subject Predicate Object. This triple is the elementary building block of RDF and is formalised as $P(S, O)$. The link it denotes is that the subject S has a predicate (or a property) P with the value of O. The predicate P can also be seen as a labelled edge between two vertices S and O (Nobel et al., 2020) (Broekstra et al., 2002).

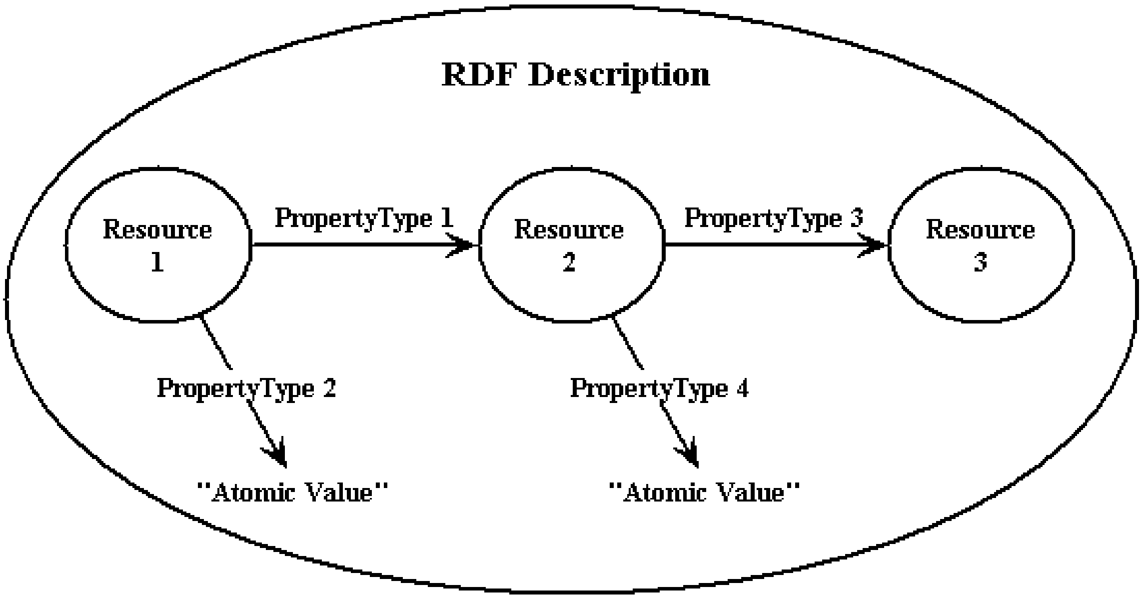

Figure 2.6: RDF description

(adapted from E. Miller, 1998 p. 3)

2.3 A-Lex

A-Lex aims to make working with all kinds of data more approachable for everyone. Therefore, a major aspect of A-Lex is its versatility (see figure 2.7). Not only can it be used for classical glossary building, but it can also be used for satisfying the information need of organisations by creating business data models using their specific natural language, e.g. a language which has developed naturally by usage (see figure 2.8). These characteristics ensure that A-Lex is capable and fit to create conceptual data models, information models, semantic networks, mind maps, concept maps, and ontologies (*meet-Alex*, 2017).

An exciting feature is that A-Lex enables people to look beyond their domain. Connecting terms from different domains give new insights into both the structure and the definition of knowledge. Exploring other domains inside the knowledge network and obtain new insights and data. Knowledge networks must either be private or public (*meet-Alex*, 2017).

Apart from the goal to make A-Lex open-source entirely, the tool aims to make data open. A vast amount of data and thus, also, data models are authoritative and also the same for everyone. For example, specific legal or insurance models are the same for every person. Building these kinds of models, with the help of crowdsourcing and making them public, will help to share data more efficiently across organisational boundaries. Models can be connected so that organisations can compare their solution models to other (roughly) similar models (*meet-Alex*, 2017).

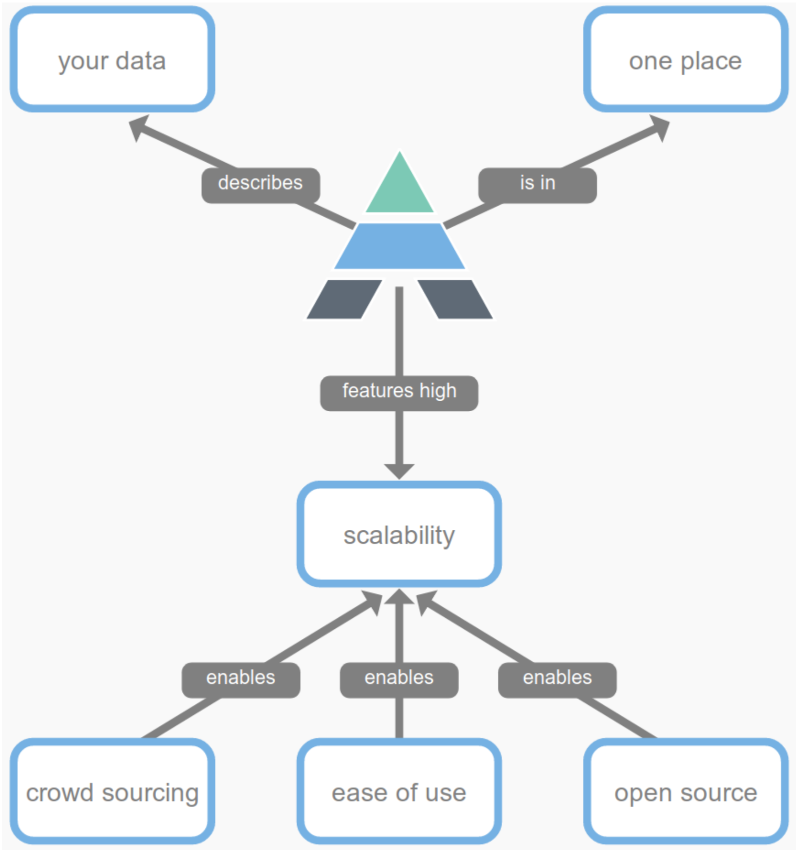

Figure 2.7: Global overview of A-Lex

(adapted from *meet-Alex*, 2017)

In A-Lex all information is organised in the form of terms. A term can be a word or also multiple words combined. For example, we can have two terms "student" and "student progression". A collection is a group of terms which are grouped. It is also possible to use terms from other collections. Therefore, multiple collections may exist which might use the same term in a different manner (both in definition and structuring). This way of working ensures management and collaboration of both self-defined and imported terms from other collections. In addition, every term has the ability to be described using text. When using the descriptions, a term can not only be described but also how it distinguishes itself from others by using and defining specific characteristics of that term (*meet-Alex*, 2017).

In order to give structure to the terms inside a collection, so-called relations are used. A relation is always between exactly two terms inside a collection and is often specific to the context of the collection in which they are defined. Different relations exist in order to give a precise definition of what the interdependence is between the two terms. The following relations exist: contains, describes, has, is a, is part of, is the role of, is same as and is a subtype of. These relation types are semantic relations based on synonyms, homonyms, antonyms, hypernyms, hyponyms holonyms, and meronyms (see section 2.1.1) (G. A. Miller et al., 1990). The combination of terms and relations creates a directed graph. This graph can either be a disconnected graph, a complete digraph or somewhere in between depending on the number of relations defined between the terms.

The open semantic relations which are not discussed in the section 2.1.1 regarding natural language, are used in A-Lex and FBM (see section 2.5). We will introduce a new semantic relationship called 'phrasal'. With A-Lex, users are allowed to define and describe their own (semantic) relations. They are the meaning of verbs or verbal phrases and are used as a way to define (new) relations. It thus uses lexical definitions as the basis for complex semantic descriptions between objects (Nobel et al., 2020). Therefore, we required a more generic term to describe these specific relations. These relations are roughly related to bridge and plain fact types in FBM and can also be used to define specific properties/attributes of object types. For example, if someone wants to make a new phrasal called "uses a". Then in FBM, we can link an object type called the farmer to an object type called "shovel". It describes a particular specific relation between the two object types, which is hard to define correctly and transparently using the default syntax of A-Lex.

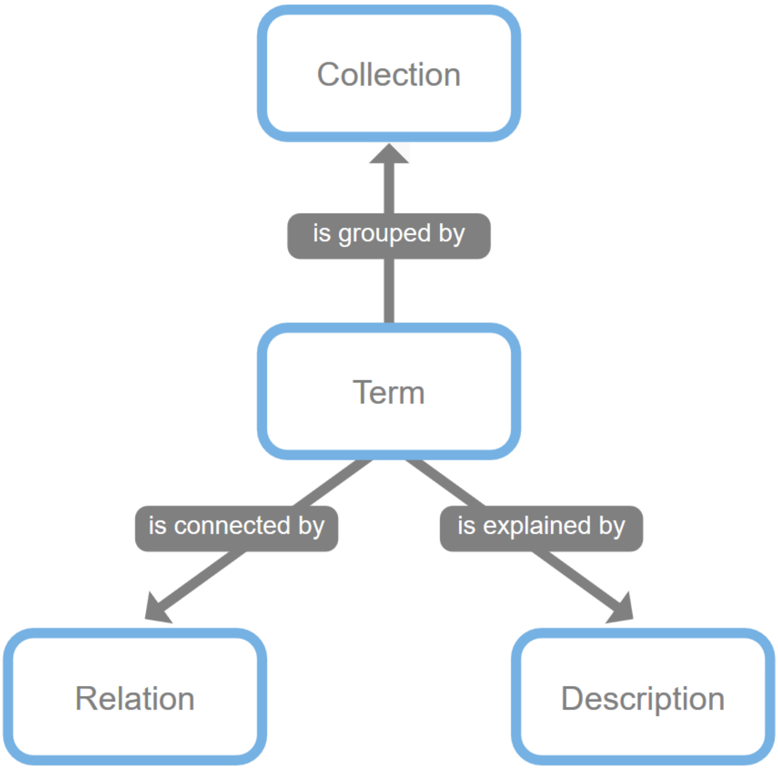

Figure 2.8: Definition of terms in A-Lex

(adapted from *meet-Alex*, 2017)

2.4 Wordnet

The Oxford English Dictionary (and other similar dictionaries) was and is of immense value to issues regarding word use and setting. However, due to the emerging disciplines of psycholinguistics and psycholexicology, the linguistic and lexical theories regarding the (synchronic) structure of lexical knowledge have evolved a lot. Linguists became aware of the need for information in a lexicon regarding the phonological, syntactic, and lexical components for everyday use and comprehension. This need led to the development of Wordnet, which can be seen as a dictionary based on psycholinguistic and psycholexical principles. Wordnet tries not to organise in terms of meaning but on word forms. Therefore, the suggestion has been made that Wordnet is a thesaurus instead of a true dictionary (G. A. Miller et al., 1990).

Wordnet is a lexical database available both offline and online. It includes nouns, verbs, adjectives, and adverbs and is grouped in so-called synsets which are groups of unordered synonyms. The way words are linked together by using conceptual-semantic and lexical relations in such a way that it can be used for both computational linguistics and natural language processing purposes.

Wordnet links words based on the 'senses' of words. Because Wordnet might seem to resemble a single large thesaurus, the proximity of different words is crucial. Furthermore, another striking feature is that Wordnet gives all the words in its lexicons so-called semantic relations with other words in the form of 'synonymous' relations. The synsets to which a word belongs also has a relationship with other synsets. Wordnet uses the same structure as A-Lex does in order to define relationships between different synsets. A word may be a member of multiple synsets if it has multiple distinct meanings. For example, the word 'rider' can mean passenger of a vehicle who is not operating that vehicle or a someone who (actively) rides an animal and thus it is present in two synsets (Princeton-University, 2019).

Figure 2.9 shows a graphical visualisation of the structure of Wordnet. The dotted rectangles are the descriptions of the synsets which are represented by the rectangles without dotted lines. The arrows between the rectangles denote the structure between different synsets.

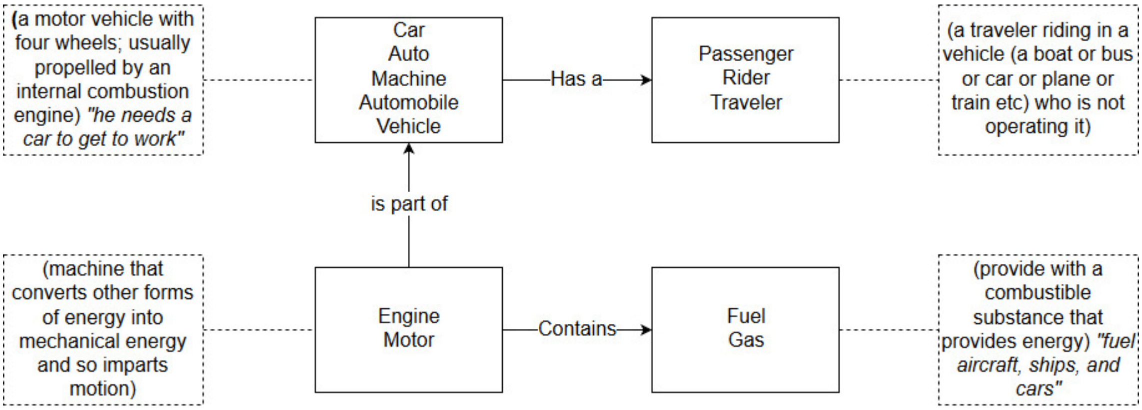

Figure 2.9: Visualisation of the Wordnet structure

Different approaches are possible in creating definitions. The classical stipulative definitions are decisions regarding the definitions of terms. A stipulative definition may or may not be useful in a particular context, but they can not be evaluated to true or false. Therefore, they are not useful in solving genuine definition conflicts. The reason for this lies in the simplification of the discourse and lack of emotive aspects (Copi, Cohen, & MacMahon, 2007, p. 88-89).

Another approach to creating definitions is the so-called lexical approach. This requires to look at the established use of a word. The result is often not a static universally applicable definition but rather a definition that depending on the context is either be true or false. This is quite useful to reduce ambiguity and explain its use (Copi et al., 2007, p. 89).

Wordnet is of immense value in computational linguistics (G. A. Miller & Hristea, 2006). In contrast to, for example, a classical Dictionary (such as the Oxford English dictionary) the origin of words in Wordnet are based on lexical definitions instead of classical stipulative or a combination of lexical and stipulative. This consequently makes computational linguistics especially interested in (formal) lexicon studies because it requires knowledge on the usage of words and how they are linguistically linked. The reason for this is because, writing text using algorithms, search text queries, and automatically organising data all become more feasible due to the requirement of access to a strict and well-formed dictionary set.

2.5 Fact-Based Modelling

A classical and formal (mathematical) way of conceptual modelling is called FBM (Fact-Based Modelling), which is an information modelling language based on set theory. It is closely related to NIAM (Natural language Information Analysis Method). Different methods based on NIAM exist, but in many cases, these methods have no formal bases and not enough attention is paid to the formal semantics. Using pictures is popular, but these can be fuzzy and unclear definitions. Without formal notations, deficiencies such as inconsistency, lack of structure, over-specification, incompleteness, ambiguity, and redundancy will emerge. These deficiencies caused by a lack of formal definition and approach do not allow for sophisticated automatic support and reasoning (van Bommel, 2019)

In this section, we will describe formal parts of the theory of FBM, which we will use to explain the foundations of A-Lex better. This theory will eventually thus be used to complete how the foundations of A-Lex are related to other concepts as part of an interdisciplinary conclusion.

2.5.1 Conceptual schemas and information structures

Information modelling aims to build an information model or a conceptual model. It only deals with functional requirements (what should the system do) and not how it should do this. It models a part of a postulated or real word which is called the Universe of Discourse (UoD) (also called the application domain or just the domain). In a broad sense, information modelling does not just include computerised aspects but also all information aspects of a business system. It is important to have an unambiguous understanding of the UoD, and therefore, it requires well-defined formal semantics which has sufficient expressive power. Furthermore, models which are not conceptual are generally not well understood by domain experts.

A conceptual schema Σ consist of both an information structure $\mathcal{I}$ and a set of population constraints $\mathcal{R}$. For the population constraints holds that $\mathcal{R} \subseteq \Gamma(\mathcal{I})$ where $\Gamma(\mathcal{I})$ is the set of all possible constraints of an information structure. This means that a valid population of a conceptual schema must be a population of the information structure and it must satisfy all constraints of $\mathcal{R}$, to put it more formally $IsPop(\Sigma, Pop) \equiv IsPop(\mathcal{I}, Pop) \wedge \forall r \in \mathcal{R}[Pop \models r]$ (van Bommel, 2019).

A visualisation of a conceptual model is not a formal description but only an illustration of the structure. Many different ways of visualising a formally defined conceptual model exist. However, the formal (mathematical) way of modelling a conceptual model has the advantage that it can be made relatively easily into other models (and visualisations). We have visualised the conceptual model and its transformations in figure 2.10:

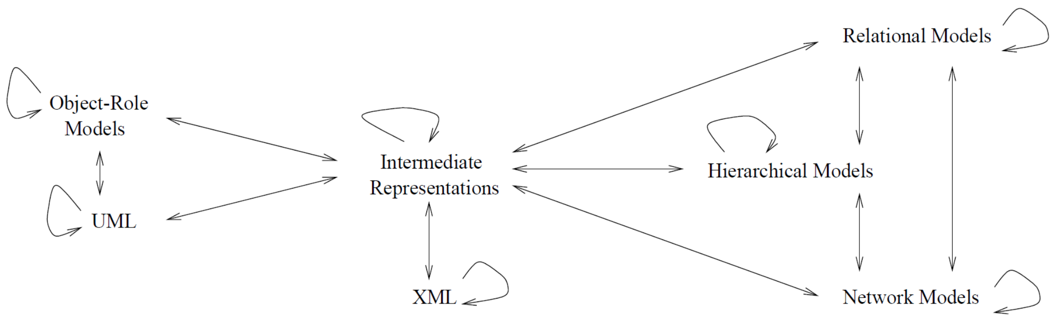

Figure 2.10: Transformation of models

(adapted from van Bommel, 2019 p. 7)

2.5.2 Population constraints

Below we will give a short introduction to FBM. Let us make a small model as an example. It shall have three so-called entity types (or vertices in a graph) and two binary fact types between them (or an edge in a graph). Formally we have the following conceptual model:

- Predicates: $\mathcal{P} = \{p_1, p_2, p_3, p_4\}$
- Entity types: $\mathcal{E} = \{A, B, C\}$
- Label types: $\mathcal{L} = \emptyset$
- Object types: $\mathcal{O} = \{A, B, C, f_1, f_2\}$

- Fact types: $\mathcal{F} = \{f_1, f_2\}$
- Fact type collection: $f_1 = \{p_1, p_2\}$, $f_2 = \{p_3, p_4\}$

Types and representation

In FBM both the fact types, as well as the entity types, are object types because for example if we only want to link a C to a combination of already existing A and B instances (so a tuple of (a, b)) then a fact type g could exist between the fact type f (which is between A and B) which we would link to C and which would enforce that only a tuple of A and B together can be linked to a C (due to fact type g having as base an entity type and another fact type). In our example, however, this is not relevant because we do not have a fact type f for which holds that its base is another fact type. It is possible to define n-ary fact types in a conceptual model. However, n-ary fact types can be reduced to binary fact types and vice versa.

Furthermore, we have so-called label types $\mathcal{L} \subseteq \mathcal{O}$, which are specific properties or attributes of object types. For example, a label type could be "name" which is linked to an object type person (which is identified by a unique number). See figure 2.11 for a visual example in the form of a small graph.

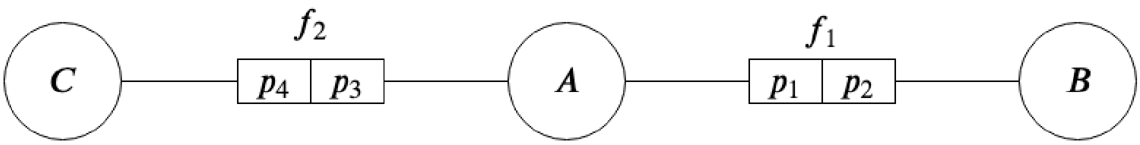

Figure 2.11: ORM example visualisation

Uniqueness

The set of predicates denoted as $\mathcal{P}$ are used to enforce special properties between object types. For example, $unique(p_1)$ would in our example mean that a valid population of A and B cannot have an instance of A, which occurs more than once. Thus if we have $pop = \{(a_1, b_1), (a_1, b_2)\})$ then this population is not valid. This population constraint can be over multiple predicates (thus the combination of certain predicates must be unique then). More formally a uniqueness constraint is a non-empty set of predictors $\tau \subseteq \mathcal{P}$. A population which satisfies the uniqueness constraint is denoted as $Pop \models unique(\tau)$. A uniqueness constraint over one fact type is defined as $\tau \subseteq f$. Although $\tau \subset f$ is more logical because the combination of all predicates in a fact type always must be unique, therefore if $\tau = f$, then the uniqueness constraint is redundant and can be omitted. It is also possible for an uniqueness constraint to exist between multiple fact types.

Total role constraint

In all populations, a constraint of an object type may be that all instances must be involved in a subset of predicates. Thus if we have $A = \{a_1, a_2, a_3\}$ and $total(p_1)$ (from figure 2.11) then the instances of f_1 should contain all instances of A at least once. We define the non-empty subset of predicates (τ) as: $\tau \neq \emptyset \wedge \tau \subseteq \mathcal{P}$. All predicates in τ must be type related ($\sim$) to each other. If object types have values in common, we will call them type related. Two predicates p and q are type-related if and only if also their bases are type related. This is defined as $Base(p) = Base(q)$. Regarding the requirement that all predicates in total role constraints must be type related, we get $\forall p, q \in \tau[p \sim q]$. Thus we can say that over some set of predicates of an object type all instances must be involved for the population to satisfy the total role constraint ($Pop \models total(\tau)$). Furthermore, just as with the uniqueness constraint, a total role constraint between different fact types may exist.

Occurrence frequency

Certain unique combinations of roles have been discussed in subsection 2.5.2 regarding uniqueness constraints. However, this can be more generalised by defining a certain minimum, maximum, range, or concrete number of combinations of roles that are allowed. For the set of predicates σ, the combinations of object type instances must occur at least n times and at most m times. This is defined as: $frequency(\sigma, n, m)$ (ter Hofstede & van Bommel, 2018).

Verbalisation of occurrence, total role, and uniqueness constraints

For both the uniqueness and the total role constraint holds that depending on the combination of both constraints, the verbalisation can differ. However, it is possible to verbalise the combinations (Halpin, 2009).

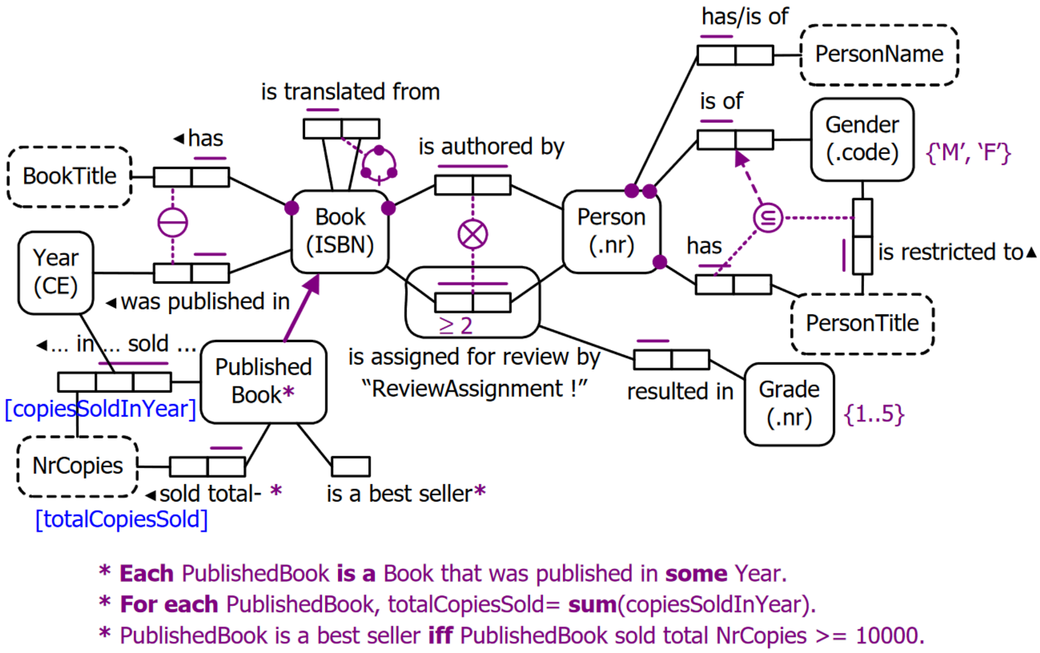

Figure 2.12: Verbalisation of occurrence, total role, and uniqueness constraints

(adapted from Halpin, 2009 p. 5)

With regard to figure 2.12 we can, for example, define the following verbalisations (not the complete list).

- Each Book is translated from at most one Book.
- It is possible that the same Book is assigned for review by more than one Person and that more than one Book is assigned for review by the same Person.
- Each Review Assignment resulted in at most one Grade.
- It is possible that the same Book is authored by more than one Person and that more than one Book is authored by the same Person.
- Each Book is authored by some Person.

(cited from (Halpin, 2009) p. 4)

We could do this for every constraint visual in the (complex) visualisation of the model. The idea behind the shown visualisation is that the population constraints in FBM can be transformed into structured natural language.

Specialisation

This relationship can be seen as the equivalence of “is a subtype of” because it denotes a more elaborated subtype. For example, a child is a more elaborated type than the more general person type and because it requires subtype defining rules. If we assume that A is a specialisation of B then we require a subtype defining rule as follows $Pop(B) = \{x \in Pop(A)|R(x)\}$.

To be more precise, specialisation is a binary non-commutative relation between two concepts or object types. If A *spec* B is the case, then it implies that A is a specialisation (or subtype) of B. In all cases must hold that a specialisation network is acyclic. If $spec^+$ is the transitive closure of *spec* then this means that $a\ spec^+\ b \Rightarrow \neg b\ spec^+\ a$ because if a could be a specialisation of b and vice versa, then a cycle would have to exist.

A subtype inherits all properties of the supertype and therefore it is required that the subtype is an entity type, more formally: $spec \subseteq \mathcal{E} \times \mathcal{O} \setminus \mathcal{L}$. A specialisation requires a pater familias (denoted as $\sqcap$) which is the object type which has no supertype. Formally the pater familias is defined as: $\sqcap(a, b) = a\ spec*\ \ b \wedge \nexists x \in \mathcal{O}[b\ spec\ x]$ where $spec*$ denotes the reflexive and transitive closure of *spec* which is not standard.

Each object type must have one pater familias thus $\forall x \in \mathcal{O}\ \exists y \in \mathcal{O}[\sqcap(x, y)]$, therefore if an object type is not a subtype it must be its own pater familias $\neg spec(x) \iff \sqcap(x, x)$ holds. Therefore non entity types (which cannot have a super type) have themselves as pater familias $x \notin \mathcal{E} \Rightarrow \sqcap(x, x)$. In fact an object type must have exactly one pater familias thus $\forall x \in \mathcal{O}\ \exists! y \in \mathcal{O}\ [\sqcap(x, y)]$. Also in the specialisation hierarchy a subtype must always have the same pater familias as its supertype thus $x\ spec\ y \Rightarrow \sqcap x = \sqcap y$. This also implies that the population of a subtype is a subset of its supertype thus we can say that $x\ spec\ y \Rightarrow pop(x) \subseteq pop(y)$ (ter Hofstede & van Bommel, 2018).

Generalisation

Although it might imply to be the inverse of specialisation, it is not because they originate from different axioms in set theory. It does not require a defining subtype rule as is the case with specialisation, but it does require that all specifiers are disjoint. Furthermore, properties are inherited upwards instead of downwards, as is the case with specialisation. This relationship can, therefore, be seen as the equivalence of “is the role of”. For example, an automobile and a motor can be seen as roles of a vehicle. The generalisation relation is defined as: $gen \subseteq \mathcal{E} \times \mathcal{O} \setminus \mathcal{L}$.

Just as with specialisation the generalisation network is also acyclic: $x\ gen^+\ y \Rightarrow \neg y\ gen^+\ x$.

Thus it also logical that the specialisation of x cannot be the same as the generalisation of x. This implies that $gen(x) \Rightarrow \neg spec(x)$ (ter Hofstede & van Bommel, 2018) (ter Hofstede, 1993)

Power type

A subset of B can become encapsulated in an instance of A. For example, a group (an instance of B) consist of one or multiple members (one or multiple instances of set A). A powerset is defined by its members, so two instances of B, which have the same instance(s) from set A are the same set and thus also the same instance of B. This is because in set theory there exists the axiom of extensionality which says that set g and set h are equivalent if they contain the same members and is formally defined as $\forall x \in g \cup h\ [x \in g \iff x \in h] \Rightarrow g = h$ (ter Hofstede & van Bommel, 2018).

This can be seen as the equivalences of the "is part of the set" when reasoning from A to B and the "contains" when reasoning from B to A (ter Hofstede & van Bommel, 2018).

Subset

This simply denotes that a set A is a subset of B thus $A \subseteq B$. This can be seen as the equivalence of "is subset of". In FBM a distinction can be made between a proper subset ($\subset$) and a subset ($\subseteq$). Other related relations such as superset ($\supset$), proper superset ($\supseteq$) and all negations ($\nsubseteq, \not\subset, \not\supset$ and $\nsupseteq$) are also possible. However, when $A \subseteq B$ then it seems that in addition explicitly defining that $B \supseteq A$ is unnecessary and redundancy. It is ofcourse possible to define additional relationships but again with respect to subsets a one way directed relation seems sufficient (ter Hofstede & van Bommel, 2018).

Equivalence and exclusion

This equivalence denotes that two sets A and B contain the same elements and therefore it can be seen as the equivalence of "is same as". Thus the previously cited axiom of extensionality must hold (ter Hofstede & van Bommel, 2018).

An exclusion constraint says that if an instance of an object type cannot be used in multiple fact types attached to that object type. For example if we have a object type A connected with a binary fact type f_1 to object type B and with another fact type f_2 to object type C, then an instance of A can either be connected to an instance of B or an instance of C but not to both. The subset of A used in fact type f_1 we will call A_{f1} and the subset of A used in fact type f_2 we will call A_{f2}. They must be completely disjoint meaning that $A_{f1} \cup A_{f2} = \emptyset$ to have a valid population with respect to the exclusion constraint (ter Hofstede & van Bommel, 2018).

2.6 Cybernetics of language in organisations

Cybernetics can be a tremendously helpful tool to describe problems of systems in the broadest sense of the term. In this section, we will try to explain the background, basics, and application of cybernetics, especially concerning natural and domain-specific language.

2.6.1 Introduction to (first-order) cybernetics

The goal of cybernetics is (according to Ashby, who was a pioneer in cybernetics/system theory) to provide effective methods for the study of intrinsically extremely complex systems. Its focus lies on 'dealing with complexity'. A system is seen as a whole of elements with relations between them showing some behaviour. Studying a system is, in most cases, a prerequisite to understanding and to control of the system. The focus of cybernetics is not on 'things' but rather on behaviour. The materiality (physical, social, mechanical etcetera) is irrelevant as long as the system involved is regular, determinate or reproducible (M. I. M. Achterbergh & J. Vriens, 2009).

This way of thinking has inspired many great scientists in many different disciplines such as biology, communication theory, physics, management, and even information sciences and computing science. Due to its general form, parts from these disciplines could not only be more elaborately explained and described but also linking different disciplines together became possible.

If we view a problem as a concrete system which shows a specific behaviour (in the sense of activities and effects (also called the abstract system)), then we can define essential variables which we want to control. For these variables, a norm/range should be drafted in order to determine if the behaviour of the system is 'wanted' or not 'wanted' (M. I. M. Achterbergh & J. Vriens, 2009). To give a small concrete example: if we have a fridge, then we can define an essential variable such as 'temperature', a range of between 0 and -20 degrees Celsius, a regulator would be the cooling element, and a disturbance would be the outside temperature. In this example hysteresis could be of influence due not having a single concrete norm value for the essential variable but rather a range for which holds that the regulator's job is to prevent the temperature from getting below or above that range by turning itself on or off (see figure 2.13).

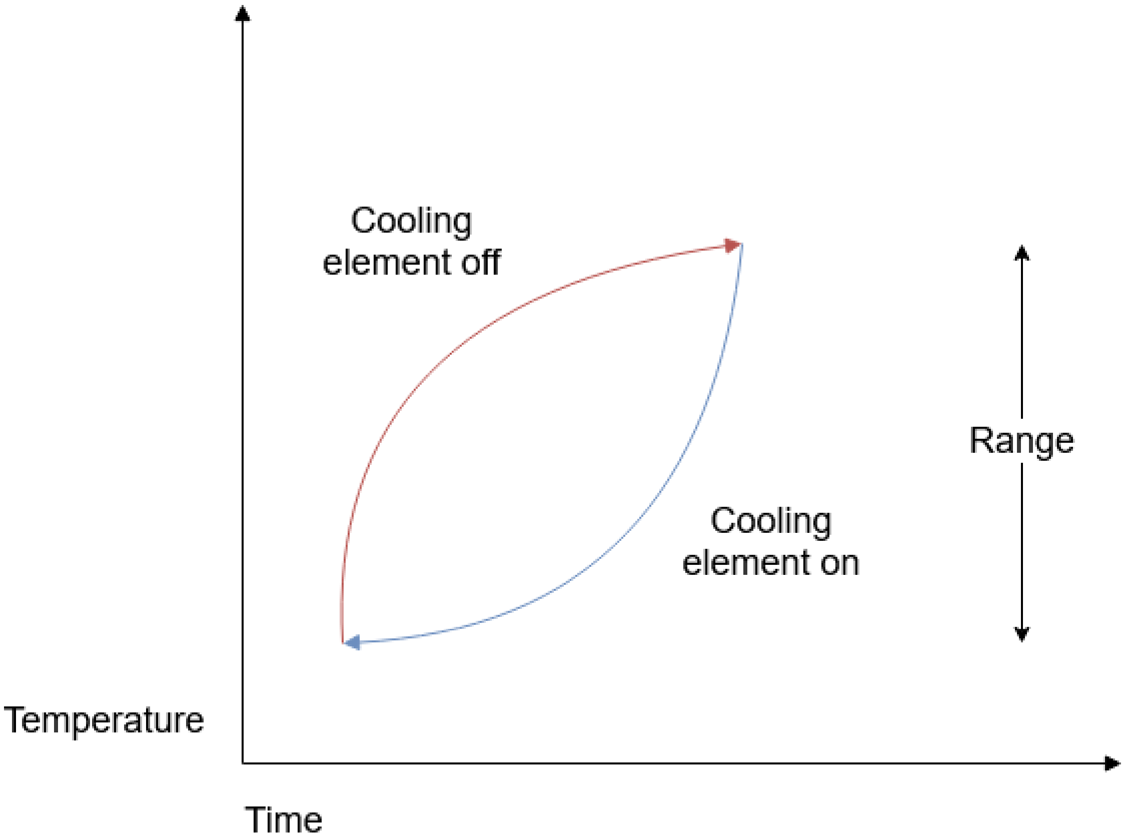

Figure 2.13: Hysteresis

Even in thermodynamics (and its computational background) the principles of entropy are an important concept on which cybernetics can be applied. Ashby's law of requisite variety states that the regulation of a system is rooted in the observation that the set of regulation states must be greater than or equal to the set of disturbances of a system. For example, a similar statement concerning thermodynamics would be "The algorithmic entropy introduced by the correction process must equal the algorithmic entropy introduced by the disturbance" (cited from (Devine, 2017). In both these similar statements, the aim is an equilibrium (a balanced state concerning disturbances and regulations within a particular range/norm. See figure 2.14) where the disturbance pushes the behaviour of the system from the equilibrium, and the correction pushes it towards the equilibrium (Devine, 2017). See figure 2.15 for an example of a simple cybernetic system.

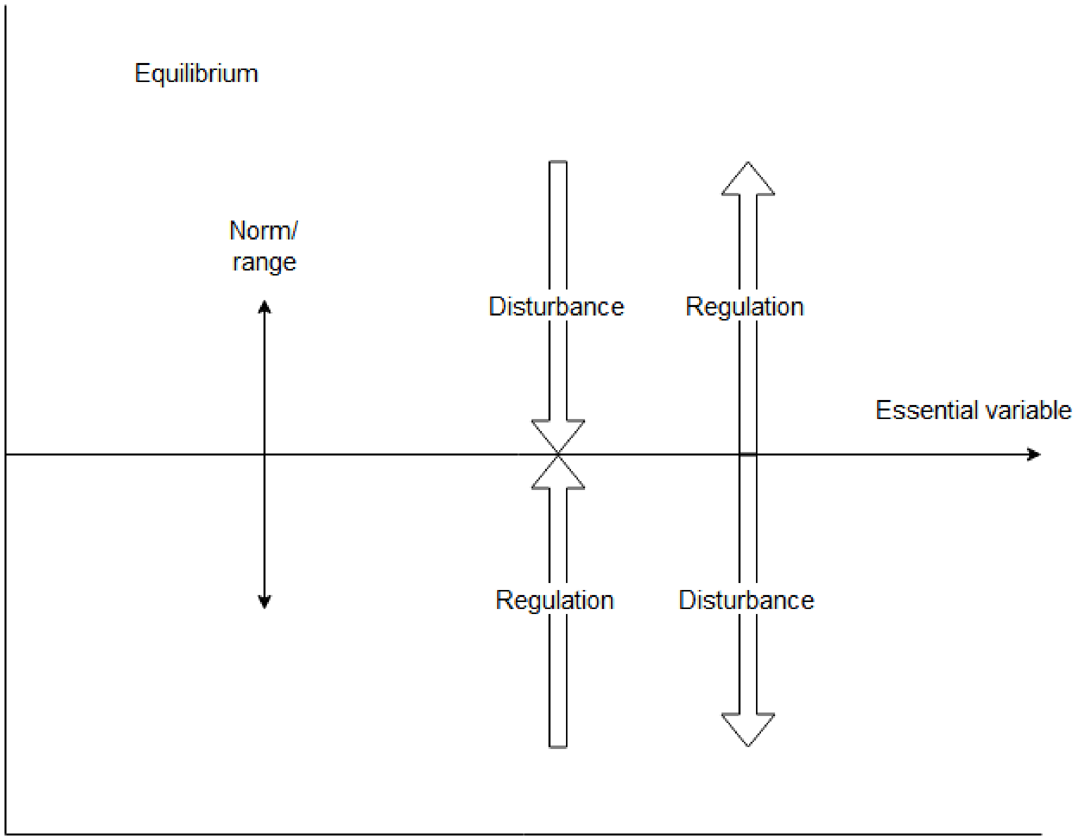

Figure 2.14: Equilibrium

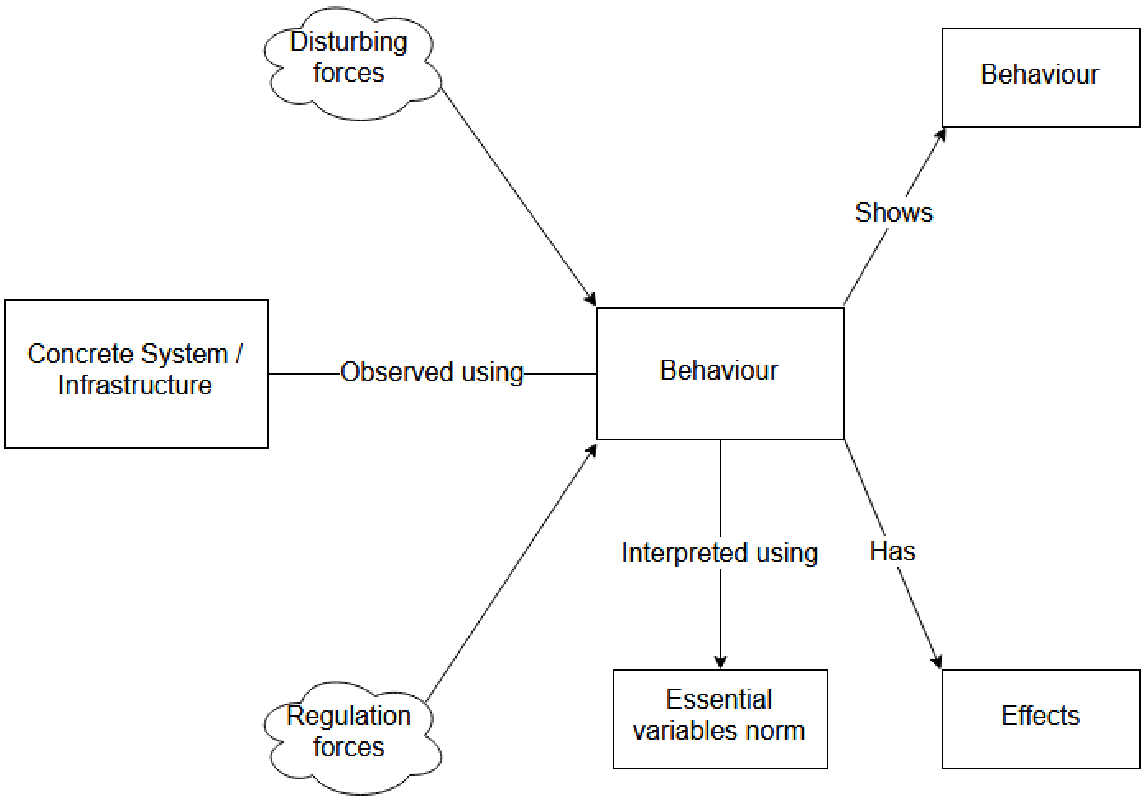

Figure 2.15: General system example

One relevant example of the power of cybernetics is the Algorithmic Information Theory (AIT) sub-field (part of computing science and mathematics), which aims to build a bridge between Alan Turing's work on computation and Claude Shannon's work on information theory (mostly a combination of natural language processing and computational linguistics) (Burgin, 1990).

2.6.2 Second order cybernetics of social systems

The question of what syntax exactly is has been answered by Alan Turing with his famous Turing machine (see also section 2.12). As has been described in (Nobel et al., 2020), the problem with dealing with lexical semantics, however, has not yet been solved. To cite the paper concerning this issue: "Apart from obviously being neuro-cognitive in nature, language is also (socio-)cultural: a shared and distributed phenomenon (Steels, 2013). Required language adaption in combination with the complexity of the environment and context in which a domain language is used (see figure 2.16) will result in necessary evolution of its concepts, terms and semantics, and inevitably leads to diversification and divergence among organisational domain languages" (Nobel et al., 2020)).

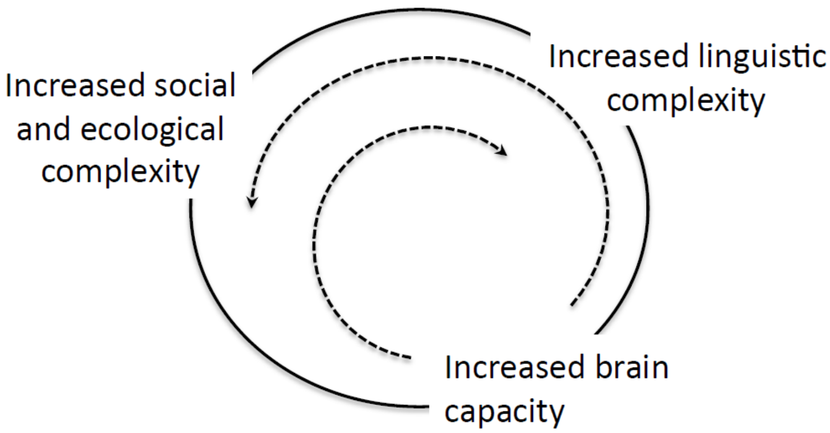

Figure 2.16: Language complexity increase

(adapted from De Beule & Stadler, 2014 p. 3)

Cybernetic principles can be applied in researching semantic domain-specific language issues through a theoretical link between information theory (or as Claude Shannon called it the "theory of communication") and the evolution of languages (De Beule & Stadler, 2014). In second-order cybernetics as defined by (among others) Niklas Luhmann, the generality of first-order cybernetics is replaced by specificity. Organisations are, in the view of Luhmann, a particular class of social system which continuously conduct 'experiments' because the outcome of a decision cannot be determined beforehand.

2.6.3 Autopoietic systems

In (second-order) cybernetics by Luhmann, the generality of cybernetics is replaced by specificity. The social 'arche' defines organisations as a particular type of social systems (see figure 2.17) which are experimenting continuously and states that there is a relation between evolution and the diversity of a language. A social system such as society but also organisations are autopoietic (self-producing) systems exactly in the same way an organism is a system of communicating cells. In an organism, the cells are the elements, and in a social system 'communication' forms the elements. The elements of an autopoietic system are the smallest elements inside that system which must produce new elements in order to sustain itself by the element it consists of. A system cannot exist without any elements because it would not be able to reproduce itself. Elements can be very complex or extremely simple. An element could be a cell in a living organism which is producing new cells or communication within an organisation which is based on (all kinds) of previous communication and produces new communication elements. When new elements are created the old interacting elements (due to which the new elements are created) become lost in the process. A system, therefore, must always have at least two elements (otherwise it could not produce new elements) (M. I. M. Achterbergh & J. Vriens, 2009).

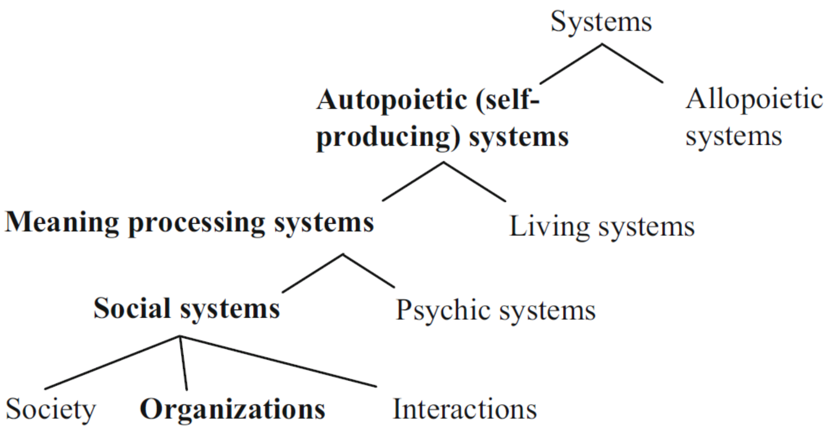

Figure 2.17: Luhmann's classification of systems

(adapted from Greshoff, 1999 p. 70)

A system has to have control functions which will prevent rogue elements from emerging and control the interactions (processes) of elements inside the system. It is imperative that the set of all possible 'complexes of things and processes' which could be produced by a system is reduced to a subset of particular elements which are likely to keep the system (and its internal processes) going. Let us define the following two sets V_p and V_l which denote the possible elements and likely (most suitable) elements respectively for which holds that $V_l \subseteq V_p$ and in most cases $1 \leq |V_l| \ll |V_p|$ will hold. If at some point in time t the system consists of S_t related elements then the structure of the system will restrict the possible set of elements that can be produced V_p to 'likely' elements V_l. A newly generated element $e \in V_l$ is accepted into the system and will now become part of the entire system and therefore for the system at time $t + 1$ holds that $e \in S_{t+1}$. This process keeps repeating during the lifetime of the system (M. I. M. Achterbergh & J. Vriens, 2009).

Information, utterance, and understanding

Communication, as the elements of a social system, is a synthesis of three components: Information, Utterance, and Understanding. These elements together form the inner structure of all forms of communications (Gabriel Calise, 2016). Between two communication partners, one is responsible for the Information and the Utterance and the other for the Understanding. Information is the focal content of the communication or in other words, the subject of the communication. The Utterance is about how the Information 'is uttered'. With Understanding, we mean that the Utterance was meant for the uttering the Information or not. For example, the cause for someone to slam a door can be to convey annoyance to another person or simply because the person feels like slamming the door. It is explicitly not about if the other person understands the meaning of the Information of the communication (M. I. M. Achterbergh & J. Vriens, 2009).

- Information → Event that selects states of a system.
- Utterance → Enunciates information.
- Understanding → Distinguishing and maintaining the difference between utterance and information.

(Gabriel Calise, 2016)

Expectations about expectations

Luhmann describes so-called expectations about expectations. If person A talks to person B, then person A has expectations about what person B will expect of person A concerning their communication (for example, a conversation). If they, for example, talk about subject X which is not related at all to subject Y then person A expects that person B expects of person A that the response of person A will be a (relevant) response about subject X and not suddenly about a non-related subject Y. However, sometimes, it is not always clear if a response is expected or not. Then if such a response is given the receiver can do two things, namely either accept this response as acceptable (and therefore also acceptable in the future and thus change the system) or not accept the response. Expectations about expectations are required to structure the production of communication by communications. Therefore, Luhmann argues that the set of options for communication is connected to prior communication (M. I. M. Achterbergh & J. Vriens, 2009).

A (social) system must have some kind of control function to prevent rogue elements from emerging but also to control the interaction (processes) of the existing elements. Decision premises is an essential term in Luhmann's theory about organisations with regard to decision making in organisations. He has defined the following nine decision premises: Membership, Communication Pathways, Decision Programs, Personnel, Position (of a member of the 'Personal'), Planning, Self-Descriptions, Organisational Culture, and Cognitive Routines. He argues that the rejection and acceptation of new decisions into an organisation depends on so-called decisions premises, and he also argues that an organisation is simply a whole of meaningful and consecutive decisions. The different decision premises in an organisation influence whether a decision is accepted as meaningful or not. Some of these decisions premises can be changed based on the decisions of the organisation (M. I. M. Achterbergh & J. Vriens, 2009).

2.6.4 Language and information system evolution

Language is not a static system but rather an incessant changing and adaptive system[3]. It adapts to the context of use, which can be incidental, situational but also long term. Evolution of a language is in most cases, inevitable regardless of whether it changes for the worse or, the better. A language (at least elements and the semantics of elements) can also form in an organisation. For example, high tech organisations which come up with new products, technologies, ideas, etcetera need to be able to find a way to communicate about these concepts. Social systems are often heavily influenced by stochastic processes. Society, organisations, families, and other human social systems are probably more influenced by these processes than most other chaotic systems (Teixeira de Melo, 2015). If the context-driven change takes place and all of the contexts are different for the involved systems, then diversity in systems and languages will take place. In data governance (such as business glossaries (language) updates and (system) descriptions have to be continuously updated in a similar way as during the evolution and development of an information system. Therefore, focusing on business-language alignment and change management/processes for both systems and languages is essential. This subsequently means that there is a clear link between the disciplines of cybernetics, linguistics, information systems, and organisational change. Without language standardisation, any other kind of standardisation can prove to be very challenging within a social system such as an organisation.

The explanation of cybernetics on the evolution of (domain-specific) languages and its inherent incessant changing and adaptive properties means that organisations should put effort into effective communication. Given that the evolution of languages is a fact of life and that organisations as social systems are heavily influenced by their (complex) environment. Organisations should, therefore, communicate (both among themselves and with other domains) and decide the meanings of their domain-

[3]Legal historian Uwe Wesel, for example, said that defining the term 'law', is like hammering a pudding to a wall. Thus, quite impossible (Hildebrandt, 2019)

specific language and preferably write them down in an organised way. Earlier, we talked about the advantages concerning the power of the crowds of the web 2.0. A-Lex gives organisations a structured approach to use the advantages of web 2.0 in order to solve seemingly inevitable problems emerging from lexical definitions due to the evolution and diversification of language in organisations.

The changing of languages and the intrinsical connection between organisations and communication means that we should focus on giving organisations (and outsiders) the possibility to study (and maybe even control) the process. If we see organisations as complex systems continuously changing their elements by adding, deleting, and updating them, we can use cybernetics as a tool and an argument simultaneously so that the continuous changes should be addressed in some way especially since many organisations appear from the outside as a black-box. The exact processes and details of that organisations are hard or even impossible to find out. Even consultants and clients often do not know all the details. This can be due to organisations secrets or simply because it is not feasible to learn all the details and processes. Therefore, the most obvious solution to this problem would be to let organisations continuously update their business language and let them model their processes. When using crowdsourcing and self-service modelling organisations can (to a comfortable degree) remove their black-box and learn from each other. This would also mean that business language can be standardised more because organisations can look at each other's definitions and processes to align better or to improve themselves.

2.7 Crowdsourcing and self-service modelling

For the relatively small group of data analysts/modellers, it is practically impossible to create the diverse and extended data models required in today's society. A form of scalability was thus required to give a wider public the chance to use and build data models in a speedy, efficient, and effective way. Almost all companies, organisations, and institution require some form of data models. It might be for designing a software system, for structuring large amounts of data or, as will be discussed in this thesis, to be able to create common lexical understanding within an organisation. Crowdsourcing is, in this case, a way to capture data and organise them (Surowiecki, 2004). Using crowdsourcing also gives rise to new problems which will be addressed in this section.

2.7.1 Crowdsourcing in other lexical applications

Wikipedia and several other internet encyclopedias use crowdsourcing to extend their lexicon. They often use an enormous public or a community to define the content of their encyclopedias. Put differently, they are using the public or a group of people to create a network of somehow related definitions (and give meaning to them) with the help of a set of pre-described rules by the encyclopedia. It is hard or even impossible to claim that these rules are genuinely (formal) syntax, but clearly, there is a link between how these kinds of encyclopedias extend their lexicon, define relations between terms, and their usage of crowdsourcing concerning the idea of how A-Lex should operate. These links can be drawn upon many other public accessible and editable data sources from open-source codebases (for example GitHub or a similar development platform) to Fandoms (encyclopedias focused on a specific topic such as science-fiction or gaming) (Nobel et al., 2020). A major difference between A-Lex/Wordnet/FBM compared to Wikipedia is that the former serves the primary purpose of lexical disambiguation.

In the early days of the internet, the traffic was mostly one-directional. Users only took and did not contribute to the content of websites. This changed with web 2.0 of which Wikipedia and social media are a great example. The traffic became bidirectional. Users did not only take information but also contributed information. This gave rise to new problems but also to a vast number of possibilities regarding user-created content and (participation in) collaborative online efforts (Muñoz-Leiva et al., 2012). Usually, the mechanism for compiling an encyclopedia is rather slow. An administrative structure is being set up, contributors are recruited, waiting for their input, compensate them (often financially), and then when the result is there, it must be edited, compiled, and distributed. This can take much time, not rarely even multiple years. However, with an internet encyclopaedia like Wikipedia, many of the previously mentioned issues are not relevant. The content of Wikipedia is continuously being assembled by a vast number of contributors who are not compensated whatsoever, are working on a voluntary basis, without background or professional qualifications checks, and errors are caught by the users (or reviewers) themselves which will (hopefully) result in edits (Nobel et al., 2020) (Goodchild, 2007).

In many articles, blogs, social media pages, and web pages, so-called tags are added. These tags or labels and the creation of the labels (tagging) are used in folksonomies. With the rise of web 2.0, a taxonomy of user-created information has emerged; hence, a portmanteau of the word 'folk' and 'taxonomy' has led to the use of folksonomy. It allows people to give meaning to terms in data which they share, consume, and generate (Gruber, 2007). The categorisation of content is a collaborative effort by generating open-ended tags which, on the contrary to professionally developed taxonomies that do not have a controlled vocabulary and are inherently open-ended. Folksonomies can also respond to and initialise changes and innovation rapidly by classifying web content (Murugesan, 2007).

Now that folksonomies have arisen and given the 'social' web an enormous impulse there are a lot of Semantic Web technologies which can be applied to the

Social Web such as a formal specification of structured data and reasoning across different data sources. This will enable technologies for searching, aggregating, and connecting people and their content (Gruber, 2007).

2.8 Different viewpoints of A-Lex, Wordnet, FBM, and (lexical) semantics

In figure 2.18, we can see an example of a rather simple data model. If we take Wordnet, for example, then we would use the terms attached to the person to define what a person is. For example a person is someone who always must have (denoted by the big dot and meaning all instances of this type must have. See subsection 2.5.2) exactly 1 unique (denoted by an arrow. See subsection 2.5.2) Social Security Number, must have exactly one unique nationality, have precisely one birth date, and can own 0, 1 or multiple houses. However, if we look at FBM, we would be interested in, for example, the relation between a person and a social security number. We do not want that a social security number could be connected to a person more than once (or the other way around), and we would want every person to have a social security number. We could check this by only looking at the composed values of fact type 1 (F_1) (also called population) and look if any of our defined constraints are violated. The goal of this model could, for example, be to look up a person in a database with a specific social security number or to find out of a given person what his/her social security number is.

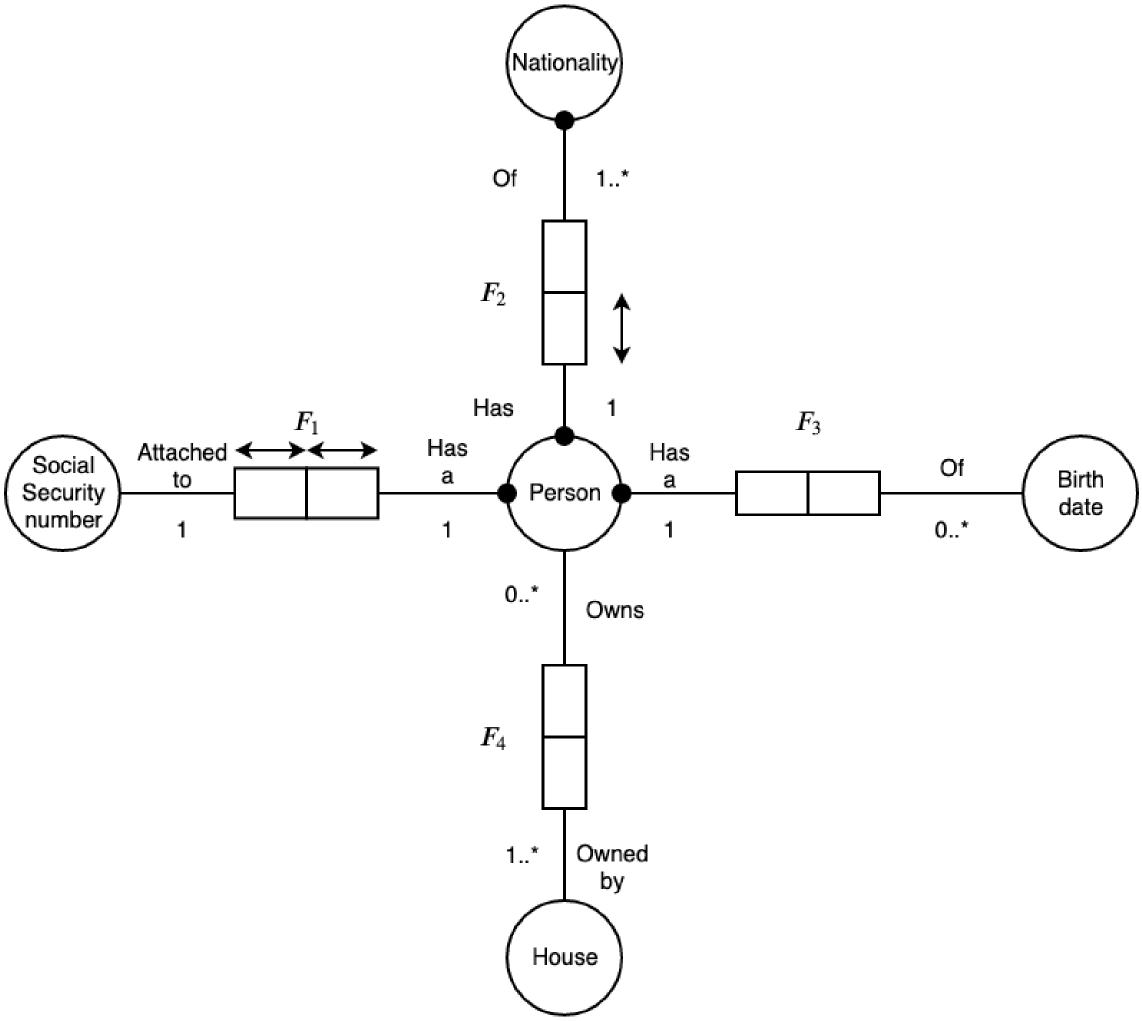

Figure 2.18: Example data model

A significant and critical difference between FBM, Wordnet, and A-Lex is that the focus point differs. FBM emphasises on the relations between objects. For example, solely the atomic value *John* from an object type *Student* is not useful for the application domain of FBM projects. On the contrary, the fact type between *Student* and *Course* is useful because it gives information in the form a composed fact type. This can be seen as a tuple of, for example, (*John*, *mathematics*). Supporting the generation of composed values is the goal of building an FBM model, and the composed values itself (which are instances of fact types and bridge types) are what is required in the application domain. The object types are facilitating the real goal of the modelling, which is defining relations (therefore Object **Relational** Modelling). RDF, and for that matter all applications that deal with lexical semantics, on the contrary to FBM are putting the emphasises on the meaning of the terms themselves. The relations between terms (the composed values) are there to facilitate and clarify the semantic meaning of the term (atomic value).

Combining parts of FBM in the way that it defines semantic relations as so-called phrasal with lexicology would be ideal. FBM originally was used to combine logic with linguistics/language. With the exception of complex axioms, this is also one of the purposes of A-Lex, using self- and well-defined relations between terms to focus not only on terms but also on the relations themselves. Using the Semantic Web as an example for A-Lex, we can learn from the way RDF is used as an easy way for unambiguous data exchange, which is both readable for humans and machines. The way the Semantic Web uses RDF in combination with other technologies such as OWL and SKOS is an example of a similar idea of how to deal with a lexical semantics.

2.9 Wittgenstein and foundations of (structured) natural language

Wittgenstein was a famous language philosopher who focused on the philosophy of language, especially regarding the boundaries of language (its limitations).

In his work 'Tractatus Logico-Philosophicus' he writes about three proposition series. The first one deals with the nature of the world, the second one deals with the nature of language, and the last one deals with the nature of logic. This critical work aims to explain the limits of the expression of thoughts using natural language (Wittgenstein, 1922).

The problem he aims to solve concerns the requirements which are needed to define a logically perfect language. He is concerned with *accurate symbolism* which means that in this form of symbolism, the 'meaning' of a sentence is something quite definite. This is, of course, on the contrary to what happens in practice. In practice, assertions and language are quite vague and imprecise. According to the preface of the book (written by Bertrand Russel), a logically perfect language: "has rules of syntax which prevent nonsense, and has single symbols which always have a definite and an unique meaning" (cited from (Wittgenstein, 1922)).

We believe that the fundamental and quite abstract theories of Wittgenstein can be used to explain both the similarities and differences between FBM, Wordnet, and lexical networks/glossaries.

2.9.1 Objects, things, entities, and the world

When defining that the facts in logical space are the 'world' (in our case the term universe is more suitable), we assume that the universe determines what is the case and therefore also what is not (1, 1.11, 1.12, 1.13, 1.14). We will define what is the case by defining atomic/elementary facts, which are declarative statements and consist of a combination of objects which can be entities or things. A thing or entity must be part of an atomic fact and has a connection with other things or entities because we cannot think of an object apart from the possibility of its connection to other things (2, 2.01, 2.011, 2.0121). Even though a thing is independent, it has a connection with an atomic fact and is therefore somewhat dependent. Knowing what an object means must also include knowing its possibilities in atomic facts, this includes not only external but also internal properties. All objects together, therefore, also give all atomic facts, and this means that all 'states of affairs' are also known (2.0122, 2.0122.1, 2.01231, 2.0124, 2.014).

Subsequently, now that we have defined an object, we can also further define complexes (such as worlds and universes). A statement about the constituent parts and propositions of a complex completely describes that complex. Objects are a substance of the world and therefore are not compound. A thing must have some properties others things do not have otherwise distinguishing them would prove to be impossible (2.021, 2.02331). An object is fixed, but its configuration is the changing variable, and this subsequently forms the atomic fact (2.0271, 2.0272). Atomic facts hang into one another like chains which ensure that the objects are combined. The way they hang onto one another determines the structure of the fact and the totality of the existing atomic facts determines the current state of the world, but they also determine which atomic facts do not exist at present. (2.032, 2.033, 2.034, 2.04) Atomic facts are independent of one another, and therefore the existence or non-existence of an atomic fact does not give any definite information about the existence or non-existence of other atomic facts. Pictures of facts are how we present facts in logical space and how we model reality. It consists of a certain configuration of facts and is a fact itself which is also called the structure. The picture is therefore linked to reality and reaches up to it (2.05, 2.06, 2.061, 2.062, 2.063, 2.1, 2.11, 2.12, 2.14, 2.141, 2.15, 2.151, 2.1511).

The picture can represent every form and therefore, can depict the world in a logically false or right way. The logical form it represents is the same as the world it represents. By showing the existence or the absence of atomic facts the picture represents a state of affairs and therefore it can be evaluated to right or wrong (true or false). This can only be checked by comparing it to reality and cannot be determined based on solely the picture itself. (2.18, 2.181, 2.182, 2.19, 2.2, 2.201, 2.202, 2.203, 2.21, 2.223, 2.224) An atomic fact must be thinkable and imaginable. All true thoughts together are pictures of the world, and the thoughts contain a possible state of affairs. What is thinkable is possible, and we cannot think anything illogical. This results in a language where we cannot present anything that contradicts logic. This means that a propositional sign is a fact which makes it a form of expression or declaration (we have seen a similar definition earlier in section 2.1.4 concerning the semiotic ladder). For example a propositional sign aRb means that a stands in a certain relation to b. The propositional sign consists of elements combined in a definite way. Facts express the sense, and the propositional sign is, therefore, a fact concealed by written expression. The mutual spatial proposition of things expresses the sense of a proposition (Wittgenstein, 1922).

2.9.2 Expressions as the irrevocable structure and only certainty

The name represents the object in a proposition, and their naming is represented by signs. It is possible to speak about signs but not to assert them. Therefore, a proposition does not mean what is but rather how is. Again a complex stands in relation to its constituent parts, which can only be explained by its description, which can be right or wrong. A proposition must be expressed in a definite and clear way. The name can not be further analysed because it is a primitive sign. The expression of a sign is signified by the signs of which it consists. What is not clear from a sign is declared by its application. An elucidation is a proposition composed of primitive signs which can only be understood when the elucidations are known before.

A proposition is what makes sense, and therefore the context of a proposition gives a name meaning. The parts of a proposition which characterise its sense, and thus also form and content, can be called an expression. The expression itself is constant, and everything else is a variable. The variables of the expression are its propositions (an expression marks a class of propositions) and are called propositional variables. This implies that we can see a proposition as a function of the static expression it contains. Multiple propositions can contain the same static expression, which is represented by the propositional variable. We will view this as $f(x)$ where the propositional variable is denoted as x. The form mixed with the content of f we may view as the expression itself, while a proposition (again an expression may encompass multiple propositions) is the combination of particular expression in combination with a specific instance of the propositional value x. A proposition given the current universe we will see as the projective relation with the world, e.g. the propositional sign (3.12, 3.13) (Wittgenstein, 1922).

2.9.3 Fundamental problems of (understanding) natural language

A word can signify different definitions. Many examples of these are homonyms (see 2.1.2). This gives rise to many fundamental problems, of which, according to Wittgenstein, philosophy is drenched in[4]. Symbolism can be used which excludes this problem. By obeying logical grammar/syntax and by assigning a symbol to a propositional sign together with its logical syntactic application, we can describe the true expression. According to type theory, no proposition can contain itself just as a function cannot contain itself. According to Wittgenstein if $F(f(x))$ could contain itself, then it would be possible that a proposition exists $F(F(f(x)))$, but this would require different meanings for the inner and outer functions F. We can define these meanings as ϕ for the inner form and ψ for the outer form. The concatenation of this will be $\psi \circ \phi \circ f(x) = \psi(\phi(f(x)))$. A function cannot be its own argument because the functional sign already contains a prototype of its argument, and this can never be the function itself. Therefore, we can conclude that just the function or proposition name F signifies nothing, but we require a clear symbol to signify the proposition. (3.323, 3.333)

A definition is a set of rules required to translate one language into the other. Meaning that every symbolism must be translatable into others using a set of rules. This set of rules is something we can see as a function. For example $x \in d_1 \iff f(x) \in d_2$. Thus, a symbol signifies that what is common to all the symbols which can be replaced using the rules of logical syntax. We can give many examples of this. A famous example are the De Morgan laws $\neg(p \wedge q) = (\neg p) \vee (\neg q)$. This kind of logical term rewriting is fundamental to truth functions (which we will discuss in 2.9.4).

A proposition defines a place in logical space and is guaranteed by its constituent parts. The totality of propositions is the language, and a propositional sign itself is the applied thought. Colloquial language can be regarded as a different object then the thought itself. Natural language disguises thought in the sense that even though we can think of every possible expression, we cannot correctly express it because the expression itself and the logic of language are two different utterly different objects[5]. (4.002, 4.003, 4.0031)

A proposition is a picture of reality as we see it and think it is. The picture aRb is

[4] According to Wittgenstein: the totality of true propositions is the total natural science and philosophy is not a natural science. Philosophy is something which does not stand beside the natural sciences but is rather an activity for the logical clarification of thoughts using elucidations.

[5] According to Wittgenstein this also means that most questions and propositions of philosophers are senseless rather than false or true. If we do not understand the logic of natural language, then we can only state that both the question and the answer are senseless or that the problem is not even an actual problem.

perceived as a picture where the sign is the portraiture of the signified. (4.023) The proposition determines reality to the extent that it can be answered with yes or no, or equivalently true or false in order to make it agreeable with reality. A proposition again is a description of a fact. The description of an object, however, is described using its external properties, where the proposition describes realities using internal properties. Therefore propositions help to see the world using logical scaffolding and thus we can determine whether or not all logical attributes and features of that proposition is true in reality.

Understanding the meaning of a proposition implies understanding what it means if the proposition is true or false. This can be done exclusively if the meaning of constituent parts is also understood. Translating one language into the other therefore requires that all constituent parts of its propositions are translated (4.023, 4.024. 4.025). If we are able to express ourselves adequately (which we do with propositions), then we must also be able to understand the simple signs (words). Essential here is that the propositions communicate new sense using old known words and also because it communicates a state of affairs, it is essential that the proposition is connected with the state of affairs. Therefore, this connection is the logical picture in which the proposition asserts something from the picture. Whether a proposition is true or false depends on the picture of reality (4.06).

An important note is that in logic $\neg\neg p = p$. However, we require a proposition to have sense independent of the facts. p does not signify true in the same way as it signifies $\neg p$ in the false sense. If we say that if $p = true$ if and only if a dot on a piece of paper is black and we say that $p = false$ if and only if it is white. Then we must first make sense of what it means for a dot to be black or white. If a dot is not black, it does not automatically mean it has all the conditions for it to be regarded as white. The determination under what conditions a dot is white or black is called the sense of a proposition. Assertion can not determine a sense of denial because it asserts or denies the sense itself (Wittgenstein, 1922).

2.9.4 Atomic facts, propositions, features, inner relations, and truth functions

Properties, atomic facts of objects and facts are, in a certain sense, the same as formal relations and relations of structures. Internal properties, however, on the contrary to external properties cannot be asserted by propositions. An internal property of a fact is also called a feature. An internal property of an object must always be possessed by that object. An internal property is therefore not expressed by a proposition. Wittgenstein's gives as an example: "there are 100 objects". According to him, this is as senseless as to say that at $2 + 2 = 4$ at 4 o'clock. He also argues that words as "complex", "fact", "function", and "number" are senseless.

Because formal concepts are presented by variables. Therefore we could say that x is the proper sign of an object. We could say that $(\exists x,\ y)$ denotes for example: "there are two objects which". Where the object can be a thing, entity etcetera.

Wittgenstein gives the example of a general logical proposition "b is a successor of a. This expression of a formal series (an ordered sequence of inner relations) means that we could denote this as aRb, $(\exists x)\ :\ aRx\ .\ xRb$, $(\exists x,\ y)\ :\ aRx\ .\ xRy\ .\ yRb$, ... Because either (we will denote succeed as $\succ$), in sequence, $x \succ a$ and then $b \succ x$ or, again in sequence, $x \succ a$, $y \succ x$, possibly multiple other $y \succ x$, after which $b \succ y$.

Depending on the truth value of an atomic proposition, the fact exists or not. If the truth value is true, it exists, and if it is false, then it does not exist. Therefore, the total number of possibilities concerning n atomic facts is $\sum_{v=0}^{n} \binom{n}{v}$. Where $\binom{n}{v} = K_n$ is the binomial coefficient[6] because it is possible for all combinations of atomic facts to exist but also not to exist. The number of combinations of n existing elementary propositions thus corresponds to the same number of truth and false possibilities.

A proposition expresses the truth agreement or disagreement with the truth possibilities of elementary propositions (see tables A.1 and A.2). Understanding propositions requires understanding all underlying elementary propositions (and the logical connectives). If we take into account the possibility of agreement and disagreement of propositions then the truth possibilities of n elementary propositions increases to $\sum_{K=0}^{K_n} \binom{K_n}{K} = L_n$. As a logical consequence, the agreement or disagreement of the elementary propositions thus expresses the final truth condition of the proposition. Alternatively, in Wittgenstein's own words: "the proposition is the expression of the truth-conditions". The L_n groups of truth conditions of truth possibilities can be ordered into a sequence as we have seen before. A tautology and a contraction, however, are without sense. Because the entire idea of truth conditions becomes invalid, it seems quite obvious that something is true if the subject is a tautology. Therefore defining it seems senseless. However, that something is true because it seems obvious is not a reason for believing something to be true.

If p follows from q and q also from p then it is the same proposition[7]. Independent propositions have no truth arguments in common with others (5.152). The structure of a proposition can be compared to another proposition by looking at its internal structure using a particular operation or function. Formal properties of propositions can be used to determine what can be used to make one proposition out of another. The previously shown series is the operation from which we have derived one proposition from the other. It shows the sense, by being an operation which shows us how we can go from one proposition to the other. It, therefore, shows the differences in forms between propositions. Thus the same operation exists which

[6]The binomial coefficient is defined as $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ for $0 \le k \le n$ and $\binom{n}{v} = 0$ for $k < 0$ or $k > n$. See also the triangle of Pascal for the values of the coefficient

[7]Wittgenstein uses $\supset$ instead of $\to$ as a symbol to denote a logical implication.

goes from q to p and from r to p because they are simply variables giving a general expression to formal relations between propositions. Again, a function cannot be its argument, but if the operation has a specific proposition as a base, but the result of that operation can be its basis. Using a specific proposition or term as a base, we can show how to progress to another. Also, according to Wittgenstein, no casual nexus exists from which we can infer one elementary proposition to another. With the operation, we do not mean infer, but to go from one proposition to another. There is not a definite beginning of an elementary proposition from which we can reach all others.

A proposition is the result of a finite number of successive truth operations on elementary propositions. The truth operations build the truth function itself. Alternatively, to quote Wittgenstein: "Every truth-operation creates from truth-functions of elementary propositions another truth-function of elementary propositions, i.e. a proposition. The result of every truth-operation on the results of truth-operations on elementary propositions is also the result of one truth operation on elementary propositions. Every proposition is the result of truth-operations on elementary propositions." (Wittgenstein, 1922).

To apply the logic to this problem, we can look at the logic from Wittgenstein's theory. The logic using elementary sentences and truth functions can help us to make A-Lex a more accurate and correct tool to define domain-specific languages. According to Wittgenstein, the general truth function has been defined as $[\bar{p}, \bar{\xi}, N(\bar{\xi})]$. Where $\bar{p}$ stands for the set of all elementary fully developed propositions, $\bar{\xi}$ stands for any subset of the propositions and $N(\bar{\xi})$ stands for the negation of all propositions in $\bar{\xi}$. With a proposition, we mean a declarative statement or sentence. Thus it is a statement declaring a fact that can be truth false, but never both. With an elementary or atomic proposition as opposed to a compound proposition, we mean a proposition with exactly one propositional variable. An example of an elementary proposition would be: "the train is late". This proposition either expresses a true or false statement. A compound proposition is the result of other compound propositions and/or atomic propositions being connected by logical connectives. For example, let p and q be two atomic or compound propositions then if we use an arbitrary logical connective which we will denote by $\bullet$ then the result r, of this compound proposition we denote as $p \bullet q = r$ will also be a compound proposition. The form of the general truth function implies that all non-elementary propositions can be derived using elementary propositions and their negations.

This way of logical thinking by Wittgenstein using truth functions can be used to explain the fundamentals of structured natural language in all of its many forms and guises. Building entities from the ground up using elementary propositions and logical connectives enables accurate, correct, and unambiguous compound propositions which we can apply in many different applications including A-Lex, FBM, and Wordnet. For example if we define that a compound entity E_1 consists of the elementary propositions (which have a SVO structure) $p,\ q$, and r and that it has the structure $p \vee q \implies \neg r$ and we define another compound entity E_2 which consists of the elementary propositions $s,\ t$, and v structured into $\neg s \vee t \wedge v$. Then a new compound entity E_3 could be formed as $E_1 \implies E_2 = (p \vee q \implies \neg r) \implies (\neg s \vee t \wedge v) = E_3$. This clearly shows that E_3 has several (recursively defined) constituent parts. It also shows that its sign is the result of the constituent signs of its constituent parts which accurately and correctly describe the sign of E_3. Furthermore, E_3 is clearly distinguishable from any other possible compound entity because it has a unique expressible sign.

2.10 Conceptual graphs for knowledge representation

In order to accurately represent conceptual schemas using semantic networks and logic, John Sowa came up with the idea of conceptual graphs (see figure 2.19). They have found many applications in artificial intelligence and computing science. Conceptual graphs have contributed to reasoning techniques, knowledge representation, and natural language semantics. The problem described in this thesis, and also of conceptual graphs, lies with the enormous expressive power of natural language, which makes it difficult to capture fully and subsequently to formalise (Sowa, 2008) (Sowa, 1992).

RDF and OWL can be seen as a proper subset of CGIF (Conceptual Graph Interchange Format). All statements in RDF and OWL can be converted into CGIF but not vice versa. The reason for this is that RDF and OWL do not support the full set of the common logic framework, which is a collection of first-order logic languages. Common logic is used for the exchange and transmission of knowledge between computers and has been approved as an ISO standard (Sowa, 2008) (Sowa, 1992). It is a general, highly expressive form of logic. Therefore it is very suitable to translate human languages into a computer language and logic (Welsh, 2018).

Conceptual graphs are very generally applicable. They can be used to model relational database queries in a similar way as FBM (Schadow et al., 2001) (Varga et al., 2010). Based on the possibility of converting natural language into a logical model, conceptual graphs are also used for ontology building (see also subsection 2.1.2) (Bouaud et al., 1994). Consequently, it would be possible to convert ontologies into relational database queries (which can be modelled/converted into FBM), but also back into structured natural language.

To give a practical example of a knowledge graph, we can look more closely at the example in figure 2.19. From this graph, we can then derive the following formula.

$(\exists x)(\exists y)(Go(x) \wedge Person(John) \wedge City(Boston) \wedge Bus(y) \wedge agent(x, John) \wedge Dest(x, Boston) \wedge Inst(x, y))$

The graph and corresponding formula evaluate to the sentence: "John is going to Boston by Bus". Thus as we can see the power of conceptual graphs is the conversion of (structured) natural language to logic and back again. A conceptual graph denotes the conceptual structure of language. The graphs have a highly regular structure on which many search, matching, and reasoning algorithms can be applied (Sowa, 2008).

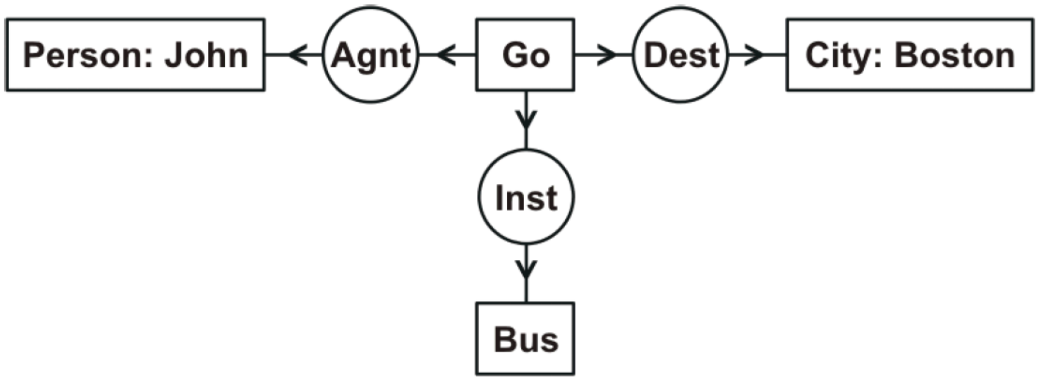

Figure 2.19: Example of a conceptual graph

(adapted from Sowa, 2008 p.2)

2.11 Beyond the picture, an alternative reality with modal reasoning

The essence of multimodal reasoning is that a model can be divided into different sets of laws and facts. For example, Dunn and Kripke both develop modal reasoning semantics which is logically equivalent to one another (Sowa, 2003).

If we take a proposition p then in some possible world/universe, this might be true and in others not. We will call such a world a possible world. We will require two operators namely the $\Box$ for necessity and $\Diamond$ and for possibility[8]. The formula $\Box p$ will be true in a world w if and only if p is true in all worlds accessible from w. The formula $\Diamond p$ will be true in a world w if and only if p is true in at least one world accessible from w. For example, we can have a sentence: "it necessarily true that

[8]The $\Box$ and $\Diamond$ operators are respectively defined as always and sometimes in temporal logic.

a car driving implies that there is a driver". We could define this in modal logic as $\Box(C \implies D)$. Another example would be that: "it is possible to play games without having a keyboard" which we could define in modal logic as $\Diamond(G \wedge \neg K)$ (Geuvers, 2017).

We will define a Kripke model $\mathcal{M} = \langle W, R, V \rangle$ where we have a set of worlds W which is non-empty, a function $R \subseteq W \times W$ for which holds that for all $w \in W$ the resulting set $R(w) \subseteq W$ are the accessible worlds of w. This function is also called the successor of w or accessibility relation. Also, we have a function V for which holds that for each world $w \in W$ the set $V(w)$ denotes the true elementary propositions of w. If we would say that proposition p is true in world w of model $\mathcal{M}$ then we would denote this as $\mathcal{M}, w, \Vdash p$. If this would be the case for the entire model, then we would get $\mathcal{M} \models p$ which also called an universal truth (Geuvers, 2017).

The idea of the multiple worlds with accessibility between one another can be used to explain and further develop the idea of collections in A-Lex. We can define specific formulas for elementary propositions which must hold in all accessible collections. For example, we could say that proposition p must hold in all collections accessible from the current collection and find out if this is indeed true or we could find out what propositions are true in all worlds. We can also use this to determine in what manner different collections are different from one another on a fundamental level. Many interesting questions regarding the diversity and dependencies between domain-specific languages can be answered using a formal theory such as modal logic. More research in different kinds of modal logic and perhaps even temporal logic (to accurately map language evolution) is required, but it is clear that for information system semantics and logical systems much can be gained (Orłowska, 1990).

2.12 On the nature of computability and decision problems

Alan Turing gave us a vital answer to what syntax is and what the properties of a formal language should be. His Turing-machine showed the limits of computability by showing what could be computed and maybe at least as important, what could not be computed. The same was shown around the same time using $\lambda-$calculus by Alonzo Church and Stephen Kleene.

Theories of computational complexity concerning the running time of algorithms (which are formal languages) within certain mathematical lower-bounds (expressed as Ω), upper-bounds (expressed as $\mathcal{O}$), and tight-bounds (expressed as Θ where $\Omega = \mathcal{O}$) have been developed. In addition, came the decision problem classes such as P, NP, PSPACE, EXPTIME, EXPSPACE etcetera[9], which defines how fast (and in higher classes in how much space) specific problems could be solved and checked[10]. Decision problems can be seen as formal languages. So whether or not a decision problem is solvable and, if so, in what complexity is an essential question in order to get an answer. Even though work remains in this field, a precise reduction method of problems has been formulated in order to recognise the problems of a particular class.

If a problem l_1 for example is NP-hard and we can in polynomial time reduce l_1 to l_2 in the form of $l_1 \leq_p l_2$ then we know that $l_2 \in$ NP-hard. The reduce operation $l_1 \leq_p l_2$ means that problem l_1 is at least as hard as problem l_2. Thus a polynomial function $f : \{0,1\}^* \rightarrow \{0,1\}^*$ exists such that $x \in l_1 \iff f(x) \in l_2$. In fact, this formula has the exact same form as the formula we discussed earlier in subsection 2.9.3 concerning the work of Wittgenstein regarding his idea of what a definition is.

Thus if we solve l_1 in polynomial time, then we can solve all NP-hard problems in polynomial time because we can reduce one problem to the other and $\leq_p$ is a transitive polynomial operation, meaning that the concatenation of multiple $\leq_p$ functions can never exceed polynomial time. However, no one has ever managed to solve any of the NP-hard problems in polynomial time.

For a problem to be $\in$ NP (this is a different property of a problem than being NP-hard) then given a particular certificate (solution) for an instance of a problem l we must be able to check in polynomial time[11] whether the solution is correct or not[12].

We can thus see why instances of decision problems are noted as languages which are in this case denoted as a binary number. Interestingly, in section 2.9.3, we talked about what a definition truly is according to Wittgenstein. He defined that a definition is a rule to translate one language into the other. This is precisely what happens with the polynomial f we used to illustrate NP-hard reductions; it translates one formal language into the other. In natural language, we use terms to define (or reduce) other terms. We have shown a set of conditions which must hold for a problem to be NP-hard, $\in$ NP or both. A similar reducing method for defining terms and their required conditions (by propositions) should be defined. If so, then we could understand terms by reducing one to another and setting up clear, unambiguous conditions which must hold for a particular term in a specific

[9]Even though it is generally assumed, for many computational complexity classes, it is not known whether or not they are disjoint. The question whether P = NP holds, for example, is a Millennium Prize Problem, which means that proof of the answer (proof or disproof) is rewarded (besides eternal fame) with a US$1 million prize (Carlson et al., 2006)

[10]In decision problems solving and checking are two distinctive actions. Often checking an answer can be done in a lower computational complexity class than solving a problem.

[11]for higher complexity classes not only time but also space becomes a measure of complexity

[12]NP-complete problems are problems which are $\in$ NP and are NP-hard.

domain. Up to some degree, this is possible using phrasals such as homonyms, synonyms, etcetera but not to a degree as can be done in decision problems. The limit of reducing terms to another is that reducing one term to another is not merely a conversion of input but instead checking what inner and outer properties of the term are the same and different. Only all inner and outer properties of a term together define the true meaning of a term required for comparison, conversion, and translation to another term.

Chapter 3

Results

Our hypothesis was that fundamentally, the levels of descriptions (lexical, relational, and constraint) are related (see the hypothesis in section 1.4 and details in section 3.2). In this chapter, we will discuss the results related to the hypotheses and to the research questions (see section 1.3).

3.1 Using logic to combine (elementary) propositions

We could use compositions of elementary but fully developed propositions and logical connectives between them to make accurate declarative statements. We could be using these compositions for the descriptions in A-Lex, which are meant for explaining a term apart from the relations with other terms to which it is connected and the context in which it is grouped through a collection, in order to make explaining terms less ambiguous. This technique, using only logical connectives, can also be used in the connections between one term and another. The structure of the elementary propositions should be a SVO. On top of this, we should ensure that all objects and subjects used in the elementary propositions are also defined as terms in the current collection or an included collection. This will ensure that the resulting proposition contains nothing but clearly defined terms which can both be an object or a subject in elementary propositions.

Using logical connectives (see tables A.1 and A.2) we can connect the previously discussed elementary propositions and their negations with any of the logical connectives. When defining terms, we can use the logical connection between them to create compound propositions, which in FBM would be called a fact type. We have previously already discussed that in FBM, every n-ary fact type can be decomposed into binary fact types on top of that. Therefore, in combination with the fact that all possible logical operators are at our disposal, the need for parenthesis in order to change the order of precedence of the logical operators is not required.

Some redundancy in elementary propositions is likely because often, a collection of terms is not a tree structure (with this we mean no cycles exist in the term graph) and therefore when reading through the structure of the collection a proposition may be used multiple times. Because natural language and logical connectives are not interchangeable one-to-one, the pronunciations given in tables A.1 and A.2 are only meant as examples. Although these pronunciations indicate the logical connection between elementary propositions p and q, depending on the formulation of the propositions, it might in some cases be more appropriate to use a different natural language expression with the same truth table values.

The goal of FBM, A-Lex, and Wordnet are similar. Even though synonyms, homonyms, and other lexical relations exist (see section 2.1.1), we aim for uncovering the sign of terms and data objects.

3.2 Wittgenstein as a foundation to combine seemingly different approaches

The viewpoints of FBM, Wordnet, and linguistics (see also section 2.8) are quite compatible with Wittgenstein's view. Terms or objects are not used on their own but are a part of a fact/connection. Defining a term or object is done with the help of its properties and facts/connections and thus by using a declarative approach. Objects are also seen as a compound of different structures, and (complex) objects can be defined in terms of their respective constituent parts. This is also how they are distinguished from each other. The variables are the different instances of objects in FBM and the different instances of terms in FBM/Wordnet (for example, an instance of the term 'farmer').

The fact types in FBM but also the relations of terms in A-Lex and Wordnet ensure that objects are combined and the totality of these facts determine the state of the system. Whether this is a database or the state of affairs concerning a domain-specific language is irrelevant. Also, no inductive or deductive reasoning can be used to determine the existence or non-existence of other atomic facts due to their complete independence. The totality of facts only tells us what has been defined in FBM, A-Lex or Wordnet. The picture we have made by defining the facts is the way we have modelled our (domain-specific) language or data model. Whether or not the picture is modelled correctly depends on the reality of the UoD and inherently by the CoD. Only by showing the existence or absence of facts and relations can we determine whether or not the picture is the correct expression of the thought. More concretely, this means that the models in FBM and the collections in A-Lex should express the thoughts of its domains and can only be checked by the domain itself.

Propositions are the bases for facts in FBM but also for relations and descriptions in Wordnet and A-Lex. We will view these propositions in a similar way as Wittgenstein did. The sign will be the structure which represents the sense of the propositions and facts. The propositional variables will be the subject in the SVO (regarding A-Lex) and a predicate in the binary fact type (in FBM). The expression of the proposition is the model structure (e.g. constraints, fact types, and bridge types) in FBM and the general propositional forms and phrasals in A-Lex and Wordnet. Propositions are built on elementary propositions. Understanding just the proposition itself by the domain is inadequate. We require the domain to understand the underlying elementary propositions. For the truth value of the propositions depends on the evaluation of the truth values of the underlying elementary propositions together with the logical connectives between the (elementary) propositions.

In FBM, A-Lex, and Wordnet realities are described using the defined internal properties. The description of the objects is realised by external properties (in the form of fact types and phrasals). The logical scaffolding of propositions (and thus of properties) helps us to define both the objects themselves as well as its relations (fact types, phrasals, etcetera). The meaning of the propositions should be clear by the domain which formulates them. This implies that the understanding what it means and what the consequences are for a model or collection in FBM and A-Lex, respectively if a proposition is true or false should be evident. This is actually the grand plan of the crowdsourcing aspect of A-Lex; to define elementary propositions and combine them by logical scaffolding, because the domain experts are the only people who can adequately express, combine, and understand them. The explanation for this is that the proposition is a picture of the reality as seen in a particular domain. Furthermore, the answer to the question of whether a proposition p evaluates to true or false should always be answered by the domain. The answer to these questions regarding the evaluations of propositions helps us to see whether the attributes and features of a proposition and object are true in the domain-specific reality and help us to compare multiple models/collections to one another (we have discussed this in 2.11).

The amount of fundamental overlap between the different approaches is clear. The article we published (Nobel et al., 2020) relates to this, we discussed (lexical) word meaning, data semantics, Business Knowledge Mapping (BKM)[1], cybernetical language evolution, and crowdsourcing in a data management context. In the article, we were wondering if there was a general theory in existence which would help us to link the different approaches together (Nobel et al., 2020). Again their applications might differ, but on the fundamental level, they are not so disjoint as one might think. In the article, we used table 3.1 in order to present the different approaches

[1]In (Nobel et al., 2020) Business Knowledge Mapping is meant as the generalised approach of which A-Lex is a part of.

and their underlying practical implementations. At higher levels of abstraction, the differences between the implementations fade away, and their theoretical foundations can be explained using the theory of Wittgenstein, Kripke, and many unnamed others. At a higher level of abstraction the related lexical, relational, and constraint view can be explained using modal and predicate calculus[2] (Sowa, 2003).

	Wordnet	**Semantic Web**	**BKM/A-Lex**	**FBM**
Rules & axioms		OWL & axioms		Constraints
Predicates		OWL & RDF		Fact types
Lexical definitions	Network & glosses	Network (Is-a) & URI+glosses	Network & glosses	Network (Is-a object types) & glosses

Table 3.1: Approaches and their practical implementations

(adapted from Nobel et al., 2020)

As has been discussed in section 2.10, it is possible to convert relational database structures into conceptual graphs and vice versa. This is especially relevant because this implies that we can convert structured natural language originating from self-service/crowdsourcing into conceptual graphs and subsequently into relational database queries which we can convert into FBM. In addition, we have shown in section 2.5 that the formalisations of FBM can be converted into structured natural language. This implies that we have begun to close the gap between the semantical, relational, and constraint levels as has been discussed in (Nobel et al., 2020). In fact, we could even extend this into the wider domain of ontology structuring/modelling and its taxonomic structure (Schadow et al., 2001) (Varga et al., 2010) (Bouaud et al., 1994).

It is insightful to think about chapter 2.12 with respect to table 3.1. The basis for λ-calculus lies in recursive functions; Wittgenstein also suggests that complexes are constituted from smaller constituent parts which recursively describe the complete complexes. Therefore, in some sense, Wittgenstein has tried the same with (structured) natural language in combination with logic. We tried to use Wittgenstein's and Kripke's theory to combine FBM and linguistic definitions of terms.

Even though we have discussed some essential results, we have not eliminated all possible limitations of our findings and the proposed approach. The progress

[2]John Sowa has developed his conceptual graphs based on predicate calculus (Sowa, 2003)

in mathematics, computing science, and formal languages (including automata) has helped to outline what is computable (and in what complexity order) and what is not. Similarly, limitations regarding the interpretation of certain propositions still exist. Apart from thoughts that cannot adequately be expressed, the proposed form of SVO may not always fit the objective and unambiguous requirements, as can be explained with Wittgenstein's view concerning the differences between natural science and philosophy (see subsection 2.9.3). As a result, the entire collection of terms and, recursively all collections which have dependencies, may endure a divergence rather than a convergence towards solid domain-specific definitions. The problem for data models in FBM, homonyms in Wordnet, and phrasals in A-Lex alike, regarding the structure of interdependencies of terms, and similarly, with collections/worlds/universes can be seen as a house of cards. Once a card is removed or falls out, the cards above will instantaneously follow.

A language which has the same expressive power as a Turing machine is called Turing-complete. In practice this means that everything which is computable could be computed using this language (Kepser, 2004). We have previously argued that Wittgenstein's theories concerning (structured) natural language are complementary to Turing's theories of computation. We would, therefore, like to introduce a new definition called: 'Wittgenstein completeness' in order to uphold the essentials of Wittgenstein's accurate symbolism and thus a way to correctly, accurately, and unambiguously describe terms in (structured) natural languages. Therefore, if we want a model or state of affairs to be Wittgenstein complete, then the previously mentioned house of cards problem should be considered very seriously to prevent any contradictions in the model. For all practical purposes, it would not always be required and practical for a model to be Wittgenstein complete unless absolute correctness and accuracy are required.

3.3 Cybernetics as an argument for sentient language evolution and communication

Cybernetics is a broadly applicable theory. In this section, we aim to bring the previously described diverse theories together and try to explain a possible new cybernetic point of view.

3.3.1 Combining different theories using cybernetics

Section 2.6 can be seen as an argument for the dynamic characteristics of domain-specific languages. Together with Wittgenstein's work, we now know that natural language does not only evolve but also that its comprehension by others can be problematic. With problematic, we mean true intent (or sign) of a term/proposition might not correspond to the understanding of the receiving party. Put differently, the information is not equivalent to the understanding of the communication (see section 2.6.3). One could, of course, quote Wittgenstein here and say that: "whereof one cannot speak, thereof one must be silent". The famous quote of Wittgenstein means that some thoughts can not be communicated adequately at all. In some instances, it is extremely difficult or even impossible to know precisely what the other person means. The meaning of someone's thought (the expression of a proposition with sense) becomes ambiguous the moment it is communicated. Even simple deductions based on observations do not give a dependable proof of someone's exact meaning. However, in many cases, it would probably be a good idea to decrease the ambiguity in the utterance of the communication (Wittgenstein, 1922).

Another interesting factor of cybernetics concerning the theories of Wittgenstein and modal logic concerns the fact that in cybernetics a system at a certain point in time has a specific state of affairs. This can be compared to the state of the universe at a particular moment in time or as a specific world in modal logic[3]. Its state of affairs continuously changes, again cybernetics does not concern itself with 'things' but instead with behaviour. It aims to control systems, and its focus lies on complexity. Its state of affairs concerns the whole of elements and its relations between them, showing some behaviour.

In section 2.4, we have discussed the origin of the meaning of terms. In classical stipulative definitions, a decision has been made on the use and definition of a term. This is on the opposite of how Wordnet and A-Lex are working. They both use lexical definitions which can be true or false in a particular context. A lexical definition is often used to reduce ambiguity and explain the use of the term (Copi et al., 2007, p. 89). The decisions in A-Lex on how terms are, the evolution and decisions made by the domain who defined them. In Wordnet, however, they define terms based on the lexical use which is more difficult to define as decisions, as they use more general definitions based on the general use of the term, and can probably be better explained using the cybernetical theories concerning general language evolution. In other words, they define terms based on sufficient general agreement concerning the semantics.

Our view on cybernetics is in line with the view that natural language is shaped by the way we conceptualise the world. In fact, in the literature, we can find strong indications that the meaning of terms has been agreed upon by the members of the domain. This, therefore, highlights the importance of agreement on domain-

[3]Temporal logic seems more suitable when talking about the state of affairs at specific moments in time. However, temporal logic is a specific form of modal logic and within the scope of this thesis we will not elaborate further on the applicability of temporal logic versus modal logic with respect to cybernetics, FBM, Wordnet, and, A-Lex.

specific semantics in organisations. It suggests that feedback concerning the success of communication is of great importance (De Beule & Vylder, 2005). Language should be a self-organising system. Also, we require an agreement concerning what information is communicated and how (the Information and Utterance parts we discussed earlier in section 2.6). This thus acknowledges our earlier statement that language is a shared and distributed cognitive and cultural phenomenon (Nobel et al., 2020) (De Beule, 2006).

3.3.2 A new cybernetical point of view

Luhmann viewed organisations as a particular type of social systems where the decisions are the elements of the system, and a number of specific decision premises as part of the supporting structure of the system exist. While as discussed earlier this is an interesting theory regarding domain-specific language evolution in organisations from an organisational theoretical point of view, we also want to show an alternative perspective which purpose is to give a more fundamental point of view. More concretely, this means that the behaviour of the system concerns the elements, which are terms consisting of logically chained elementary propositions or (negated) elementary propositions, interlinked with each other by relations (phrasals and other semantic relationships). The system, of course, cannot exist without elements, and the combination of elements must create new elements. In other words, the combination of terms into propositions using the SVO structure and the subsequent newly linked propositions, create new elements in the system.

The old elements might not disappear, but if the element is more than just an elementary proposition defining a term, then it changes in the sense that it has been given an additional context and is therefore not wholly equivalent to its previous variant. This might seem like a minor or insignificant difference. Nevertheless, it is now used in an additional context (and therefore in some proposition or phrasal) and its sign is not entirely the same due to a change of the meaning (actually signs) of the terms constituent parts. The totality of the signs combined is again a sign[4]. We do make the assumption here that part of the constituent part of an (elementary) proposition is its context. It is thus an autopoietic system in the sense that it creates new elements based on its smallest non-divisible elements.

This new view on how to place domain-specific language evolution in the context of cybernetics might not be developed according to general cybernetic requirements, but it might be more informative concerning the explanation of domain-specific language evolution on the fundamental level as we have described it earlier.

[4]Although not we did not discuss this in the thesis. This statement fits within the critical concept of recursion, which is essential in cybernetics. This concept can also also be applied to signs. The top sign is recursively defined by its constituent signs.

3.4 Applying a sturdy logical foundation combined with a cybernetical view and the power of the web 2.0

The previously stated ideas concerning building terms from (elementary propositions) and phrasals combined with the need to model the domain-specific language evolution using the power of the web 2.0 can be combined. As has been stated before in section 2.10, the idea of dissecting propositions has been suggested before by John Sowa. In section 2.9, we have shown the theory of Wittgenstein concerning the foundations of terms and propositions. In addition to his theory, we have talked about the idea of modal logic and Kripke semantics in section 2.11. In section 2.7 we have talked about crowdsourcing, self-service modelling, and the power of the web and about first-order cybernetics and cybernetics in organisations in section 2.6. We have talked about the semantic web in section 2.2, about FBM in section 2.5, and about natural language, semantic relations, and phrasals in section 2.1. Lastly, in section 2.12, we discussed the link between computability (in particular decision problems) and their link with respect to our view and way of dealing with (structured) natural language and term definitions.

In this thesis, with its broad range of discussed theories, we aim to build a bridge between the theories of Wittgenstein's masterpiece, modal logic in the form of Kripke Semantics, John Sowa's work on conceptual graphs, first- and second-order cybernetics, crowdsourcing, the semantic web, and the web 2.0. Using user content creation capabilities (which has been made possible by the web 2.0) by the domain itself (instead of using IT experts), many new things are possible because the domain-specific knowledge/language can now directly be applied and modelled. If we combine these possibilities with the theories of Wittgenstein and the work of John Sowa the prospect of correctly and continuously modelling domain-specific languages, by the domain itself, with the possibility of learning from other domains and therefore reducing the divergence of the languages might become a reality.

Chapter 4

Conclusion

The first research question was: *How does A-Lex relate to fields like lexical databases (Wordnet), the semantic web, and Fact-Based Modelling (FBM)?*. We have related A-Lex with Wordnet and FBM by showing that on the fundamental level, the different approaches (lexical, semantical, and constraint) are actually quite similar to each other. On the fundamental level, correctly defining words or correctly modelling a database model is not different at all. Both use the SVO structure to link multiple entities to each other using a fact type. The totality of the 'world' or language are the facts. Even though the applications of the different approaches differ substantially, several parts of these approaches are actually interchangeable due to their similar foundations. The work of Wittgenstein has contributed greatly to this insight. His work on the nature of (structured) natural language has proven to be very insightful.

The second research question was: *How can self-service, community-based modelling be related to theory, integrated with the answer of question 1?*. We have discussed the need to involve the public in defining the semantics. This also means that in order to have all the required knowledge to define domain-specific languages, we require the complete involvement of the domain. This means that we require the domain itself to build the domain-specific language model. The power of the web 2.0 and its crowdsourcing capabilities are enormous, especially using the technologies of the semantic web (such as RDF). Wikipedia and several other internet encyclopaedias have extended their lexicon exponentially due to the user's content-creating capabilities brought forth by the web 2.0. For an application such as A-Lex, the power of the web 2.0 in combination with the proposed way of defining terms in this thesis is a great combination to show the possibilities of adequately defining the ever-changing domain-specific languages. The semantic web and its related technologies might prove an inspiration for this.

The third research question was: *How can we sketch an integrating formalisation of theory, to bridge the gap between the use of (structured) natural language in A-Lex,* *FBM, and logic?*. We have combined a seemingly diverse set of concepts and theories. The semantic web and the conceptual graphs of John Sowa (which are based on RDF) have in no small extent ensured for proper communications and taken into account a large part of the problems which arose from the UoD's and CoD's. The problems of dealing with the lexical semantics can be explained using cybernetics, symbolism (including Wittgenstein), and Kripke. In addition, the formalisation of FBM and other (modelling) techniques can be used to assist the problem of accurate symbolism in information systems and domain-specific languages. More concretely, due to the common logic on which conceptual graphs are based, it is possible for them to be converted back into structured natural language, relational database queries (thus also into FBM), but also into ontologies. We have discussed a number of formal relations from FBM in section 2.5. These have been fully formalised and can be converted into structured natural language. These discussed examples thus can be used to further formalise certain other relations between objects/terms. In addition, it can be seen as an example of our proposal concerning how logic can assist in unambiguous structured natural language statements and how it can help to bridge the gap between the lexical, relational, and constraints views.

We have shown with cybernetics and the theory of Wittgenstein that domain-specific languages are not a static entity but rather a diverse and incessantly changing dynamic entity prone to changes in its complex environment. The different universes (or worlds in Wittgenstein and Kripke's words) and communities of discourse are well-known and maybe well-defined for a few people from their domain, but because the domain-specific language is not formally defined on paper or with a tool, correct and understandable communications for all parties involved remains in many cases quite difficult. The notions of uniquely expressing signs are rooted in the theory of Wittgenstein (see section 2.9.2) and have an essential application in semiotics (see subsection 2.1.4). In information retrieval, for example, the sign of a term from a query is crucial to retrieve the requested and expected information. Wittgenstein has shown us what could be expressed. He showed us what sense is, but also, and perhaps even just as important, what senseless, and without sense is. He did this with the use of explaining symbolism, expressions, facts, pictures, and not to forget (elementary) propositions. Wittgenstein tells us that the totality of propositions is the language, while Turing has defined what syntax exactly is by showing us what was accepted and decidable (computable) using a Turing machine. This was the main inspiration why we came up with a new definition called: 'Wittgenstein completeness' which denotes the absolute correctness and accuracy of structured natural language definitions without any contradictory statements.

In subsection 2.6.2 we referred to (Nobel et al., 2020) where is stated that the issue of dealing with lexical semantics has not yet been completely solved. This may be true, but in a certain sense, Wittgenstein has shown us what lexical semantics exactly are. What remains is how we deal with this complex issue. If we look back at the semiotic ladder in figure 2.3, then the first two steps lay respectively out of human influence and the scope of this thesis. However, the third and fourth step are defined by Turing and Wittgenstein, respectively. The last two steps are part of the cybernetic evolution of language and all that comes with the broad and complex disciplines of first- and second-order cybernetics complemented with Kripke semantics. Therefore, in a sense, the theories of Turing, Wittgenstein, Cybernetics, and Kripke are complementary to one another and explain all steps of the semiotic ladder except the first two.

Chapter 5

Discussion

In this chapter, we will talk about the limitations of the proposed solutions and how future research might omit these (at least partly). Possible ideas for future research also include subjects which have not been addressed extensively enough in this thesis and could be the subject for future research.

5.1 Limitations

In this thesis, we have combined a lot of seemingly different (and formal) theories. We do not claim to give a solid and definitive proof of anything. We aimed solely to accomplish sufficient theoretical depth and explanation to provide a well-founded new hypothetical and theoretical view.

5.2 Improvements and future research

5.2.1 Boolean retrieval models and 2d search

PubMed[1] is an example of a search engine which uses a boolean retrieval model. Thus it combines terms with operators such as NOT, AND, and OR (Lu et al., 2009). The limitation of this approach, however, is that many people are not familiar with logical connectives and their truth tables. Thus concerning usability and user-friendliness connecting propositions using logical operators and perhaps even using existential and universal quantifiers are not the best options. However, the precision that can be achieved is very high. The main advantage regarding boolean retrieval models is its high efficiency, accuracy (if used correctly), predictability, and easiness of explanation (especially regarding the search results). This is especially true when

[1]Pubmed is a search engine specific for medical articles and research.

assuming that all query terms are equivalent in terms of relevance and that term relevance is binary (Croft et al., 2009).

We suggest that in future research, the work of Tony Russell-Rose concerning 2dsearch is considered. He developed a tool and conducted research regarding how we can let users precisely define and combine search terms using boolean algebra but in a more user-friendly approach. It was initially intended for information retrieval queries, but we believe that this can also be applied to build domain-specific definitions of terms (Russell-Rose & Gooch, 2018).

5.2.2 Web 2.0

A crucial problem remains how we can use the power of the web 2.0 and the public to use the structured way of propositions and logical connectives to define terms of domain-specific languages. Although we talked about the use of the web 2.0, a clear plan should be made on how this concept can be used to its fullest extent to the benefit of A-Lex. The implementation of the semantic web and its technologies might be an inspiration.

5.2.3 Conceptual graphs

The work of John Sowa could be used for the theoretical framework, especially his work on RDF, knowledge representations/ontologies, conceptual graphs, and logic. He has written many books and articles concerning logically capturing and modelling (structured) natural language (see section 2.10). His extensive work might prove useful for improving and extending the proposed solutions.

5.2.4 Other related work

Wittgenstein produced two significant works in his life. The first one, and the only one published during his lifetime was the Tractatus Logico-Philosophicus which we have extensively discussed in this thesis. The second one is called Philosophical Investigations and talks about language games. These games are used to explain the use of natural language in everyday life. This work might also be beneficial to understand how language is used to define the semantics of natural languages. Also, it might be a good idea to see to what extent Wittgenstein's second work is compatible with the idea of Luhmann regarding expectations about expectations (see subsection 2.6.3).

Other leading figures which might prove to be an inspiration for future work might be Putnam, Russel, and Leibnitz but also a more general concept such as the philosophical methodology called Ordinary Language Philosophy (to which Wittgenstein also belonged).

Bibliography

Bouaud, J., Bachimont, B., Charlet, J., & Zweigenbaum, P. (1994). Acquisition and structuring of an ontology within conceptual graphs. In *Proceedings of iccs* (Vol. 94, pp. 1–25).

Broekstra, J., Kampman, A., & Harmelen, F. (2002, 06). Sesame: A generic architecture for storing and querying rdf and rdf schema. In (Vol. 2342, p. 54-68).

Burgin, M. S. (1990, Jul 01). Generalized kolmogorov complexity and other dual complexity measures. *Cybernetics*, *26*(4), 481–490. Retrieved from `https://doi.org/10.1007/BF01068189` doi: 10.1007/BF01068189

Carlson, J. A., Jaffe, A., & Wiles, A. (2006). *The millennium prize problems*. American Mathematical Society Providence, RI.

Code, G. (n.d.). *synonym.* Retrieved 2019-07-15, from `https://code.google.com/archive/p/synonym/`

Copi, I., Cohen, C., & MacMahon, K. (2007). *Introduction to logic, 2011, boston, new york, etc.* Pearson.

Croft, W. B., Metzler, D., & Strohman, T. (2009). Search engines - information retrieval in practice..

De Beule, J. (2006). Simulating the syntax and semantics of linguistic constructions about time. In *Evolutionary epistemology, language and culture* (pp. 407–428). Springer.

De Beule, J., & Stadler, K. (2014). An evolutionary cybernetics perspective on language and coordination. *New Ideas in Psychology*, *32*(Complete), 118-130. doi: 10.1016/j.newideapsych.2013.03.003

De Beule, J., & Vylder, B. (2005, 01). Does language shape the way we conceptualize the world? In (p. 335-336).

Definition. (n.d.). *English Oxford Living Dictionaries.* Oxford English Dictionary.

Devine, S. D. (2017). *Algorithmic information theory: review for physicists and natural scientists.* Victoria University of Wellington.

Ferreira, J., Noble, J., & Biddle, R. (2007, 01). The semiotics of usage-centred design. *Project Management and Risk Management in Complex Projects: Studies in Organisational Semiotics*, 211-229.

Gabriel Calise, S. (2016, 05). Rethinking niklas luhmann's theory of the phylogenesis and ontogenesis of language in light of systemic-functional linguistics. *Revija za sociologiju*, *45*, 223-248. doi: 10.5613/rzs.45.3.1

Geuvers, H. (2017, 11). *Formal Reasoning.* Retrieved from `https://www.cs.ru.nl/ freek/courses/fr-2017/`

Goodchild, M. (2007, 01). Citizen as voluntary sensors: Spatial data infrastructure in the world of web 2.0. *International Journal of Spatial Data Infrastructures Research*, *2*, 24-32.

Greshoff, R. (1999). *Die theoretischen konzeptionen des sozialen von max weber und niklas luhmann im vergleich.* doi: 10.1007/978-3-322-97095-4

Gruber, T. (2007, 01). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems*, *3*, 1-11.

Halpin, T. (2009). *Object-role modeling.* Neumont University, USA. Retrieved from `http://www.orm.net/pdf/EncDBS.pdf`

Hesse, W., & A. Verrijn-Stuart, A. (2000, 01). Towards a theory of information systems: The frisco approach. In (p. 81-91).

Hildebrandt, M. (2019, 6 2). 2. law, democracy, and the rule of law. In *Law for computer scientists.* Retrieved from `https://lawforcomputerscientists.pubpub.org/pub/hx1s8lbt` (https://lawforcomputerscientists.pubpub.org/pub/hx1s8lbt)

Hoppenbrouwers, S. (2018). *Formal Languages, Natural Languages, and the Web.* Retrieved from `Obtained from Radboud University, course NWI-IMK012`

Hoppenbrouwers, S. J. B. A. (2003). *Freezing language, Conceptualisation Processes across ICT-Supported Organisations.* Retrieved from `https://ru.on.worldcat.org/oclc/6893058293`

Hébert, L. (2018). Elements of semiotics.. Retrieved from `http://www.signosemio.com/elements-of-semiotics.asp` (Online-Resource)

Insfran, E. (2003, 10). A requirements engineering approach for object-oriented conceptual modelling..

Jambhulkar, S. V., & Karale, S. J. (2016, Nov). Semantic web application generation using protégé tool. In *2016 online international conference on green engineering and technologies (ic-get)* (p. 1-5). doi: 10.1109/GET.2016.7916686

Jayawardana, V., & Jayawardana, V. (2017, Mar). Ontology generation and visualization with protégé. Medium. Retrieved from `https://medium.com/@vindulajayawardana/ontology-generation-and-visualization-with-protégé-6df0af9955e0`

Jurafsky, D., & Martin, J. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recogni-*

tion (Vol. 2).
Kepser, S. (2004). A simple proof for the turing-completeness of xslt and xquery. In *Extreme markup languages®*.
Khatri, V., & Brown, C. V. (2010, January). Designing data governance. *Commun. ACM*, *53*(1), 148–152. Retrieved from http://doi.acm.org.ru.idm.oclc.org/10.1145/1629175.1629210 doi: 10.1145/1629175.1629210
King, J. B. (1993, Feb 01). Learning to solve the right problems: The case of nuclear power in america. *Journal of Business Ethics*, *12*(2), 105–116. Retrieved from https://doi.org/10.1007/BF00871930 doi: 10.1007/BF00871930
Lannan Gallery. (2019). *Early 20th-century ship's telegraph.* Retrieved from https://lannangallery.com/products/early-20th-century-ships-telegraph ([Online; accessed June 22, 2019])
Lu, Z., Kim, W., & Wilbur, W. J. (2009, Feb 01). Evaluation of query expansion using mesh in pubmed. *Information Retrieval*, *12*(1), 69–80. Retrieved from https://doi.org/10.1007/s10791-008-9074-8 doi: 10.1007/s10791-008-9074-8
meet-alex. (2017). Retrieved 2019-10-08, from http://www.meet-alex.org/
Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, *6*(1), 1–12.
Miller, B., Ranum, D., Elkner, J., Wentworth, P., Downey, A., Meyers, C., & Mitchell, D. (2014). How to think like a computer scientist: Interactive edition. *Runestone,*.
Miller, E. (1998). An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, *25*(1), 15-19. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/bult.105 doi: 10.1002/bult.105
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990, 12). Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, *3*(4), 235-244. Retrieved from https://doi.org/10.1093/ijl/3.4.235 doi: 10.1093/ijl/3.4.235
Miller, G. A., & Hristea, F. (2006). Wordnet nouns: Classes and instances. *Computational Linguistics*, *32*(1), 1-3. Retrieved from https://doi.org/10.1162/coli.2006.32.1.1 doi: 10.1162/coli.2006.32.1.1
M. I. M. Achterbergh, J., & J. Vriens, D. (2009). *Organisations: Social systems conducting experiments.* doi: 10.1007/978-3-642-00110-9
Murugesan, S. (2007, 08). Understanding web 2.0. *IT Professional*, *9*, 34 - 41. doi:

10.1109/MITP.2007.78
Muñoz-Leiva, F., Hernández-Méndez, J., & Sánchez-Fernández, J. (2012, 11). Generalising user behaviour in online travel sites through the travel 2.0 website acceptance model. *Online Information Review*, *36*, 879-902. doi: 10.1108/14684521211287945
Nobel, T., Hoppenbrouwers, S., Pleijsant, J. M., & Ouborg, M. (2020). Word meaning, data semantics, crowd sourcing, and the bkm/a-lex approach. , To appear in: Christophe Debruyne, Hervé Panetto et al., On the Move to Meaningful Internet Systems: OTM 2019 Workshops. Confederated International Workshops, Rhodes, October 2019, Revised Selected Papers.
Orłowska, E. (1990, Jun 01). Kripke semantics for knowledge representation logics. *Studia Logica*, *49*(2), 255–272. Retrieved from `https://doi.org/10.1007/BF00935602` doi: 10.1007/BF00935602
Pleijsant, J. M. (2019). Scale your information modelling effort using the power of the crowd. In C. Debruyne, H. Panetto, W. Guédria, P. Bollen, I. Ciuciu, & R. Meersman (Eds.), *On the move to meaningful internet systems: Otm 2018 workshops* (pp. 159–170). Cham: Springer International Publishing.
Princeton-University. (2019). What is wordnet?. Retrieved from `https://wordnet.princeton.edu/` (Online-Resource)
Russell-Rose, T., & Gooch, P. (2018). 2dsearch: A visual approach to search strategy formulation. In *Desires.*
Schadow, G., Barnes, M., & Mcdonald, C. (2001, 02). Representing and querying conceptual graphs with relational database management systems is possible. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 598-602.
Sowa, J. F. (1992). Conceptual graphs summary. *Conceptual Structures: current research and practice*, *3*, 66.
Sowa, J. F. (2003). Laws, facts, and contexts: Foundations for multimodal reasoning..
Sowa, J. F. (2008). Conceptual graphs. *Foundations of Artificial Intelligence*, *3*, 213–237.
Steels, L. (2013, 03). Self-organization and selection in cultural language evolution. doi: 10.1075/ais.3.02ste
Surowiecki, J. (2004, 01). The wisdom of crowds: Why the many are smarter than the few. london. *Abacus*, 295.
Sutton, T. (2015). *Creating a similarity graph from wordnet.* Retrieved 2019-07-15, from `https://slideplayer.com/slide/4454912/`
Teixeira de Melo, A. (2015, 11). From 1st and 2nd order cybernetics to complexity: Building grounded theories of family complexity in family therapy process research. *Melo, A. T. & Alarcão, M. (2015). From 1st and 2nd order cybernetics to Complexity: Building grounded theories of family complexity in family ther-*

apy process research. Human Systems: The Journal of Therapy, Consultation and Training, 26(1), 86-102..

ter Hofstede, A. (1993). *Information modelling in data intensive domains.* Retrieved from `https://ru.on.worldcat.org/oclc/66247580`

ter Hofstede, A., & van Bommel, P. (2018, Nov). *Formal syntax and semantics of advanced information models.* Retrieved from `Obtained from Radboud University, course NWI-IBC020`

van Bommel, P. (2019). *Foundations of information systems.* Retrieved from `Obtained from Radboud University, course NWI-I00035`

Varga, V. Z., Sacarea, C., & Takacs, A. (2010). Conceptual graphs based representation and querying of databases. *2010 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)*, *3*, 1-6.

vor der Brück, T. (2010, 01). Learning deep semantic patterns for hypernymy extraction following the minimum description length principle..

Welsh, R. (2018, July 18). *Current conversations in ai: John sowa.* Retrieved from `https://www.youtube.com/watch?v=YLimd97OT1Q`

Wittgenstein, L. (1922). *Tractatus logico-philosophicus.*

Appendix A

Truth tables of logical connectives

The two tables below are meant to illustrate structured natural language expressions of logical connectives. It should be noted that many other ways of expressing the logical connectives in structured natural language exist.

p	q	$\overline{\vee}$	$\not\leftarrow$	$\not\rightarrow$	$\veebar$	$\overline{\wedge}$
T	T	F	F	F	F	F
T	F	F	F	T	T	T
F	T	F	T	F	T	T
F	F	T	F	F	F	T

Logic gate	NOR		NIMPLY	XOR	NAND
Expression	neither...nor	not...but	but not	either, but not both	not both

Table A.1: Logical connectives, gates, and example expressions part 1

p	q	$\wedge$	$\leftrightarrow$	$\rightarrow$	$\leftarrow$	$\vee$
T	T	T	T	T	T	T
T	F	F	F	F	T	T
F	T	F	F	T	F	T
F	F	F	T	T	T	F

Logic gate	AND	XNOR	IMPLY		OR
Expression	and	if and only if	implies	is implied by	or/and

Table A.2: Logical connectives, gates, and example expressions part 2