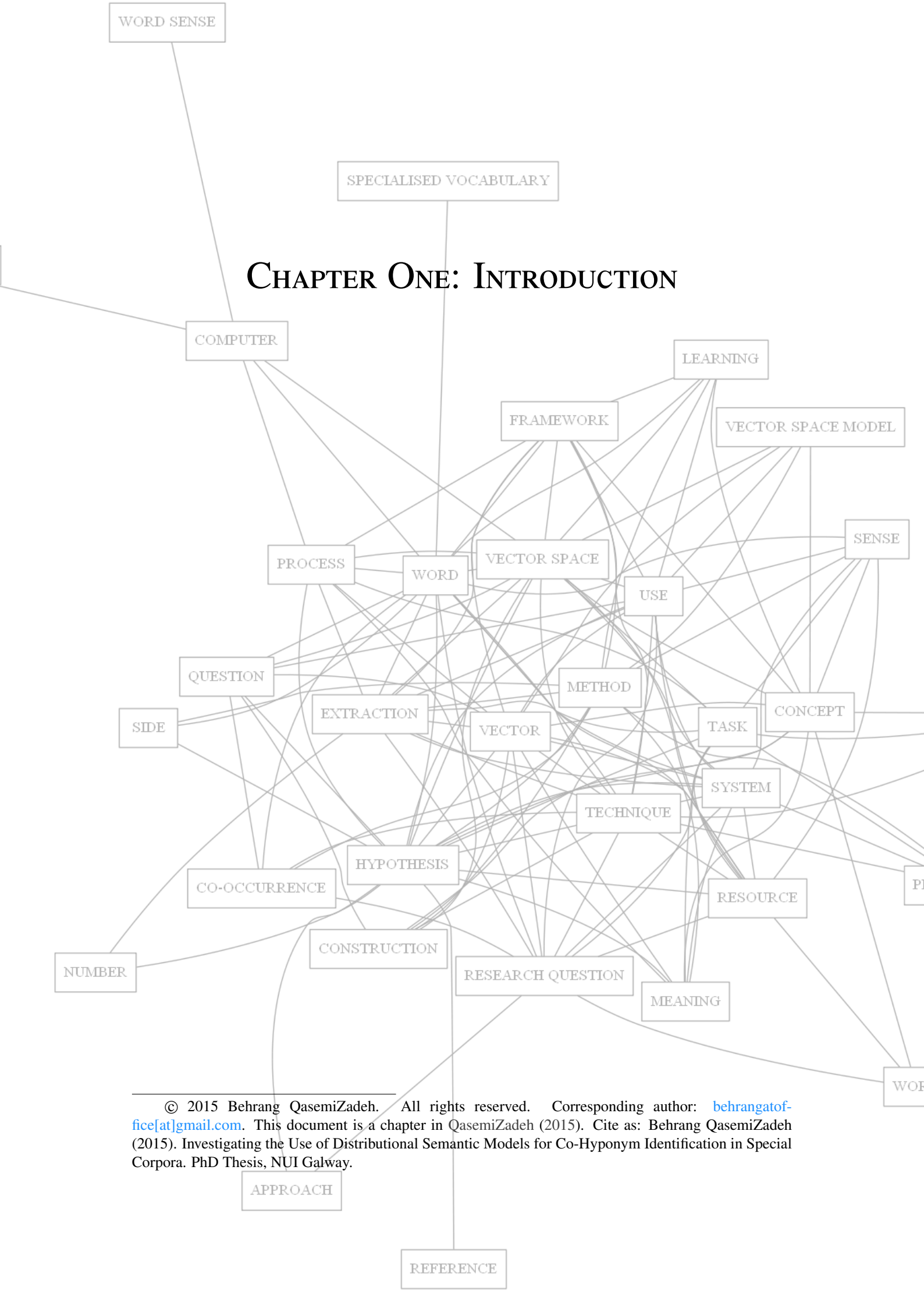

Chapter One: Introduction

 Corresponding author: behrangatoffice[at]gmail.com. This document is a chapter in QasemiZadeh (2015). Cite as: Behrang QasemiZadeh (2015). Investigating the Use of Distributional Semantic Models for Co-Hyponym Identification in Special Corpora. PhD Thesis, NUI Galway.

This page is intentionally left blank.

Contents

This page is intentionally left blank.

List of Figures

This page is intentionally left blank.

Chapter 1

Introduction

1.1 Motivation

Directly accessing human thoughts and transferring the knowledge they possess to machines is still far beyond the reach of technology.[1] Language—and thus text—is still the main vehicle for knowledge dissemination. An ever-increasing amount of text data in our digital era manifests the fluid nature of knowledge and its rapid growth. However, capturing knowledge from text and representing it in a machine-accessible format is a tedious and time-consuming problem. Since the early days of commercial computers, this has resulted in difficulties in developing knowledge-based systems—as is still best described by the term *knowledge acquisition bottleneck* coined by Feigenbaum (1980).

Automated text analysis techniques have thus been developed to facilitate the process of *knowledge acquisition* from text and to improve the productivity of knowledge workers.[2] Evidently, the development of these methods has evolved into several multidisciplinary research areas. In these research, the study of knowledge and its relationship to language is a common theme. *Concepts* are often seen as the constituents of knowledge; disputes about their nature, structure, and relationship to language and linguistic communication, however, have led to different ways of formulating research questions in these studies.[3] Disregarding these differences, the essence of the problem has remained the same: bridging the *semantic gap* between text and machine-accessible knowledge structures (see Brewster, 2008, chap. 2 for a thorough perspective).

In the study of language structure and its relationship with knowledge, much attention has been paid to lexical units known as *terms*. Human knowledge is an expression of a plurality of domains of knowledge. In each domain, terms constitute a *specialised vocabulary* to communicate knowledge.[4] Since concepts are abstract mental objects that cannot be sensed, terms are often seen as labels to access salient concepts in a domain knowledge (L'Homme and Bernier-Colborne, 2012). As a result, identifying terms and constructing terminological resources can be considered as a stepping-stone for constructing domain-specific knowledge bases. For instance, Brewster et al. (2009) suggest that identifying terms is the *key step* for building a *domain ontology*. The discipline of *terminology*, and its sub-discipline computational terminology, has developed as a result of the systematic study of terms (see Chapter 3).

Specialised vocabularies are invented mainly to reduce lexical ambiguity. General language words are inherently vague due to their envisaged function in natural language communication systems—that is, a *finite* set of words are used to communicate *innumerable* concepts.[5] To alleviate ambiguity in the process of knowledge dissemination (e.g., technical and scientific writing), special attention is paid to *lexical cohesion* (e.g., as em-

[1] Such as depicted in *Star Trek* by the *Vulcan mind meld* and the *Marijne VII beings* communication ability; however, a similar technology is not yet available to the *computer access and retrieval system* in the 29th century (Roddenberry, n.d.).

[2] Or, breaking the knowledge acquisition bottleneck, as put by the artificial intelligence community.

[3] See Margolis and Laurence (2014), for a gentle philosophical explanation.

[4] This perspective is maintained throughout this thesis. Hence, in this thesis, it is assumed that the interpretation of the meanings of a term is bounded to a particular domain knowledge.

[5] The ambiguity of words is not limited to *polysemy*; see Murphy (2002, chap. 11, p. 404) for an elaboration of the meaning of the word *vague* in this context.

phasised in *technical writing pedagogy*).[1] In achieving this goal (i.e., lexical cohesion) and to ensure precision in communication, the invention of terms for reducing lexical ambiguity is a dominant mechanism employed in technical writing.[2]

In this process, the collection of documents that represents a domain knowledge, as a whole, constitutes the discourse in which meanings of terms are interpreted.[3] As such, lexical cohesion is established over the corpus and not individual documents or text segments.[4] Empirical studies in natural language processing—particularly, *word sense disambiguation*—support this argument. Results obtained based on generalisations of the so-called *one sense per discourse* (OSD) hypothesis by Gale et al. (1992) are well-known examples.[5] Accordingly, Martinez and Agirre (2000) show that the OSD hypothesis is strongly held in corpora that share a related genre or topic. Similarly, enhances in the performance of word sense disambiguation algorithms as a result of domain-adaptation are also evidence that support the proposed argument (e.g., see Chan and Ng, 2007).

In computational terminology, *automatic term recognition* (ATR) techniques are often at the centre of attention. ATR techniques are developed as an (assistive) tool for extracting terms from text and maintaining up-to-date inventories of specialised vocabularies. ATR algorithms do not specify semantic relationships between terms. The input of ATR is often a *domain-specific corpus*,[6] and the output is an unstructured set of terms. These terms signify a broad spectrum of concepts from the domain knowledge that they represent. However, in many applications (e.g., in *ontology-based information systems*[7]), the extracted terms are required to be organised to meet demands or to enhance performances of information systems. An analogy of this convention is the method employed in the *Princeton WordNet* lexical database (Fellbaum, 1998) for organising words.

WordNet distinguishes between *word* and *concept*: a *word* is a lexical form of a *concept* (or *meaning*). The relationship between words and concepts is assumed to be many-to-many. Hence, *synonymy* is one of the main relationships employed to organise words.[8] In WordNet, words that refer to the same concept are synonymous and organised as one *synset* (Miller et al., 1990). In turn, the synonym relation between words and constructing synsets can be seen as the mechanism employed to denote concepts.[9] In contrast, Miller et al. define another set of relationships between 'word meanings' (i.e., concepts or

[1] For example, see Halliday and Hasan (2013, chap. 6).

[2] In general language a similar mechanism is used, too, perhaps using *compounding*: 'The process of forming a word by combining two or more existing words (Trask, 2013)'.

[3] Note that *what constitute this whole* and the *discourse* is a subject of study and a research question in itself (e.g., see Wilks and Brewster, 2009, chap. 4).

[4] Also, see the complementary perspective given based on Zellig Harris's work in Section 1.3.

[5] As cited by Wilks and Tait (2005), *Karen Spärk Jones* must be acknowledged as the pioneer of introducing ideas of this nature.

[6] For an account of the term domain-specific (or, *special*) corpus see Section 1.3. Also, note that depending on the application and availability of information resources, an ATR algorithm can use additional background knowledge, such as an existing terminological resource—see Chapter 3.

[7] Or, the classic *property assignment* (slot filling) task in Minsky's (1974) frame-based knowledge representation systems.

[8] Inarguably, Jones is the originator of the discussion about the relationship between semantic classes and the synonymy relationship between words (see Jones, 1986).

[9] Synonymy and *synset construction* are two sides of the same coin, as Wilks and Tait (2005) explain.

synsets in WordNet). Among these relations, the *hyponymy–hypernymy* is a transitive and asymmetrical relationship between synsets employed to organise general English *nouns*. The result is a hierarchical structure (i.e., a taxonomy), in which a hyponym synset is classified below its superordinate.[1]

This thesis suggests an organisation of terms based on *co-hyponymy* relationships between them, in analogy to the role that the synonymy relationship plays for organising words in WordNet. Terms and their corresponding concepts are usually organised into semantic categories; each category characterises a group of terms from 'similar' concepts in a domain knowledge—that is, a *type-of* or *is-a* relationship between a set of terms and their superordinate.[2] Terms organised under a particular hypernym are in a *co-hyponymy* relationship simply because they are hyponym of the same hypernym. For example, in an application, one may consider terms such as *corpus*, *dictionary*, *bilingual lexicon*, and so on as co-hyponyms under the hypernym *language resource* (see Figure 5.1).[3]

Using co-hyponymy as a basis for organising terminologies can be motivated by at least two observations:

a) *Persistency*: that is, many practical applications of the co-hyponymy relationships (which have emerged under various names and for diverse reasons, as is abridged in the following paragraphs); and,

b) *Regularity*: that is, in a specialised vocabulary, the co-hyponymy relationship between terms is more *frequent* than other types of relationships such as synonymy.

The latter is a direct outcome of the deliberate act of reducing lexical ambiguity in domain knowledge dissemination and in adopted perspectives in terminology (see Chapter 3). Although a synonymy relationship between terms exists (mainly as a function of *term variation* such as addressed by Freixa, 2006), to a large extent synonymy is (and to an extent polysemy) less frequent than co-hyponymy in terminological resources. In turn, the synset-based mechanism employed in WordNet is not effective for organising entries of a terminological resource, at least as a conceptual denotation (categorisation) mechanism.[4]

The overture proposed in the above paragraphs leads us to an important, though indirect outcome, of the presented study. Organising terms by characterising co-hyponymy relationships can be seen as a step towards bridging the semantic gap between the three elements a) lexical knowledge,[5] b) a conceptual representation of a domain knowledge, and c) a quantitative interpretation of meaning of terms in a specialised discourse. Given

[1] See also Resnik's (1993) elaboration on the *class-based* approach to lexical relationships.

[2] The study of the nature of this *kinds-sorts* relationship and how it is established (e.g., as examined by Carlson, 1980), unfortunately and although quite relevant, is beyond the scope of this thesis. A recent stimulating discussion on *kind-level* and *object-level* nominals can be found in Acquaviva (2014). Also, an applied perspective in the context of knowledge engineering is given by Cimiano et al. (2013). This thesis deliberately does not distinguish between the delicate difference between form and concept.

[3] This discussion is further extended in Chapter 5. As explained in Section 5.1, in the context of mapping a vocabulary to a domain ontology, terms that are *reified* to same ontological references are considered co-hyponyms.

[4] The recursive nature of hyponym–hypernym relationship can result in a controversy: at a very fine level of conceptual granularity, perhaps, there is no difference between synonymy and co-hyponymy.

[5] If one insists that it is different from the knowledge itself.

this perspective, this thesis is an investigation of vector-based distributional representations of terms in order to form a quantitative model of kinds-sorts that resembles a 'correlate to conceptual representations[1] (as nicely put by McNally, 2015)'.[2]

The proposed co-hyponymy-based mechanism for organising specialised vocabularies, in turn, paves the road towards a *class-based approach* to the manipulation of terms on the basis of their distributions in domain-specific corpora (i.e., in a similar fashion that Resnik (1993) and Brown et al. (1992) suggest for words in general language). The list of literature that motivates the identification of co-hyponym terms is beyond the references listed in this section; the emphasis that *Adrienne Lehrer* puts on the structure of vocabulary and its relationship to meaning is particularly worthwhile mentioning (e.g., see Lehrer, 1978). It is also important to note that co-hyponymy is not sufficient for capturing all the semantics in a specialised vocabulary,[3] but it is an essential relationship for extending the inventory of relationships that address a number of practical problems in knowledge engineering.

Section 1.2 continues this discussion from a computational perspective, followed by the complementary view of natural language processing in Section 1.3. Section 1.4 enumerates the practical research questions investigated in this thesis. A summary of contributions is listed in Section 1.5. Section 1.6 provides readers with information about the structure of this thesis.

1.2 Implied Computational Challenges: A Solution

Although Section 1.1 promotes a novel perspective for organising terminologies based on their distributional similarities in corpora (as with other researchers such as McNally and Herbelo (2015)), the extraction of co-hyponym terms is not a new task by all means. The identification of co-hyponymy relationships as a linguistic phenomenon has been addressed previously to meet demands in various use-cases—ranging from entity recognition and term classification methods to taxonomy learning tasks (see also the complementary introduction in Chapter 5).

The most established examples of methods that, in fact, extract co-hyponyms are *entity taggers*. Typically, lexical items of a certain *type* are annotated manually in a corpus. In this context, *type* is the hypernym or the superordinate, and annotated lexical items or entities are a group of co-hyponyms. The corpus is then employed to develop an entity tagger often in the form of a sequence classifier. These methods rely on manually annotated data, in which each mention of a term and its concept category (i.e., the hypernym) must be annotated. *Bio-entity taggers* are familiar examples of this type. Provided that enough training data is available, a reasonable performance can be attained in these recognition tasks (e.g., see report in Kim et al., 2004).

[1] Again, if we can conceive such thing without language.

[2] See also Agres et al. (2015) who apply a similar principle to investigate conceptual relationships in the context of music creativity (cognition).

[3] For example, similar to the problems resulted from *is-a overload* (as described by Guarino, 1998) and as implied by the term *tennis problem* in the context of the WordNet organisation (e.g., as explained recently by Nimb et al., 2013).

Apart from entity taggers that identify co-hyponyms, as described in Chapter 3, the co-hyponymy identification has also been addressed by a number of methods known as *term classification* (e.g., see Nigel et al., 1999). Given a taxonomy, term classification techniques, similar to entity taggers, often employ a supervised learning classification method to label terms with their hypernyms. Apart from delicate differences between previously introduced methods, they lack a number of features. These methods often do not provide a model of terms that can be used as their (intermediate) semantic representation of terms. The output is often a label, often without a degree of similarity between terms and with no built-in mechanism for representation of conceptual structures. In addition, in these methods, the dynamic nature of the co-hyponymy relationship between terms is largely ignored.

In a study, Lamp and Milton (2012) describe that the employed schema for term categorisation (i.e., the co-hyponym groups) not only changes by the dynamic of a domain knowledge, but also by the way that terms are shared and used at a specific given point in time. Hence, in a given categorisation of terms, change is inevitable—not only from a *diachronic* perspective, but also on a *synchronic* level and depending on the parties involved in the communication process. Comparably, it may be required to organise an existing terminological resource in order to address the constantly changing demands of an information system. This problem has been largely overlooked in methods previously proposed for knowledge acquisition from text (and, the identification of co-hyponym terms).

The major research challenges to develop a mechanism to address the problems mentioned above can be summarised as follows:

1) The mechanism must identify co-hyponymy relationships between terms—that is, the association of a term to a particular hypernym or a category of concepts.
2) The mechanism must be capable of capturing the dynamic nature of the co-hyponym groups in a domain knowledge (e.g., as in Lamp and Milton, 2012).
3) The mechanism must be capable of resembling the conceptual structure of a domain knowledge in some sense (see Section 1.1).

The first challenge, in general, is non-trivial since terms cannot be distinguished explicitly from lexical units that are not a term. Co-hyponym terms in particular can not be distinguished from other terms. Devising such a mechanism implies a level of *text understanding*. Therefore, it is an open research question. The second and third challenge listed above rule out the use of previously employed techniques such as entity tagging for finding and encoding co-hyponymy relationships between terms. Entity tagging and other supervised methods are too rigid to be used as an approach to reflect the dynamic of co-hyponym groups and to reflect various co-existing conceptualisation structures (e.g., manual annotations must be revised, the underlying classifiers must be retrained, or even a new classifier must be added to find and represent a new co-hyponym group).

As illustrated in Figure 1.1, identifying a group of co-hyponym terms in a terminological resource is equivalent to charactering a subset of valid terms. Evidently, from a computational perspective, the co-hyponym identification can be boiled down to a classification task. As suggested above, this formulation of the problem has been adopted in a number of previously proposed methods (e.g., see Nigel et al., 1999; Afzal et al.,

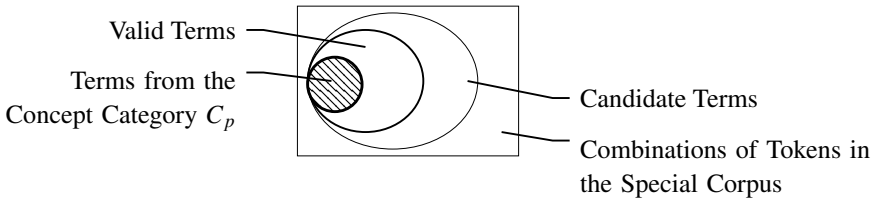

Figure 1.1: Venn diagram that illustrates the relationships among candidate terms, valid terms, and a particular category of terms C_p. ATR targets the extraction of candidate terms and the identification of valid terms. However, the proposed term classification task targets the identification of co-hyponym terms—that is, a subset of valid terms.

2008; Kovačević et al., 2012). However, in contrast to these methods and in order to address the research challenges itemised above, this thesis proposes a justification of the co-hyponym identification task in the general framework of *distributional semantics* and using a *similarity-based reasoning* process that employs *memory-based learning*. In turn, the proposed methodology is evaluated systematically.

I assume that the association of a term to a category of concepts (i.e., a co-hyponym group) can be characterised with respect to its co-occurrence relationships in the corpus. Such being the case, I hypothesise that terms from similar concept categories tend to have similar distributional properties. In order to quantify these distributional similarities, I employ vector spaces: a mathematically well-defined framework, which has been widely used in text processing (Turney and Pantel, 2010). In a vector space, candidate terms are represented by vectors in a way that the coordinates of the vector determine the correlation between candidate terms and the collected co-occurrence frequencies. Consequently, the proximity of candidate terms can be used to compare their distributional similarities. The result, as implied by Schütze (1993) and delineated later by Widdows (2004) and Sahlgren (2006), is a geometric metaphor of meaning: a semantic space that is, accordingly, called a *term-space model*.

In this term-space model, the task is to identify a particular *paradigmatic* relationship between terms—that is, co-hyponymy. It is assumed that each group of co-hyponym terms can be characterised using a set of *reference terms* or examples (shown by R_s)—that is, a small number of terms (e.g., 100) that are annotated with their corresponding hypernym (i.e., concept category). The distance between vectors that represent candidate terms and the vectors that represent R_s is assumed to determine the association of candidate terms to the group of co-hyponyms represented by R_s. This *similarity-based reasoning framework* is then implemented based on the principles of Daelemans and van den Bosch's (2010) memory-based learning—that is, using an instance-based k-nearest neighbours (k-nn) algorithm, as described later in Chapter 5. Notably, k-nn introduces a technique for similarity-based reasoning that can meet the requirements imposed by the dynamic nature of co-hyponym groups (i.e., the ability to update the rational behind the reasoning process at any time during the use of system with minimum effort). To reflect changes in the structure of co-hyponym groups, it is only required to update R_s—that is, to provide a new set of examples.

The use of this proposed method, however, is hampered by two major (related) obstacles:

1. *the curse of dimensionality*: In the proposed term-space model, due to the Zipfian distribution of words in text, vectors that represent candidate terms are usually high dimensional and sparse—that is, most of the elements of the vectors are zero. The high dimensionality of vectors hinders computation and diminishes the method's performance; the sparsity of vectors is likely to diminish the discriminatory power of a constructed term space model (see Chapter 2).
2. *the inflexibility of models to accommodate updates*: In addition, changes in the documents that represent a domain knowledge or adding new candidate terms, inevitably demands changes in the structure of the vector space that represent the domain knowledge. Previous methods employ the so-called *one-dimension-per-context-element* (see Chapter 2). Put simply, in these methods of vector space construction, the structure of vectors is firmly controlled by the input text-data. The basis of vectors (i.e., informally their dimension) is determined by the words that co-occur with terms. An update in a model (i.e., changes in the collection of documents or terms) demands a change in all the vectors since new dimensions must be appended or removed from the model. This is not acceptable considering the fact that models usually are large in size and updates are frequently necessary to reflect the dynamic of a domain knowledge.

In the presented study, special attention is paid to these problems. As a result, so-called *incremental techniques* using *random projections* are proposed to avoid the obstacles listed above (see Chapters 4 and 5).

As explained thoroughly in the following Section 1.3, in distributional analyses of languages, a major research is the study of co-occurrence relationships with respect to a targeted task (here, co-hyponymy identification). For example, in rule-based information extraction methodologies, the task of a researcher can be to identify and then characterise linguistic patterns in a formal language, such as *regular expressions* or more sophisticated grammar rules. In distributional methods, a similar effort is required; however, in another form and using mathematical tools other than rules. Although a distributional model is built automatically, research is still required to:

a) define the way these models must be constructed;
b) and then to (b) set variable parameters of the envisaged model (e.g., see the proposed research questions in Section 1.4 and the evaluation parameters discussed in Section 5.3, Chapter 5).

Evaluation of distributional models in general, and, in particular, the proposed distributional model for identifying co-hyponym terms, in a way that the interdependencies between parameters are assessed, remains an untouched area of research. Evidently, a distributional model, such as the one proposed in this thesis, is a multi-parameter system in which the interdependence between parameters is not known. In previous research, this fact has often been overlooked; hence, parameters of a model have been mostly evaluated independently of each other. To address this problem, much of the work in this thesis is devoted towards a holistic evaluation of the constructed models.

1.3 The Natural Language Processing Perspective

The motivation for this study can also be described from the perspective of *natural language processing*. Natural languages are certainly the most important vehicles for information creation and dissemination. Consequently, natural language processing has emerged as an important interdisciplinary research field that melds linguistics with computer and information science. The major objective of research in this area has been to establish an abstract system that characterises natural language. The interpretation of this abstract system must enable computers to represent, store, access, process, and unlock information that is encoded in natural languages, for instance as explained in the motivation for this thesis.

In contrast to research topics such as *human language technology*—which pursues the ultimate goal of natural language communication between man and machine similar to man-to-man communication—or, for example, computational cognitive science and psycholinguistics—which study the underlying mechanisms of understanding language in the human mind—natural language processing is modestly concerned with finding a suitable model of language to fulfil a particular task. Although all these areas of research discern the problem of natural language understanding and the meaning of meanings, in natural language processing the focus is on practical applications. To achieve practicality, then, natural language processing deliberately simplifies aspects of natural language.[1]

The foundation of natural language processing and the method proposed in this thesis can be traced back to as early as the 1950s and the growing availability of commercial computers. On one side, computers facilitated processing language corpora (i.e., a collection of text data); on the other side, using computers for information processing stimulated the need for building computable models of language. The product was the formation of a strong *empiricist*[2] approach towards analysing languages and the development of a set of data-driven techniques for their automatic processing—what are nowadays referred to as statistical natural language processing and corpus-based methods.

Simply put, these methods validate hypotheses about different aspects of natural language—such as, morphology (i.e., the structure of words), syntax (i.e., the structure of sentences), and semantics (i.e., the structure of meanings)—by collecting evidence from corpora (for an overview of these methods and their applications see, e.g., Tognini-Bonelli, 2001; Wilson and McEnery, 1996). The ever-increasing processing power of computers has made these empiricist approaches a dominant technique for realising goals set by natural language processing research.

A number of prominent researchers[3] have contributed towards establishing theoretical frameworks that can be used to explain these corpus-based, data-driven methods—see, for example, the inventory of the names listed in Jones and Kay (1973) and Moskovich

[1] In research literature, terms such as natural language processing and human language technology are often used interchangeably. The aim here is to contrast the objectives of these related areas of research. Also, it is worth mentioning that these research topics are reciprocal in their relationships, that is, research findings in one area are often employed to support claims or stimulate activities in the other. The term *computational linguistics*, perhaps, is the best representative of the aggregation of these research topics.

[2] In the sense that knowledge is elucidated upon 'sense experience' (Markie, 2015).

[3] Conceivably, of an equal importance.

(1976). In the context of this thesis, however, theoretical articulations by Zellig Harris (1909–1992) are relied upon, namely, Harris's (1954) *distributional hypothesis* and his idea of *sublanguages* (see, e.g., Harris, 1968, p. 154). As it is best described by Nevin (2002, Foreword, italics are added):

> The consequence *of Harris's theories* is that the work of linguistic analysis can be carried out only in respect to co-occurrence relations in the data of language—what had come to be called distributional analysis.

Harris's (1954) distributional hypothesis is often employed to justify a contemporary research trend in computational semantics that characterises itself by the name *distributional semantics*. As it is described in Chapter 2, distributional semantic methods use a data-driven approach for modelling and interpreting the meanings of linguistic entities such as words, phrases, and sentences. In these methods, the meanings of these entities are a function of their usage in language corpora.

Compared to the distributional hypothesis, Harris's idea of sublanguages is, perhaps, understated. Similar to the notion of substructure in mathematics, Harris argued that a certain subset of sentences in a general natural language can form a sublanguage if and only if it 'is closed under some operations' of the general natural language (the *closure* property):

> A subset of the sentences of a language forms a sublanguage of that language if it is closed under some operations of the language: e.g., if when two members of a subset are operated on, as by and or because, the resultant is also a member of that subset (Harris, 1998, p. 34).

According to Harris, in a sublanguage, information is expressed by the repeated use of limited sentence types and word classes. Therefore, once these types and classes are determined from an *analysis of sample documents*, they can be used to build a structure for the information that will be extracted from the analysis of new sample texts. Despite shortcomings—for example, as stated by Kittredge and Lehrberger (1982), the lack of an adequate definition—and harsh and contradictory critics,[1] Harris's (1968) sublanguages idea provides a theoretical basis for the corpus-based processing of (domain-specific) natural language texts. The notion of sublanguages, particularly, has been employed to justify the generalisation of findings from a *limited* number of observations in a reference corpus to the unseen and unlimited text data that is not the reference corpus.[2]

[1] Compare, for example, reviews by Wheeler (1983) and Nevin (1984): Wheeler concluded that

> The work *of Harris* does not help us with semantics, it is not mathematics, and it comes late to the problems of syntax (Wheeler, 1983, italics added).

Nevin (1984), however, suggested that sublanguages 'are essential to an understanding of semantics of natural language'.

[2] As repeatedly stated throughout this thesis, Harris is neither the first nor the only linguist who promotes the structuralist perspective of language through the functional distributional analysis of words. Similar philosophical perspectives are presented in the work of Jost Trier (1894–1970). In many respects, the notion of *word (semantic) fields* as Trier (1934) put forward is similar to Harris's sublanguages (perhaps, only a terminological difference. For example, compare this section with explanations given in Gliozzo and Strapparava, 2009). See also Chapter 2.

Since then, Harris's perspective has influenced a substantial amount of research on the automatic analysis of language. Notably, Harris's doctoral student Naomi Sager perfected and applied the idea of sublanguages to real-world applications (see, e.g., Sager, 1975). The influence of the idea of sublanguages can be further traced in the work of Sager's collaborators such as Carol Friedman, Ralph Grishman, and her doctoral student Jerry Hobbs (e.g., see chapters of Grishman and Kittredge, 2014). Through the series of DARPA's founded Message Understanding Conferences,[1] the idea of sublanguages eventually emerged as today's modern information extraction technology (see Hobbs and Riloff, 2010, for an overview of the state of the art in information extraction).

The use of this sublanguages idea is not limited to information extraction. Languages that are used in specialised communicative contexts (which from now on will be called *specialised languages*) and, respectively, the corpora that represent them (which following the suggested guidelines by Sinclair (1996), will be called *special corpora* or *domain-specific corpora*) are the most definite examples of sublanguages (see, e.g., the recent study in Temnikova et al., 2014). For example, as stated by Harris (2002), in order to reflect the information's structure in a specialised knowledge domain, a special language (e.g., the language of science writing) conforms not only to particular structures—for instance, syntactic and discourse structure—but also uses a *specialised vocabulary*.[2]

As discussed in Section 1.1, the entries of this specialised vocabulary (also known as a terminological resource) are often called *terms* and have been the subject of study in the discipline of *terminology*. Whereas traditional terminology investigated terms as self-subsisting linguistic entities, independent of their usage in text, the idea of sublanguages has encouraged the study of *terms* in *context*, as stated by Pearson (1998).[3] Disregarding the theoretical motivations, special corpora and terminological resources have been a vibrant topic in the broad domain of natural language processing and, in particular, the emerging multi disciplinary research field of *computational terminology*.

Accordingly, in this thesis, among research topics in computational terminology, the application of corpus-based methods for extracting co-hyponym terms is revisited using the aforementioned theoretical framework of Harris's distributional hypothesis and sublanguages and the mathematical framework of *real normed vector spaces*. The proposed method is then evaluated in the systematic way that is encouraged by advances in distributional semantics.

1.4 Research Questions

To investigate the hypothesis proposed in this thesis—that is, co-hyponym terms share similar distirbutional properties that can be employed to organise a specialised vocabulary—a number of research questions must be addressed. The first and foremost question—similar to other applications of distributional methods—is:

[1] See http://www-nlpir.nist.gov/related_projects/muc/.

[2] The notion of sublanguages can be approached from other perspectives, for example, see the short note and references in Karlgren (1993).

[3] Please note that the study of terms in context has been suggested by several other motivations and theories (e.g., see Faber and L'Homme, 2014).

- *What kind of co-occurrence relationships among relationships must be collected to form a suitable model to characterise the targeted structure?*

As is explained in Chapter 2, previous research in distributional semantics suggests that a *paradigmatic* relationship, such as the one targeted in this thesis, can be distinguished by collecting co-occurrence frequencies from *small* windows of text in the vicinity of candidate terms. This knowledge results in another research question:

- *What is the best configuration for this window of text?*

The question above can be broken down into several sub-research questions. However, as explained in Chapter 2 and stated in the previous research (e.g., see Baroni and Lenci, 2010; Sahlgren, 2008), at least three questions can be asked:[1]

RQ 1.1 *In which direction, regarding the position of the candidate terms, must this window of text be stretched?*

1. only to the left side of a candidate term to collect the co-occurrences of the candidate term with preceding words;
2. only to the right side to collect co-occurrences with the succeeding words; or
3. around the candidate term—that is, in both left and right directions?

RQ 1.2 *What is the best size for this window of text—for example, one or two tokens, or bigger sizes, such as six or seven?*

RQ 1.3 *Is the order of words in this window of text important; and, does encoding the sequential order of words improve the discriminatory power of models?*

After collecting the co-occurrences, several other questions arise regarding the use of the suggested similarity-based reasoning framework:

RQ 2.1 *What kind of similarity measure performs better?*

RQ 2.2 *What is the role of neighbourhood-size selection—that is, the value of k in the memory-based learning framework?*

Another question can be asked with respect to the size of corpus, namely:

RQ 3 *Is the size of the corpus used for collecting co-occurrences important? Is bigger, better?*

Last but not least:

RQ 4 *Are the obtained results consistent across concept-categories?*

[1] See additional questions in Chapter 6.

Apart from the questions listed above, a major research concern that is investigated in this study deals with the *curse of dimensionality* and the design of scalable methods for the construction of vector space models. Whereas a technique such as *truncated singular value decomposition* is mathematically well-defined, its application is limited by the resource required for its computation, particularly when dealing with big text data. In contrast, the alternative scalable technique named *random indexing* lacks adequate mathematical justifications. In this thesis, this argument is formulated by

RQ 5 *What are the mathematical justifications of random indexing in particular, and in general, incremental methods of vector spaces construction?*

The aforementioned research questions result in the scientific contributions that are described in the next section.

1.5 Summary of Contributions

Based on the principles of distributional semantics, a method for identifying co-hyponym terms in a terminological resource is proposed. The association of terms to a category of concepts, hence, the co-hyponymy relationship, is modelled as a paradigmatic relationship in a vector space model. The construction of this model is carried out automatically and at a reduced dimensionality using an incremental, thus, scalable methodology. Using minimal supervision and given a small set of examples from the targeted category of concepts, the association of terms to the concept category are computed using an example-based k-nearest neighbour classifier (see Chapter 5).

The methodology is then evaluated in the systematic way that is encouraged by advances in distributional semantics. In order to answer each of the questions asked in the previous section, several experiments are designed and performed. The outcome of these experiments confirms the validity of the proposed hypothesis and method. Each set of experiments targets answering a set of questions that are asked above (i.e., Sections 5.4.1 to 5.4.4 in Chapter 5). In turn, in Section 5.5, the observations from these experiments are discussed and a summary of the findings is provided. Based on these observations, in Chapter 6 a set of guidelines that can be used in similar tasks is proposed.

The *random indexing* technique is studied and the method's incremental procedure is explained mathematically. This study provides a theoretical guideline for setting the method's parameters which has not been previously proposed. To support the theoretical findings, the results from a set of experiments are reported. Using the proposed delineation, the random indexing method is generalised and a novel technique called *random Manhattan integer indexing* is proposed. This method can be employed for the incremental construction of ℓ_1*-normed term-spaces* at a reduced dimensionality (see Chapter 4). The method, therefore, can be used to improve the performance of distributional semantic models when similarities between vectors are measured using the city block (or, the Manhattan) distance.

The contributions listed above are discussed further in Section 6.1.

1.6 Thesis Structure

The remainder of this thesis is organised in three parts:

Part One: Background

Chapter 2 is a practical guide that walks the reader through the basics of distributional semantic methods: how they work and how they can be expressed—or formalised—in computers. More precisely, as suggested in Section 1.4, the vector space mathematics will be described and employed. In this framework, the major processes are explained, from the construction of a model through the distillation of results. The reader who is familiar with these concepts can thus safely skip this chapter. Chapter 3 introduces computational terminology and reviews methods of term extraction and classification. In doing so, the common mechanism of term extraction techniques are discussed using the jargon that is introduced in Chapter 2.

Part Two: Core Research

Chapter 4 introduces random projection techniques and their applications in natural language processing. In this chapter, the random indexing technique is revisited and justified mathematically. This justification is employed to provide a set of guidelines for setting the method's parameters. A novel technique called *random Manhattan indexing*, and its enhanced version called *random Manhattan integer indexing*, are then introduced. The discussions in this chapter are accompanied by a series of experiments to support the theoretical discussions.

The main methodology for identifying and scoring co-hyponym terms are then introduced and evaluated in Chapter 5. After introducing the methodology, the evaluation framework is laid out. The section in the remainder of this chapter, targets a particular set of research questions that are proposed earlier. The discussions in this chapter are connected to the explanations in the previous chapters; hence, the reader can start with this chapter and follow the provided pointers for relevant elaboration in other parts of the document. In addition, results from the experiments are connected to the original research questions described in this chapter.

Part Three: Epilogue

Chapter 6 concludes this thesis by providing a summary of findings. The lessons learned are discussed and additional questions that are faced during this study are presented as possible future research.

Reference List

Acquaviva, P. (2014). The roots of nominality, the nominality of roots. In Alexiadou, A., Borer, H., and Schafer, F., editors, *The Syntax of Roots and the Roots of Syntax*, volume 51 of *Oxford Studies in Theoretical Linguistics*, pages 33–57. Oxford University Press. 6

Afzal, H., Stevens, R., and Nenadic, G. (2008). Towards semantic annotation of bioinformatics services: Building a controlled vocabulary. In Salakoski, T., Schuhmann, D. R., and Pyysalo, S., editors, *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 5–12, Turku, Finland. Turku Centre for Computer Science (TUCS). 8

Agres, K., McGregor, S., Purver, M., and Wiggins, G. (2015). Conceptualizing creativity: From distributional semantics to conceptual spaces. In Toivonen, H., Colton, S., Cook, M., and Ventura, D., editors, *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, pages 118–125, Utah, USA. The Association for Computational Creativity, Brigham Young University. 7

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721. 14

Brewster, C., Jupp, S., Luciano, J. S., Shotton, D., Stevens, R. D., and Zhang, Z. (2009). Issues in learning an ontology from text. *BMC Bioinformatics*, 10(Suppl 5):S1. 4

Brewster, C. A. (2008). *Mind the Gap: Bridging from Text to Ontological Knowledge*. PhD thesis, University of Sheffield. 4

Brown, P. F., de Souza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479. 7

Carlson, G. N. (1980). *Reference to kinds in English*. Outstanding Dissertations in Linguistics. Garland Publishing, rev. version of author's thesis, university of massachusetts, amherst, 1977 edition. 6

Chan, Y. S. and Ng, H. T. (2007). Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics. 5

Cimiano, P., McCrae, J., Buitelaar, P., and Montiel-Ponsoda, E. (2013). On the role of senses in the ontology-lexicon. In Oltramari, A., Vossen, P., Qin, L., and Hovy, E., editors, *New Trends of Research in Ontologies and Lexical Resources*, Theory and Applications of Natural Language Processing, pages 43–62. Springer Berlin Heidelberg. 6

Daelemans, W. and van den Bosch, A. (2010). Memory-based learning. In Clark, A., Fox, C., and Lappin, S., editors, *The Handbook of Computational Linguistics and Natural Language Processing*, pages 154–179. Wiley-Blackwell. 9

Faber, P. and L'Homme, M.-C. (2014). Lexical semantic approaches to terminology: An introduction. *Terminology*, 20(2):143–150. 13

Feigenbaum, E. A. (1980). Knowledge engineering: The applied side of artificial intelligence. Technical Report STAN-CS-80-812 (HPP-80-21), Computer Science Department, Stanford University. 4

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press. 5

Freixa, J. (2006). Causes of denominative variation in terminology. *Terminology*, 12(1):51–77. 6

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of Speech and Natural Language Workshop (HLT'92)*, pages 233–237, Harriman, New York. Morgan Kaufmann Publishers. 5

Gliozzo, A. and Strapparava, C. (2009). Semantic domains. In *Semantic Domains in Computational Linguistics*, pages 13–32. Springer Berlin Heidelberg. 12

Grishman, R. and Kittredge, R., editors (2014). *Analyzing language in restricted domains: Sublanguage description and processing*. Psychology Press, New York, NY, US. First published 1986 by Lawrence Erlbaum Associates. 13

Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. In Calzolari, N., Choukri, K., Hoege, H., Maegaard, B., Mariani, J., Municio, A. M., and Zampolli, A., editors, *First International Conference on Language Resources and Evaluation*, Granada, Spain. European Language Resources Association. 7

Halliday, M. A. K. and Hasan, R. (2013). *Cohesion in English*. English Language Series. Routledge. First published by Longman Group in 1976. 5

Harris, Z. (1968). *Mathematical structures of language*. Number 21 in Interscience tracts in pure and applied mathematics. John Wiley and Sons. 12

Harris, Z. (1998). *Language and information*. Columbia University Press. 12

Harris, Z. S. (1954). Distributional structure. *Word, The Journal of the International Linguistic Association*, 10:146–162. 12

Harris, Z. S. (2002). The structure of science information. *Journal of Biomedical Informatics*, 35(4):215–221. Sublanguage - Zellig Harris Memorial. 13

Herbelo, A. (2015). Mr Darcy and Mr Toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 151–161, London, UK. Association for Computational Linguistics. 7

Hobbs, J. R. and Riloff, E. (2010). Information extraction. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921. 13

Jones, K. S. (1986). *Synonymy And Semantic Classification*, volume 1 of *Edinburgh Information Technology Series*. Edinburgh University Press. The book comprises Jones's Ph.D. thesis, which is approved in 1964 at the University of Cambridge. 5

Jones, K. S. and Kay, M. (1973). *Linguistics and information science*. Academic Press. 11

Karlgren, J. (1993). Sublanguages and registers: A note on terminology. *Interacting with Computers*, 5(3):348–350. 13

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In Collier, N., Ruch, P., and Nazarenko, A., editors, *JNLPBA: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, pages 70–75, Geneva, Switzerland. Association for Computational Linguistics. 7

Kittredge, R. and Lehrberger, J. (1982). Variation and homogeneity of sublanguages. In *Sublanguage: Studies of Language in Restricted Semantic Domains*. Walter de Gruyter. 12

Kovačević, A., Konjović, Z., Milosavljević, B., and Nenadic, G. (2012). Mining methodologies from NLP publications: A case study in automatic terminology recognition. *Computer Speech and Language*, 26(2):105–126. 9

Lamp, J. W. and Milton, S. K. (2012). The social life of categories: An empirical study of term categorization. *Applied Ontology*, 7(4):449–470. 8

Lehrer, A. (1978). Structures of the lexicon and transfer of meaning. *Lingua*, 45(2):95–123. 7

L'Homme, M.-C. and Bernier-Colborne, G. (2012). Terms as labels for concepts, terms as lexical units: A comparative analysis in ontologies and specialized dictionaries. *Applied Ontology*, 7(4):387–400. 4

Margolis, E. and Laurence, S. (2014). Concepts. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Center for the Study of Language and Information (CSLI), spring 2014 edition. 4

Markie, P. (2015). Rationalism vs. Empiricism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, spring 2015 edition. 11

Martinez, D. and Agirre, E. (2000). One sense per collocation and genre/topic variations. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP)*, pages 207–215, Hong Kong, China. Association for Computational Linguistics. 5

McNally, L. (2015). Kinds, descriptions of kinds, concepts, and distributions. Technical report, Universitat Pompeu Fabra. First Presented at Workshop Bridging Formal and Conceptual Semantics (BRIDGE-14). 7

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244. 5

Minsky, M. (1974). A framework for representing knowledge. Artificial Intelligence Lab Publications AIM-306, Massachusetts Institute of Technology, Cambridge, MA, USA. 5

Moskovich, W. (1976). Perspective paper: Quantitative linguistics. In *Natural Language in Information Science*, pages 57–74. Skriptor. 11

Murphy, G. L. (2002). *The Big Book of Concepts*. The MIT Press. 4

Nevin, B., editor (2002). *The Legacy of Zellig Harris: Language and Information Into the 21st Century*, volume 1: Philosophy of science, syntax and semantics of *Amsterdam Studies in the Theory and History of Linguistic Sc*. John Benjamins Publishing Company. 12

Nevin, B. E. (1984). [review of the book *A Grammar of English on Mathematical Principles*, by Zellig Harris]. *Computational Linguistics*, 10:203–211. Formerly the American Journal of Computational Linguistics. 12

Nigel, C. N., Collier, N., and Tsujii, J. (1999). Automatic term identification and classification in biology texts. In *Proceedings of the 5th Natural Language Pacific Rim Symposium (NLPRS'99)*, pages 369–374, Beijing, China. 8

Nimb, S., Pedersen, B. S., Braasch, A., Sorensen, N. H., and Troelsgard, T. (2013). Enriching a wordnet from a thesaurus. In Borin, L., Fjeld, R. V., Forsberg, M., Nimb, S., Nugues, P., and Pedersen, B. S., editors, *Proceedings of the Workshop on Lexical Semantic Resources for NLP at NODALIDA 2013*, volume 19 of *NEALT Proceedings Series*, Oslo, Norway. Linkoping University Electronic Press. 7

Pearson, J. (1998). *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. John Benjamins, Amsterdam, The Netherlands. 13

QasemiZadeh, B. (2015). *Investigating the Use of Distributional Semantic Models for Co-Hyponym Identification in Special Corpora*. PhD thesis, National University of Ireland, Galway. i

Resnik, P. (1993). *Selection and information: a class-based approach to lexical relationships*. PhD thesis, University of Pennsylvania. 6, 7

Roddenberry, G. (1965-Now). Star trek. American science fiction entertainment franchise. 4

Sager, N. (1975). Sublanguage grammars in science information processing. *Journal of the American Society for Information Science*, 26:10–16. 13

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University. 9

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54. 14

Schütze, H. (1993). Word space. In Hanson, S., Cowan, J., and Giles, C., editors, *Advances in Neural Information Processing Systems 5 (NIPS 1992)*, pages 895–902, San Francisco, CA, USA. Morgan-Kaufmann. 9

Sinclair, J. (1996). Preliminary recommendations on corpus typology. Technical Report EAG–TCWG–CTYP/P, Expert Advisory Group on Language Engineering Standards (EAGLES). 13

Temnikova, I., Jr., W. A. B., Hailu, N. D., Nikolova, I., Mcenery, T., Kilgarriff, A., Angelova, G., and Cohen, K. B. (2014). Sublanguage corpus analysis toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1714–1718, Reykjavik, Iceland. European Language Resources Association. 13

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*, volume 6 of *Studies in Computational Linguistics*. John Benjamins. 11

Trask, L. R. (2013). *A Dictionary of Grammatical Terms in Linguistics*. Routledge. First published 1992. 5

Trier, J. (1934). Das sprachliche feld: Eine auseinandersetzung. *Neue Fachbuecher fuer Wissenschaft und Jugendbildung*, 10:428–449. 12

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188. 9

Wheeler, E. S. (1983). [review of the book *A Grammar of English on Mathematical Principles*, by Zellig Harris]. *Computers and the Humanities*, 17(2):88–92. 12

Widdows, D. (2004). *Geometry and Meaning*. Number 172 in CSLI Lecture Notes. CSLI Publications, Stanford, CA. 9

Wilks, Y. A. and Brewster, C. A. (2009). *Natural Language Processing as a Foundation of the Semantic Web*, volume 1 of *Foundation and Trends® in Web Science*. now Publishing Inc. 5

Wilks, Y. A. and Tait, J. I. (2005). A retrospective view of synonymy and semantic classification. In Tait, J. I., editor, *Charting a New Course: Natural Language Processing and Information Retrieval: Essays in Honour of Karen Sparck Jones*, volume 16 of *The Kluwer International Series on Information Retrieval*, pages 1–11. Springer Netherlands. 5

Wilson, A. and McEnery, T. (1996). *Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press, 2nd edition. 11