Cyberspatial Semantics

The behaviour of prepositions in software development forums

Dylan Bonga
Supervisors: Tejaswini Deoskar, Joost Zwarts

Utrecht University

University: Utrecht University
Department: Linguistics
Date: August 27, 2019

Acknowledgment

I owe my thanks to Tejaswini Deoskar and Joost Zwarts, not only for creating this project, but also for guiding me through it. Their different fields of expertise gave interesting insights that were renewing and very helpful. Further, they allowed me to grow as a researcher, by giving me space to solve my own problems, but also by challenging me. Without their help and supervision, this project would not have succeeded to exist.

I am also grateful for the support of my fellow students in the Linguistics research master program in Utrecht. My days in the 'basement' would have been twice as long, if I did not have them to discuss, vent, and lunch with. Especially, thanks are due to James Teasdale for owing me drinks after exploiting my native language intuitions. Also, he did a pretty decent job making up the title. In addition, thanks are due to Jonathan Ben Kamp, for pursuing the same goals together. This kept me on track and significantly decreased my paperwork. I also want to thank Leonie Barabas-Weil, for helping me get through the final weeks of writing. Further, I owe gratitude to the teachers in my program for challenging me to become a better researcher, but also for getting me into shape for the thesis.

Lastly, I thank my parents, family, and friends for being supportive when needed. Additionally, I thank them for being a distraction when needed. While discussions and challenges are nice in their own way, I could not have finished the thesis if it was only discussions and challenges.

The data collected in this thesis is available on request. Please do not hesitate to send me an e-mail at dylanbonga@gmail.com.

Abstract

In this thesis, I look at the use of two types of prepositions in forum posts on software development, because these prepositions are used 'atypically'. Atypical use refers to a preposition occurring with a verb it would not typically occur with in Standard English. According to Jackendoff (1983), the prepositions examined in this thesis are used to express a *bounded (goal) path* ('to', 'into', 'onto', and 'against') or to express a *route* ('over', 'through', 'via', and 'across'). I attempt to answer the question to what extent spatial prepositions within the language domain of software development are used similarly to the use in Standard English.

To answer this question, I have collected data from a subforum of the UNIX forum, namely the Advanced and Expert Users forum. These data were gathered using web scraping, a method to collect data throughout a website. These data were segmented, cleaned up, and filtered. A total of 1,825 sentences remained where a verb and preposition occurred together that would not occur together outside this domain.

Each preposition was considered individually, in order to establish whether its use was 'atypical' or not. I found that Jackendoff (1983)'s approach accounted for a great part of the data, but that there was overlap between the verbs used with 'to' on the one hand, and 'into' and 'onto' on the other hand. Jackendoff (1983)'s approach was therefore extended by assuming that in some cases the prepositional object was underspecified, which lead the speaker to choose a more neutral preposition. This also applies to the overlap that was found for the prepositions 'via' and 'through'.

Contents

Chapter 1

Introduction

The amount of data available online is enormous, expanding with new data every day. This is mostly done in online communities, which contribute new data on daily basis. Successful online communities have committed users, a mutual purpose and active reenactment of existing policies, according to De Souza and Preece (2004). Furthermore, Nguyen and Rosé (2011) carefully show that members of an online community, especially long term members, use forum-specific jargon or slang. The forms of language introduced by online communities make it necessary to collect this data for linguistic analysis. This does not only concern new lexical items, such as the gamer term *permadeath* (meaning a character is dead permanently and cannot be used anymore), but also adding new meanings to existing lexical items, such as 'camping' in example (1).[1] In this context, 'camping' means staying in an obscured place for the purpose of ambushing other players.

(1) 'You're better off sneaking, hiding and camping.'

Furthermore, online communities do not only influence the meaning of content words. The examples in (2) show that in online communities concerning software development, prepositions are used differently as well.[2] In (2a), the preposition 'against' is used together with the verb 'link', however these do not typically occur together in spoken English. The same holds for 'into' with 'boot' and 'onto' with 'append', in (2b) and (2c) respectively. The examples in (2) are interesting, because they show that in this specific context a verb and preposition show atypical behaviour.

(2) a. 'you cannot dynamically link a Solaris shared library against Unix libraries.'
    b. 'it all failed to boot into single user mode.'
    c. 'this prevents ssh from appending onto known_hosts'

Taking software development as a subject for linguistic research seems trivial. However, there is a valid reason to research this domain of language. Since the launch of the World Wide Web

[1] Retrieved from https://forums.battlefield.com/en-us/
[2] Retrieved from https://www.unix.com/unix-for-advanced-and-expert-users/

in 1991, people all over the world became connected through their computers. New software was rapidly developed to send more and more data to other users of the Web, but also to make computers faster or automatize processes that were previously done manually. Over the last 28 years, this field has been rapidly expanding in many directions.

Because of this rapid expansion, language expanded as well. As shown above, new verb-preposition pairs that did not occur in Standard English. In order to further study the new forms of language that are developing alongside the expansion of the software domain, data must be collected from a source where users discuss software. There are different media that may be relevant, such as a platform containing documentation, such as GitHub, and question answering forums, such as the UNIX forum.

The prepositions used in (2) usually describe spatial relations. This is interesting because the examples describe events that occur within *cyberspace*. The term *cyberspace* originates from William Gibson's 1984 novel *Neuromancer* and was adapted in many different lines of research with different interpretations. Here, I will adopt the definition given by Dodge and Kitchin (2003, p.1), namely that *cyberspace* "refers to the *conceptual space* within information and communication technologies, rather than technology itself."

1.1 Current study

This thesis will look at the use of prepositions in *cyberspace*. More specifically, it will look at atypical use of prepositions in a forum on software documentation, like the examples in (2). Here, atypical use is defined as a preposition that occurs with a verb it would not typically occur with in Standard English. Assuming that this behaviour is not limited to the preposition 'against', this thesis will be based on the prepositions listed in (3). These prepositions are often used for spatial reference. According to Jackendoff (1983), these prepositions can be divided into those that describe a *bounded path* ('to', 'into', 'onto', 'against') and those that describe a *route path* ('across', 'over', 'through', 'via').

(3) into, onto, to, across, over, through, via, against

This thesis revolves around the question to what extent spatial prepositions within the language domain of software development are used similarly to the use in Standard English. In other words, to what extent do these preposition describe *paths* within *cyberspace*, and if so, does this usage correspond to the type of *paths* Jackendoff (1983) describes?

Researching this question in particular gives insight in the use of prepositions within this language domain, but it also evaluates the applicability of current theories on new data. Further, it gives insight to software development as a language domain. This is useful, because software development is a rapidly expanding field that introduces new forms of language. Additionally, it gives perspective on the interaction between prepositions and verbs (or verb classes), but also on prepositions and verbs individually.

On a broader scale, this research relates to the question how language can differ between domains or communities and how this difference manifests in online communication. It can give insight in how some online communities develop independently of other communities. Secondly, it introduces new language domains to linguistic research that were previously untouched. Thirdly, on a broader scale, this type of research may also contribute to machine learning approaches to linguistics. Because newly gathered data may improve these models, adding new words and structures to the training data.

This thesis is a pilot study, since it is the first study of its kind on this topic. The aim of this thesis is to study the extension of prepositional meaning in a different domain and to study the feasibility of such research.

This study focuses on the prepositions listed in (3). While these prepositions are not among the most frequent prepositions, they are more easy to analyse than highly frequent prepositions, such as 'in' or 'of'.[3] According to Aitchison and Lewis (2003) and Fenk-Oczlon et al. (2010), relatively high frequency can lead to *bleaching*, or unsemanticization, and polysemy. This means that highly frequent words may lose their meaning or develop several different meanings, which both make analysis more complicated. Additionally, the prepositions in (3) are frequent enough to show interesting patterns. Previously done research on these prepositions will be reviewed in chapter 2.

The examples in (2) imply language change, in the sense that new senses are added to existing verbs or prepositions. Assuming language changes gradually, based on Anttila (1989) and Hopper and Traugott (2003), I expect that the verb's senses might change, but the prepositions will not, because prepositions are a functional category, and verbs are not. Related to the main question, I hypothesize that the prepositions in the domain of software development are used similarly to the usage in Standard English. Further, I hypothesize that Jackendoff (1983)'s theory on spatial prepositions holds for this language domain. However, since some prepositional objects in this language domain are abstract, as is shown in (4), speakers have to make certain assumptions, which leads to variation in the use of the prepositions. However, still, I expect that these assumptions will be similar for different users, which means that the variation will be more or less consistent.

(4) 'make sure your fallback sshd is connected to <u>your ssh-session</u>'

To examine the main question, there are a number of subquestions that need to be answered. Firstly, how can spatial prepositions be characterized for Standard English by existing literature. In order to make clear whether the current theories on spatial prepositions apply to the prepositions used in *cyberspace*, it must be clear what these theories actually claim. Chapter 2 will discuss the literature on spatial semantics, as well as elaborate on the distinction between prepositions that are associated with *goal-paths* and those that are associated with *routes*.

Secondly, one must determine what kind of data is relevant to research the main question and how these data can be collected. The source of the data can have many forms, such as a forum or

[3]For an elaborate overview of the frequencies of different prepositions, see Leech et al. (2014)

a chat history. It is important to find an appropriate data source to gather data from, because it is unclear at what frequency verbs occur in this language domain with the prepositions examined in this thesis. More concretely, using an inappropriate data source might prove it difficult to find enough data for an analysis.

Evidence was gathered from a question answering forum by means of web scraping. This is method is preferred, because elicitation experiments at this point are not possible, since the magnitude of the 'atypical use' of prepositions was unknown previous to this thesis. Additionally, linguistic intuitions are tricky, because it is difficult to define which speakers are fitting to give intuitions and finding enough speakers makes this even more challenging. Web scraping, however, is a convenient alternative, as the data already exists and only needs to be extracted from its website.

The data source, the web scraping method, and the data pre-processing will be further explained in chapter 3. In order to see how the prepositions are used, all data are filtered. The corpus resulting from the web scraping will also be presented in chapter 3. Additionally, a subcorpus will be presented containing only the data relevant to the main question of this thesis.

Thirdly, these data will be related to the characterization of spatial prepositions made in chapter 2. This analysis will be shown in chapter 4. This is further divided in (1) whether goal-prepositions are used similarly inside this language domain and (2) whether route-prepositions are used similarly. With all the subquestions answered, I will answer the main question to this thesis, as well as discuss the implications of this research and opportunities for follow-up research in chapter 5.

Chapter 2

Spatial Semantics

This chapter gives an overview of the research done on spatial semantics and how this research can be linked to the use of prepositions. As Zlatev (2007) describes, extensive research has been done on this subject, for example by Fillmore (1968), Talmy (1983), Langacker (1987), and Jackendoff (1983, 1987), and more recently by Zlatev (2003), Denis et al. (2003), and Jackendoff (2012). Zlatev (2007) reasons that this subject has been so thoroughly researched for two reasons. Firstly, spatial relations (amongst other relations) seem to be universal across humans. For example, Mandler (1992) argues that infants preverbally already derive spatial meaning, like containment and support, from what they perceive visually. The second reason is that there are strong parallels between spatial semantics and other semantic domains. According to Pütz and Dirven (2011), space as a conceptual domain influences other conceptual domains in language that are more complex. One example is time, like in phrases such as 'at the end of the day'.

2.1 Semantic primitives

Zlatev (2007) gives an overview of cognitive linguistic research done on spatial semantics. He describes a set of semantic primitives, which are *trajector*, *landmark*, *frame of reference* (*FoR*), *viewpoint*, *region*, *path*, *direction*, and *motion*. These primitives occur under different names in the literature, in addition to other terms that are not listed here. The terms that are listed here are 'present in almost all descriptions of spatial semantics' (Zlatev, 2007, p.7). While some authors might see these terms as conceptual building blocks, often aiming to give a description of cognition, rather than language, I merely use them as descriptive notions, because I focus on language, rather than cognitive modelling. These notions are necessary, for instance, to distinguish between the entity that is located or moved and the entity this location or movement is related to. In this section, the notions *trajector* and *landmark* will be discussed in subsection 2.1.1. Subsection 2.1.2 will elaborate on the notions *region* and *path*. Then, subsection 2.1.3 will provide more detail on the notions *motion* and *direction*. Finally, subsection 2.1.4 will briefly go into *FoR* and *viewpoint*.

2.1.1 Entities

Spatial expressions usually reflect a spatial relation between two entities, an entity that has its location or movement described and an entity to which that location of movement is related. The entities can be physical objects, animate and inanimate, as well as events, as can be seen in (1).

(1) a. [$_{Animate}$ John] sat in the kitchen.
b. [$_{Inanimate}$ The lamp] stood in the kitchen.
c. [$_{Event}$ Fighting] occurred all over town.

As Talmy (1983) observes, it is crucial to distinguish different entities within a spatial expression, in order to express the type of spatial relation. Talmy (1983) shows that there is an asymmetry between two objects in a spatial relation, like in (2a). Therefore, he borrows the terms Figure (for 'bike' in (2a)) and Ground (for house in (2a)) from Gestalt psychology. The type of spatial relation follows naturally, according to Talmy (1983), because the Figure has certain distinguishable aspects that compare to the Ground. For instance, Talmy (1983) argues that the Figure is typically more movable, smaller or more salient, while the Ground is typically more permanently located, bigger, or less salient. This also explains why (2b) is more marked than (2a).

(2) a. The bike is near the house. (Talmy, 1983, p.230)
b. The house is near the bike. (Talmy, 1983, p.230)

Similarly, other authors might use the terms *trajector* and *landmark* (Langacker, 1987) or *referent* and *relatum* (Miller and Johnson-Laird, 1976), for Figure and Ground, respectively. However, Jackendoff (1983) does not make a similar distinction, although he does use the term *referential object*. Rather, he describes both Figure and Ground as *things* or objects (Jackendoff, 1987) that can be used as an argument of a function, such as a PLACE-function or PATH-function. The grammatical function of the *thing* depends on what function they are used in and which argument role the *thing* fills. Therefore, Jackendoff (1983)'s framework would still predict a difference between the sentences in (2).

The asymmetry between (2a) and (2b) is mostly of pragmatical nature. It follows from Talmy (1983) that the car as Figure (as in (2a)) is conceptually different from the car as Ground (as in (2b)). Similarly, according to Jackendoff (1983), the difference is derived by the argument's place in logical form. He describes that Ground-like *things* "are unattended and relatively less vivid" (Jackendoff, 1983, p.42). Furthermore, it is worth noting that for spatial referential sentences, both Talmy (1983) and Jackendoff (1983) only focus on sentences that contain both a Figure and Ground. I will use the terms Figure and Ground, not only to make clear predictions for the asymmetry in (2), but also because they are more useful as descriptive terms than *thing*.

2.1.2 Region and Path

The relation between the Figure and Ground can be based on a *region* or on a *path*. In the words of Jackendoff (1983), this respectively means that the Figure is located at a PLACE or translocates

along a PATH. This section will look into these two relations in further detail.

The locative relation between Figure and Ground is based on a *region* or *place*. On the one hand, according to Jackendoff (1983), *region* can be a PLACE. (4a) shows a PLACE ('there'), which Jackendoff (1983) refers to as "intransitive PPs". One the other hand, *region* can be a PLACE-function with a *thing* as an argument, such as in (3). The argument, *thing Y*, of the function is what Talmy (1983) considers the Ground. Here, (3) shows the expanded PLACE-function from Jackendoff (1983), which does not apply to "intransitive PPs". This is not a problem, because "intransitive PPs" are not considered in this thesis, so the expanded function is used here. Example (4b) shows how the PLACE-function can be applied to give the conceptual structure of the spatial relation.

(3) $[_{Place}\, x] \rightarrow [_{Place}$ PLACE-FUNCTION $([_{Thing}\, y])]$ (Jackendoff, 1983, p.162)

(4) a. John stood there.
b. John stood on the table.
$[_{Place}$ ON $([_{Thing}$ TABLE$])]$

At the same time, this function on its own does not give a complete overlap with Zlatev (2007)'s description of *region*. Moreover, the function in (3) does not specify what requirements must be met for a *thing* to be an argument, or for any NP to be a thing. This is problematic for prepositions such as 'in', that require an NP that can function as a container as prepositional object. The function can be useful, when either there are specific boundaries to what arguments the function is applicable, and, as Zlatev (2007) rightfully comments, when the universality of this function is questioned, since the argument specification should be highly language-specific.

Another view on *region* comes from Svorou (1994). She defines *region* as 'a conceptual structure which is determined by our structured knowledge about physical, perceptual, interactional and functional attributes associated with specific entities' (Svorou, 1994, p.15). This means that Svorou (1994) regards *region* as an aspect of the Ground. Firstly, a Ground can have *interior region*, which makes it a container like box or bowl. Secondly, a Ground can have *exterior region*, which applies to anything with external boundaries, such as tabletops and trees, but also humans. Lastly, Svorou (1994) describes entities that are a *region*, rather than have a *region*, like continents and countries. When describing a position "in the room" or " out the room", this would mean that the Ground ('room') has *interior region*. Describing a position "in front of the room", it shows that the Ground ('room') has *exterior region*. Svorou (1994) extends this set of *regions* with subregions such as FRONT-*region*, but also *paths*, which are described as movement along a *region*. For instance, with the preposition 'through' the Figure moves along the THROUGH-*region*.

Similarly, Jackendoff (1983) also describes *region* and *paths*, but as two separated primitives. As mentioned above, the locative relation between Figure and Ground is based on *region*, and most agree that there also should be a primitive *path*, to describe the directional relation. Like the PLACE-function, Jackendoff (1983) also describes a function for *paths*.

According to Jackendoff (1983), *path* can be divided into three subcategories: *bounded path*, *direction*, and *route*. The most important difference is the presence or absence of a start or end of

the path. What divides *bounded path* and *directions* from *routes*, is the presence of a *source* or *goal*, i.e., a point where the translocation starts or ends. Crucially, this begin or endpoint is part of a *bounded path*, whereas *directions* only extend towards it (for unspecified distance). *Routes* take the reference object as interior to the *path*.

Bounded paths can be divided in *source-paths* and *goal-paths* (Jackendoff, 1983, p.165). *Source-paths* are usually expressed by the preposition 'from', and *goal-paths* are expressed by prepositions like 'to'. Jackendoff (1983) notes that the preposition 'from' "can be followed by many PLACE-prepositions to express conceptually complex sources, whereas the path-function TO tends to combine with PLACE-functions into a single lexical item" (Jackendoff, 1983, p.165). *Routes* are described using several different prepositions, such as 'by', 'along', and 'over'. *Routes* can be generalized as a function, namely 'VIA + place-function'. In other words, the preposition 'through' can be defined as VIA + IN.[1] *Directions* are expressed with prepositions like 'towards' or 'away from'.

In (5), the PATH-function is divided into two functions, one taking a *thing Y* as argument (in (5a)) and one taking a *place Y* as argument (in (5b)).[2] The PATH-FUNCTION for both can be TO, FROM, TOWARD, AWAY-FROM, and VIA. These functions show that paths can take a *thing* or a PLACE as an argument, which is shown in (6). Here, sentences are given with the conceptual structure of the spatial relation. Examples (6a) and (6b) describe a bounded path, with the PATH-function taking different arguments. Example (6c) denotes a *route*, where the PATH-function takes a PLACE-function as an argument. Note that the Figure in (6b) remains underneath the Ground, while in (6c) it does not. Lastly, Jackendoff (1983) describes how a PLACE-function can also take a *path* as an argument, shown in (7).

(5) a. $[_{Path}\, x] \rightarrow [_{Path}$ PATH-FUNCTION $([_{Thing}\, y])]$
b. $[_{Path}\, x] \rightarrow [_{Path}$ PATH-FUNCTION $([_{Place}\, y])]$

(6) a. The mouse ran to the table.
$[_{Path}$ TO $([_{Thing}$ TABLE$])]$
b. The mouse went under the table. (Jackendoff, 1983, p.166)
$[_{Path}$ TO $([_{Place}$ UNDER $([_{Thing}$ TABLE$])])]$
c. The mouse went under the table. (Jackendoff, 1983, p.166)
$[_{Path}$ VIA $([_{Place}$ UNDER $([_{Thing}$ TABLE$])])]$

(7) Up the hill (Jackendoff, 1983, p.167)
$[_{Place}$ ON $([_{Path}$ UP $([_{Thing}$ HILL$])])]$

Note that, the *direction* path is not a semantic primitive in Jackendoff (1983), it is rather a subcategory of *path*. There is mention of *direction* as a broader mental concept in his work, however

[1] Although Jackendoff (1983) technically says 'VIA + INSIDE', Jackendoff (1990) analyses 'through' as 'VIA + IN'. Therefore, I will use this latter function, in order to avoid confusion.

[2] These functions are simplified version of the function given in Jackendoff (1983, p.166). While using these simplified functions makes it impossible to give the accurate description of multiple paths within one sentence, this is not necessary for the scope of this thesis. Thus, the simplified functions will suffice.

he does not go into detail on this. The semantic primitive *direction* is further explained in the next section.

While both Jackendoff (1983) and Svorou (1994) give an approach that allows for *regions* and *paths*, I side with Jackendoff (1983) here. Primarily because his theory makes specific predictions for individual English prepositions, but also because having a separate primitive *path* is helpful to make certain generalizations. This is shown in (8), where the region is interior, but the *path* is exactly opposite.

(8) a. John went out of the room. (Zlatev, 2007, p.11)
b. John went into the room. (Zlatev, 2007, p.11)

2.1.3 Motion and Direction

According to Jackendoff (1983), *places* and *paths* are an important part of expressing *motion*. While, as explained in section 2.1.2, location or translocation often is expressed by the preposition, *motion* is often expressed by the verb.

One could divide *motion* into the presence of perceived motion (event) and absence of perceived motion (state/stative), translocation and location, respectively. This is in line with the generalization made by Jackendoff (1983), which says that the PATH-function usually pairs with an event, whereas the PLACE-function usually pairs with a state. Additionally, there could be 'fictive' motion, such that there is no physical motion in the literal sense (Talmy, 1996), as is illustrated in (9a). Talmy (1996) argues that speakers experience a certain sense of motion with these constructions. This, as Zlatev (2007) correctly criticizes, is mostly based on (the author's perception of) cognitive processes, rather than based on linguistic input. Langacker (1987), on the other hand, phrases it differently. He argues that terms used to express *motion* and other spatial concepts are often extended to nonspatial domains (Langacker, 1987, p.168), like in (9b).

(9) a. The office is through the corridor. (Zlatev, 2007, p.12)
b. He can go quickly from one mood to another. (Langacker, 1987, p.168)

The examples (9a) and (9b) show that the presence or absence of perceived movement is not an adequate criterion for whether there is *motion* linguistically speaking. This is especially relevant for this thesis, as in many computational processes there is no movement visible (on the screen), while at the same time it can be argued that there is an event occurring. For example, in the process of (10), the 'loop' could still be performed without there being any visible trace on the screen. The only evidence of the program "reading" every sentence, is the output (every saved sentence). Additionally, in non-spatial sentences, such as (11a) and (11b), Beavers (2008) shows that *paths* can occur with strictly punctual verbs, where the absence or presence of *motion* is underspecified. This is shown for the verb 'stun', which can take many physical forms, compare (11a) and (11b). I will return to the difference between 'to' and 'into' in section 2.2.

(10) I looped through the file and saved every sentence containing a preposition.

(11) a. John stunned Mary *to / into silence by just looking at her.
b. John stunned Mary *to / into silence with his new dance moves.

For an alternative approach to *motion*, I turn to Jackendoff (1983)'s distinction between *states* and *events*. There are two grammaticallity tests that differentiate between *states* and *events*. Firstly, only *events* can be introduced by 'what happened was that', while *states* cannot, as in (12). Secondly, according to Jackendoff (1983), *events* must be expressed by present progressive aspect, whereas *states* can also be simple present tense, as shown in (13).

(12) a. What happened was that Mary drank all the champagne. (event)
b. * What happened was that Mary was in the kitchen. (state)

(13) a. Mary is running / *runs in the backyard (event)
b. Mary is sitting / sits in the backyard (state)

Jackendoff (1983) also shows that there are states where the Figure does express motion, like in (14a), which is similar to the examples in (9a). He also shows that there are events where the Figure does not express motion, as in (14b).

(14) a. The highway extends from Denver to Indianapolis. (state) (Jackendoff, 1983, p.172)
b. The bird is staying/*stays in its nest. (event) (Jackendoff, 1983, p.172)

Section 2.1.2 showed that Jackendoff (1983) regards *direction* as a type of *path*, as do other authors like Hawkins (1984). However, Zlatev (2007) argues against this. He motivates separation between the two by arguing that languages like German and English do not combine paths with motions, except for directions. This occurs with verbs such as 'sink', illustrated in (15), given that English being a satellite-framed language, i.e. the language uses a particle or prefix to encode the path of motion. According to Zlatev (2007), (15b) should be fine and (15c) should be over-specifying, given that a *direction* should be combined with the verb.

I consulted a native speaker of English, who rendered both (15b) and (15c) as grammatical. Moreover, the native speaker indicated that the meaning of (15b) was more similar to (15a), in the sense that when no explicit path is given, (15b) is the more probable meaning of (15a). Additionally, the informant found (15c) not over-specifying and added that using 'toward' is only appropriate if the ship never reached the bottom, or the speaker does not know if it did.

(15) a. The ship sinks.
b. The ship sinks to the bottom of the ocean. (bounded path)
c. The ship sinks toward the bottom of the ocean. (direction path)

Either way, the example given in (15c) argues that *direction* is an important notion, but it does not argue it cannot be a subcategory of *path*. One could simply argue that only *direction paths* can be combined with *motions* in these languages, which means that there is no overt realisation of the *direction path*. Zlatev (2007), however, gives another argument in favor of having a distinct notion of *direction*, illustrated with example (16).

(16) a. John flew over the bridge. (Zlatev, 2007, p.17)
b. John walked over the bridge. (Zlatev, 2007, p.17)

The Figure in (16a) can cross the Ground in any direction, whereas the Figure in (16b) can only cross the Ground along the surface of the bridge. The main difference between the reading of 'over', here crossing an object via the upper end, is at what side the path begins. In other words, only the *direction* of the Figure differs. Therefore, it is important to keep *direction* as a primitive.[3]

2.1.4 Frame of Reference and Viewpoint

Zlatev (2007) notes that *FoR* has been defined and applied in many different ways in the literature, such as Levinson (1996)'s approach. *FoR* is described by Levinson (1996) as a way of specifying angles on the horizontal axes. In English, according to Levinson (1996), this is done by using the features or axes of the Ground or the viewer's body as a reference point (or *Viewpoint*). He describes three different *Frames of Reference*. However, for this thesis these notions are not of sufficient importance to evaluate these different views. Because the spatial expressions occur on the screen and not in the physical world, I will cautiously assume that the *FoR* can take two forms. On the one hand, it can be *intrinsic FoR*, which has the speaker as 'origin'. This means that objects are perceived with the viewer as reference point, as in (17a). On the other hand, it can be *relative FoR*, which has the Figure as 'origin'. This is shown in (17b).

(17) a. Sit behind me (Zlatev, 2007, p.9)
b. He is in front of the bush (Zlatev, 2007, p.9)

2.2 Prepositions

The notions defined in section 2.1 are not only applicable to prepositions. There are various spatial relations possible without the use of prepositions, such as (18). However, given the aim of this thesis, the analysis here (and the thesis in general) only concerns data that includes the use of prepositions. Additionally, the prepositions can specify different types of relations, such as a spatial or temporal relation, based on either a *region* or a *path*. For instance, 'into' can also describe a temporal path, as in the sentence 'he partied well into the morning'. Jackendoff (1983) relates these kind of sentences to the *Thematic Relations Hypothesis*, formulated by Gruber (1965). This hypothesis states that "in any semantic field of [EVENTS] and [STATES], the principal event-, state-, path-, and place-functions are a subset of those used for the analysis of spatial location and motion" (Jackendoff, 1983, p.188). This means that a different semantic field, in this case *Identificational* (Gruber, 1965, Jackendoff, 1983), uses *paths* like they are used in spatial relations. The functions for *paths* (and for places) described in chapter 2 therefore still apply. An example

[3]Furthermore, *direction* is an important part of vector-based analyses (cf. O'Keefe (1996), Zwarts (1997)). However, in order to keep this chapter clear, I will not go into this.

is given in (19), which was adapted from Jackendoff (1983, p.194). In the example, the conceptual structure is shown for the *goal-path*, according to Jackendoff (1983).

(18) Sue entered the room.
[$_{Event}$ GO ([SUE], [$_{Path}$ TO ([$_{Place}$ IN ([$_{Place}$ ROOM])])])][4]

(19) The coach changed into a pumpkin.
[$_{Event}$ GO$_{Ident}$ ([COACH], [$_{Path}$ TO$_{Ident}$ ([PUMPKIN])])]

Likewise, spatial relations can be expressed without using a Ground or prepositional object (here: PO). However, given the scope of this thesis, I will focus on data that do contain a PO in the PP. The structure that is typically found is shown in (20a). This is also the syntactic structure I will primarily focus on. Furthermore, I will go into spatial utterances with and without a Figure, which in this construction is often the direct object (compare (20b) and (20c)).

(20) a. Subj Verb (DirObj) [$_{PP}$ Prep PO]
b. Peter threw the ball [$_{PP}$ to Bill]
c. Peter threw [$_{PP}$ to Bill]

Following the semantic primitives that are illustrated in 2.1, one can argue that the prepositions relevant to this study describe a *path*. Denis et al. (2003) argue that prepositions can be divided into *static*, or location-denoting, and *dynamic*, or path-denoting. This is in line with the division between PLACE and PATH (Jackendoff, 1983). Denis et al. (2003) first test the hypothesis that the difference between *static* and *dynamic prepositions* is that only *static* prepositions can be used with stative or positional verbs. Lakoff (1966) has several tests to check whether a verb is stative or not. Example (21) shows three of the tests, where the stative verb 'know' is ungrammatical in these sentences. Additionally, in a study on positional verbs in Nen, Evans (2014) briefly describes that positional verbs denote a 'posture', such as 'lay' or 'stand', or denote a spatial disposition with respect to the ground, as in (22).

(21) a. I am running/*knowing the answer.
b. What I did was run/*know the answer.
c. I persuaded John to run/*know the answer.

(22) John wedged himself between two old ladies in the train.

The hypothesis Denis et al. (2003) tested is problematic for prepositions such as 'across', as is shown in (23). Because this hypothesis does not hold, Denis et al. (2003) generalize that static prepositions can also denote a *bounded* path, when combined with a motion verb, as in (24). Furthermore, they argue that in fact all motion verbs select for a *path* argument, which enables the verb to "enforce this type onto their complement PP" (Denis et al., 2003, p.4).

[4]Jackendoff (1983) uses GO as a traversal function for events.

(23) a. John walked across the field.
b. John lay across the bed.

(24) Bill arrived at the party. (Denis et al., 2003, p.124)

Some of the prepositions examined in this thesis would be described as *dynamic* by Denis et al. (2003), such as 'into', 'onto', and 'to'. The other prepositions need to be accounted for as well. Denis et al. (2003) argued that only *static* prepositions can occur with stative or positional verbs ('be', 'stand'), as well as with verbs of inherent direction ('enter', 'arrive'). Example (25) shows the positional verb 'to stand', the verb of inherent direction 'to fall', and the manner of motion verb 'to jump' (32c). This is shown in (25a) and (25b), where the use of 'at' results in a fully grammatical sentence, whereas the use of 'to' does not. Secondly, they showed that verbs that express manner of motion ('jump', 'run') can occur with both *static* and *dynamic* prepositions, as is shown in (25c).

(25) a. Billy stood at/*to the bar.
b. Billy arrived at/*to the bar.
c. Billy ran at/to the wall.

Table 2.1 shows what prepositions can occur with what type of verbs. The most important difference is shown in the first column, since only *static* prepositions can occur with a stative or positional verb. This implies that both 'against' and 'across' are actually *static* prepositions. They pattern similarly to 'at' in (25). Therefore, I assume that they are *static* and thus will only denote a *path* when used in combination with a motion verb.

Preposition	Positional	VIDM	MoM
across	+	+	+
against	+	+	+
into	-	+	+
onto	-	+	+
over	-	-	+
through	-	+	+
to	-	-	+
via	-	+	+

Table 2.1: Denis et al. (2003)'s types of verbs each preposition can occur with.

Furthermore, the table also shows that both *static* and *dynamic* prepositions can occur with manner of motion verbs, as the last column shows only pluses. This is in line with the claim made by Denis et al. (2003). Additionally, the Table also shows that some of the *dynamic* prepositions ('into', 'onto', 'through', 'via') are able to occur with a verb of inherent directed motion, while others are not ('over', 'to'). However, I will not further pursue this issue, as it is not relevant to this thesis.

Since *static* prepositions can still denote a *path*, I will make a division between prepositions that can denote a *goal-path* (onto, into, to, against) and prepositions that possibly denote a *route* (through, over, via, across). The first group will be discussed in further detail in section 2.2.1, the second in section 2.2.2.

2.2.1 Bounded paths

Described by Jackendoff (1983) as *bounded paths*, the prepositions addressed in this subsection are used to express a path with a defined begin point or endpoint. Note that the prepositions described in this section are associated with a *goal* only, not with a *source*. Therefore, I will refer to these kind of *paths* as *goal-paths*, rather than *bounded paths*.

To, into, onto

As mentioned in 2.1.2, Jackendoff (1983) describes that *goal-paths* can be conceptualized like a TO-function, here specified in (26). The preposition 'to' is depicted in (26a), with a corresponding example in (27a). On the other hand, 'into' and 'onto' are a combination of the PATH-function (TO) and a PLACE-function (IN/ON) as its argument. The function is given in (26b), with the corresponding example in (27b). However, as Denis et al. (2003) correctly criticize, the functions in (26) do not make any restrictions as to what argument can be used as *thing Y*, as mentioned before in subsection 2.1.2. This is an issue that is more clearly shown in (27c), if it has the reading that the Figure ends up inside the hill. If the function is not given any restrictions, it does not account for situations that are pragmatically impossible (given that the hill in (27c) is a regular hill with no hidden tunnels).

(26) a. $[\text{PATH}] \rightarrow [_{Path}\ \text{TO}\ ([_{Thing}\ \text{Y}\ ])]$
b. $[\text{PATH}] \rightarrow [_{Path}\ \text{TO}\ ([_{Place}\ \text{Y}\ ])]$

(27) a. 'Max walked to the house.'
$[_{Path}$ TO ($[_{Thing}$ HOUSE])$]$
b. 'Max walked into the house.'
$[_{Path}$ TO ($[_{Place}$ IN ($[_{Thing}$ HOUSE])])$]$
c. # 'Max walked into the hill.'
$[_{Path}$ TO ($[_{Place}$ IN ($[_{Thing}$ HILL])])$]$

Denis et al. (2003) propose to make a further distinction between *bounded paths*. This is done to account for the issue shown in (28)[5], namely that (28a) and (28b) are ungrammatical with 'to', but not with 'into' and 'onto' respectively. The difference according to Denis et al. (2003), between 'into' and 'onto' on the one hand and 'to' on the other hand, is that 'into' and 'onto' describe a *minimal path*, whereas 'to' describes an *extended path*. A *minimal path* can be described as a transition from source (here: all points considered not the end-point) to goal (here: all points

[5]See Gehrke (2007) for an analysis of *directional 'in'* and *directional 'on'* in (28a) and (28b) respectively.

considered the end-point). They describe that the *extended path* is internally more complex, as it contains at least three subparts.

(28) a. John stepped *to / in(to) the theater.
b. John stepped *to / on(to) the crate.
c. John walked to / into the theater.
d. # John stood just outside the theater. Then, he walked to the theater.

This approach is more extensively explained by Beavers (2008), who explains the difference between durative and punctual events. Punctual events only have a start point and an endpoint, which makes these events non-gradable or non-scalar. At the same time durative events also have a medial subevent, which makes these events gradable or scalar. This scalarity is also applicable to *paths*, according to Beavers (2008). The preposition 'to' can only occur in durative events (Beavers, 2008), and can only combine with other durative events, such as durative verbs. Verbs like 'walk', in (28c), are durative and thus this sentence is grammatical and coherent. However, in a sentence like (28d), "the context inherently involves a simplex path" (Beavers, 2008, p.15), which corresponds to Denis et al. (2003)'s *minimal path*. In other words, the path is non-gradable, which is not allowed with the gradable 'to'. Beavers (2008) proposes a test to distinguish between durative and punctual events, or gradable and non-gradable scales, with the test in (29). Beavers (2008) argues that the human subject in (29a) is only capable of drawing in a sequenced manner, this event has to be durative, which is why using the adverb 'instantly' makes the sentence incoherent.

(29) a. John drew a circle (#instantly). (Beavers, 2008)
b. The (special new) printer drew a circle (instantly). (Beavers, 2008)

(30) a. John walked to the restaurant (#instantly).
b. John walked into the restaurant (instantly).
c. John walked to the stage (#instantly).
d. John walked onto the stage (instantly).

As for 'into' and 'onto', in (28a) and (28b) respectively, these prepositions are underspecified for gradability and, thus, can occur with different verbs.[6] The difference between 'to' (gradable) and 'into' and 'onto' (underspecified) is reflected in the difference between (30a) and (30b), and between (30c) and (30d).

In order to describe the difference between 'into' and 'onto', first the difference between 'in' and 'on' must be described. The preposition 'in' implies containment (Landau and Jackendoff, 1993, Jackendoff, 2012). In other words, the prepositional object of a PP with 'in' must be able to serve as a container. This can be a literal, physical container, such as 'in a box', as well as a more abstract container, such as 'he fitted well in that group'. The *region* the Figure is located, is

[6] For a more detailed analysis, see Beavers (2008).

interior to the Ground. The prepositions 'into' therefore expresses a *path* that ends interior to the Ground.

The preposition 'on', on the other hand, demands that the prepositional object has a surface on which the Figure can place itself (Talmy, 1983, Landau and Jackendoff, 1993, Jackendoff, 2012). This is not necessarily a flat surface, nor necessarily a horizontal one, as can be seen in example (31). The Figure is located exterior to the Ground and makes contact with it. The preposition 'into' expresses a *path* that ends in contact with the Ground and exterior to the Ground.

(31) a. The elephant balanced on the ball.
b. Mario fell on(to) the spikes.
c. I will stick a post-it onto the board.

Against

As mentioned above, 'against' is a *static* preposition, rather than a *dynamic* preposition, in terms of Denis et al. (2003). The pattern for 'at' in (25) is the same as for 'against' in (32), as it can occur with all three types of verbs. This also means that it can denote a *path*, but only when occurring with a motion verb, like in (32a) and (32c).

(32) a. Billy stood against the bar.
b. Billy fell against the bar.
c. Billy ran against the wall.

Being *static* in Denis et al. (2003)'s terms corresponds to Jackendoff and Landau (1991). Their chapter describes how objects in spatial relations are encoded by language. They note that 'against' is similar to 'on', as it expresses a *region*. Additionally, they describe that 'against' is exterior and makes contact. Furthermore, they describe that this contact is often an exertion of force, as can be seen in (33). (33a) shows mild force, as gravity pulls the Figure ('mother') over, but the Figure is stopped by the Ground ('cupboard'). Example (33b) shows that 'against' can also be used with a different type force, namely a car being driven. Jackendoff and Landau (1991) note that 'into' also can be used to express exterior contact, but only with considerable force (compare (34a) and (34b)).

(33) a. Mother leaned against the cupboard.
b. Mother drove the car against the wall.

(34) a. * Mother leaned into the cupboard.
b. Mother drove the car into the wall.

Rhee (2002) describes various historical uses of 'against', based on the Oxford English Dictionary. On the one hand, Rhee (2002) describes meanings and uses of 'against' that are similar Jackendoff and Landau (1991)'s description, such as 'toward and in direct contact with' or 'into direct

collision with'. Other descriptions are related, such as 'towards' and 'in strict spatial proximity', where, crucially, there is no contact. On the other hand, the description can be exactly opposite to Jackendoff and Landau (1991)'s description. One of the readings Rhee (2002) describes is 'in opposite direction to the course of anything'. While the examples in Rhee (2002) are somewhat dated to generalize to Modern English, 'against' can still be used as 'in opposite direction', such as in (35). This is closely connected to a non-spatial reading of 'against', which is 'being opposed', as in (35b).

(35) a. Fred is swimming against the current.

b. Fred is against tipping the waitress.

2.2.2 Routes

As explained in 2.1.2, *routes* are associated with a spatial relation where the Figure at some point along the path was nearby or inside the Ground. For example, when a Figure moves 'through a tunnel', the *source* and *goal* of the path are unknown, but there is certainty that at some point the Figure is inside the Ground ('tunnel'), whereas it was outside before and after. Jackendoff (1983) describes this as 'VIA + PLACE'. While this could easily be implemented for 'through' (VIA + IN) and 'via' (VIA + AT). The prepositions 'over' and 'across' are less straightforward, because they are more ambiguous.

Across

Like 'against', 'across' is a *static* preposition in Denis et al. (2003)'s terms. This can be derived from (36a-36c), since they are all fully grammatical. Examples (36b) and (36c) contain a *path*, as the preposition is combined with a motion verb.

(36) a. Billy stood across the bridge.

b. Billy went across the bridge.

c. Billy ran across the wall.

However, 'across' is also ambiguous. This is shown, for instance, by Xu and Badler (2000). In their study, they explain how to implement directional prepositions and their elicited movement in a 3D agent animation system. They show implementations for several different prepositions, such as 'near', 'along', 'across', and 'over'. Xu and Badler (2000) found that it is necessary to split the trajectories involving 'across' in two categories, in order to successfully implement this preposition into their model. These two categories imply that either the Figure translocates via the inside of the Ground ('walk across the room') or via the outside of the Ground ('walk across the bridge').

Over

Xu and Badler (2000) also describe that 'over' is more difficult to implement in their model, since it optionally has altitude as a third dimension. Interestingly, this is not always the case, as is shown by the asymmetry found between (37) and (38), where the first pair (37) has a very different meaning and the second pair (38) does not. This indicates that there are at least two readings of 'over', the first being translocation via the upper end of the Ground (VIA + TOP SURFACE) and the second being 'through'-like. Note that the Figure does not necessarily make contact with the Ground, as in (39).

(37) a. John went over the mountains.
b. John went through the mountains.

(38) a. John went over the slides for his presentation.
b. John went through the slides for his presentation.

(39) John flew over the mountains.

Through, via

In the line of Jackendoff (1983), the preposition 'through' can be seen as VIA + IN. Section 2.2.1 already covered 'in'. For 'through', VIA + IN means that the Figure translocates into a container and out from it. This can be inside an object that has clear physical boundaries, such as 'through a tunnel', but also an object where the boundaries are less clear, such as 'through a group of people'.

One issue with the preposition 'through' is that it is difficult to distinguish between its spatial use and its instrumental use. An example is given in (40). Example (41) shows that the spatial 'through' can be replaced by the preposition 'via', whereas the instrumental 'through' can be replaced by the preposition 'with'.

(40) a. Sophie accessed the building through the door.
b. Sophie accessed the building through her access card.

(41) a. Sophie accessed the building via/*with the door.
b. Sophie accessed the building *via/with her access card.

The preposition 'via' can be seen as VIA + AT. It involves a Figure traversing a path and either approximating the Ground or going through the Ground. This is shown by the ambiguity of (42). The sentence can either mean that the Figure (*John*) was interior or exterior to the Ground of 'via' ('the house'). The ambiguity mentioned here possibly derives from AT.

(42) John went to the yard via the house.

As mentioned in chapter 1, high frequency leads to *bleaching* and polysemy. This is also the case with 'at', which can have many different meanings, as is shown in (43). In (43a) shows the Figure is in proximity the Ground, inside or outside. The Ground in (43b) does not have an interior or exterior, but also shows a proximity to the Ground. However, (43c) implies that the Figure is within the borders of the Ground. For the preposition 'at', the Ground only needs to bounded in extent (Jackendoff and Landau, 1991, Landau and Jackendoff, 1993). In other words, the Ground cannot extend infinitely and has to have boundaries. The polysemy of 'at' will also be present for 'via', given its function VIA + AT.

(43) a. I will meet you at Melbourne central. Vasardani et al. (2017, p.4)
b. John is at the tram stop. Vasardani et al. (2017, p.4)
c. The butterflies live at the park. Vasardani et al. (2017, p.4)

Interestingly, the prepositions that denote a *routes* can also be used to indicate the instrument used for communication. This is shown in (44). While 'over' and 'through' can be used for both e-mail and phone calls, an informant showed slight preference for 'across the phone' over 'across e-mail', as well as 'via e-mail' over 'via the phone', the preferred form is underlined in (44c) and (44d).

(44) a. We discussed it over e-mail / the phone.
b. We discussed it through e-mail / the phone.
c. We discussed it across e-mail / <u>the phone</u>.
d. We discussed it via <u>e-mail</u> / the phone.

2.3 Cyberspace

Interestingly, the prepositional objects in *cyberspace* do not always have a visible surface, nor do they always show an apparent way of containment. However, when the prepositional object does show either of these properties, I suspect a great preference for the preposition that typically goes with that property. For example, a 'folder' can contain 'files', thus it serves as a container. The 'folder', like many other aspects of the user-interface, is used as a metaphor to make the usage of a computer more efficient and understandable. However, in computer memory, a folder is represented differently and not a strictly container. Therefore, I expect that the user-interface representation influences the selection of the preposition.

However, whenever it is unclear what the spatial properties of the prepositional object are, I expect that several different prepositions are used. More specifically, I expect that the speakers, when in doubt, use a preposition that is more general instead of more specific. Because the more specific makes assumptions on the prepositional object that the speaker does not want to make, they rather use a preposition that is too general. Therefore, I also expect overlap, because some speakers have more experience with certain objects and are, thus, more certain about its spatial properties.

2.4 Conclusion

In this chapter, I have described views on semantic primitives that are constant in studies on spatial semantics. Most spatial relations in language occur between a Figure and Ground, where the Figure is located at or translocates in relation to the Ground. For the data of this thesis, I assume that the *FoR* in the data will almost exclusively be a *intrinsic FoR*, with the speaker as Viewpoint or an *relative FoR* that is object-oriented, where the speaker is not involved. The spatial relations can be expressed in different ways, such as by using prepositions.

This chapter has also shown that 'against' and 'across' are *static* prepositions, whereas the other prepositions examined in this thesis ('into', 'onto', 'to', 'through', 'over', and 'via') are *dynamic*. This means that 'against' and 'across' need a motion verb in order to be able to denote a *path*. Furthermore, the prepositions addressed in this thesis can then be divided into prepositions used to express a *goal* (to, into, onto, against) and prepositions used to express a *route* (through, over, via, across). While *bounded paths* have a start or endpoint, the prepositions focused on here are only goal-oriented. Moreover, a division between *minimal paths* and *extended paths*, as has been done by Denis et al. (2003), is expedient, since Jackendoff (1983)'s analysis could not account for the differences formulated in (28). As was shown in 2.2, the prepositions can have different readings, depending on the verb and the prepositional object. Interestingly, it is possible to use all four prepositions expressing a *route* in a construction where the prepositional object is an instrument of communication.

Chapter 3

Data Collection

This chapter will elaborate on the data source used in this thesis. The medium that is examined in this thesis, namely forums, will be specified in section 3.1. In order to collect the data, web scraping is used. This method allows a user to gather and collect data from a website by *crawling* through the website and *scraping* the relevant data. Section 3.2 will supply more detail on why this method is used, on how it works, and on which software is used. The raw data are then pre-processed, which is described in section 3.3. Finally, the resulting dataset is presented in section 3.4.

3.1 Language domain

Crystal (2011) describes in more detail how online data can be similar to both speech and writing. In online data, a person is expected to respond as soon as possible. However, like in writing, an immediate response is expected, but not required. Given that there is no delay in delivering the message (as there would be in delivering a letter), this could lead to several follow up messages requesting an answer. The urge for replies is heavily dependant on the used medium. Additionally, online data mostly missed what Crystal (2011) describes as simultaneous feedback, such as facial expressions and body language. Some media account attempt to cope with this by allowing emoji's or pictures to be added.

3.1.1 Questing answering forums

There is not one way to handle online data, since it includes many different genres of data. In this thesis, question answering forums are considered as a primary data source. Interaction on these forums is as follows: one person opens a question page, or thread, and asks their question. Then other users of the forum can respond to that question. For some forums, being allowed to ask or respond to a question requires being a member of the forum.

Although the user asking the question wants a response promptly, the discourse is not considered time governed, like writing. The responding user can take as much time as they want for

answering, although taking to long may defeat the purpose of answering, since the questioning user might forget about posting or they might find a different solution themselves. Ideally, each thread contains exactly one discourse.

Most forums nowadays support the use of emojis, however, given the purpose of the forum (question answering), these are disregarded as a substitute for simultaneous feedback here. While they might prove beneficial for analysis in other media, the use of emojis in the forum that is used in this study is minimal. Therefore, I will not pursue this issue any further.

Forums often have rules and policies that are enforced upon the users by moderators and administrators. These policies can be extensive in forums where privacy and safety are important, such as support group forums for people with a specific disease. There are also forums where there is a bare minimum of rules, giving the users to discuss anything in whatever manner they would like. Question answering forums have a very clear goal, namely to have the question answered with an appropriate answer. To enforce this, moderators might delete messages that are counterproductive, as they give no answer or provide false information. Although these messages can be altered or deleted by other users, this is mostly done by moderators or administrators to enforce rules.

3.1.2 UNIX forum

The UNIX forum is an online forum on software development that is highly useful for this thesis. The forum is created to provide technical support for users of UNIX and Linux systems, so they share their problems and discuss different solutions. The goal of the users is straightforward, to find help or to give help to other users. It is difficult to measure how committed users are to give answers in these threads, however most threads have various extensive replies. That is to say, there are various users that take time to give the solution, but also give an explanation why they think that is an appropriate solution, or discuss different solutions that might be relevant.

As mentioned in subsection 3.1.1, forums have different policies that make the user experience more pleasant. Likewise, the UNIX forum has an administrator and a team of moderators that enforce these policies. Browsing the website can show messages of them ‘merging duplicate threads’ or forbidding users to ‘bump’ their thread, which is a way of raising a thread to the attention of other users, in order to get an answer quicker. Furthermore, the website’s privacy regulation explicitly states that “it is been [their] long-standing policy not to permit any personal or private data in public posts”, personal data such as email addresses are deleted from posts by the moderators. Because of these policies and because of the committed enforcement by the administrator and moderators, the UNIX forum in general is a valuable resource.

Additionally, the UNIX forum has different subforums, each devoted to a specific group of users or a specific type of problems. The administrator and moderators may choose to move threads to different subforums, because they are more relevant there or more likely to receive an answer in a different subforum. The advantage of using different subforums is that, for instance, novice and advanced users are separated, making it easier for each group to browse relevant threads.

Because the subject of the subforums differs, this study will focus on one subforum in particular, namely the subforum for advanced and expert users (AEU forum).[1] This subforum contains about 8,800 threads, containing 36,700 posts at the end of the scraping process (02-01-2019). In addition to the reasons mentioned above, this specific subforum is also used for other reasons: the target group, the size of the subforum, and the accessibility.

Firstly, the target group for the AEU forum is advanced and expert users. Because their posts are directed at other advanced users, it is more likely that they will use the language appropriate to this language domain. In other words, they do not have to simplify their answers, because they expect the other users to be on the same level. Possibly, this in turn will also tempt other expert users to contribute, because they are up for the challenge. In other subforums, where novice users are active, there are substantially more threads, discouraging the expert user to go through all of them, because their degree of expertise is most likely not needed there.

Secondly, the subforum turned out to be of sufficient size for this thesis. While this was not an initial criterion, it is helpful to consider the size of the subforum beforehand. Some of the other subforums are bigger and might provide more data, but it is likely that the threads on novice problems have a simpler and shorter answer. In other words, the amount of posts in the AEU forum may be lower, but the size of the posts might be bigger. Furthermore, other subforums of equal size are only on coding and scripting, which leads to posts consisting of mostly computer code, lacking natural language. While the AEU forum might have been too small, additional data could have been gathered from different forums that were also deemed suitable.

Thirdly, the UNIX forum is accessible for non-members. In other words, one does not need to create a membership to the forum to access the threads. This is highly convenient for using a web scraper. While this is initially a more general criterion, it also possibly leads to more involved users. Given that the AEU forum is separated from other subforums that include content on personal matters, Holtz et al. (2012) points out that the relative anonymity of forums lifts certain social constraints. Because the AEU forum is separated, its users are probably more willing to supply information in a thread.

3.2 Data collection method

The data of this thesis are extracted from the AEU forum using a web scraper. This is a computational tool that consists of two parts, a web crawler and a scraper. Simplified, a web crawler is used to 'crawl' through a website, i.e., go through a HTML-page and collect all hyperlinks to other pages. A scraper can open a hyperlink and collect data from the corresponding website, the features of which the programmer can specify. Web scraping was used, because elicitation experiments and linguistic intuitions are not applicable at this point, as mentioned in chapter 1. Furthermore, this method is a very efficient way of gathered a great amount of data, mainly because one person can construct the web scraper and because the program can run without supervision.

[1] Accessible at https://www.unix.com/unix-for-advanced-and-expert-users/

3.2.1 Crawler

As mentioned above, a web crawler was used to navigate through the website. The crawler used here is based on the crawler described by Lawson (2015). It takes an initial website as a start point, or seed, which in this case is the home page of the AEU forum. First, it opens and downloads the HTML-code that is used to construct this page. Then it gathers all hyperlinks from the HTML-code, which lead to other web pages. On those new web pages, it repeats this process. The crawler used here works on two levels: the main page and the thread. Because the scraper processes every link that was collected by the crawler, the crawler was specified in several ways in order to avoid duplicates.

Firstly, it was specified to only collect links that lead to a location within the same forum, because otherwise the scraper would gather data from different websites or subforums. These links occur on both the main page and within the threads. By doing this, all other parts of the UNIX forum were excluded, as well as other web pages. This was done by matching the AEU forum's address (URL) to every found link using regular expressions.

Secondly, the crawler was allowed to collect links that lead from the main page to a thread, but not links that lead from one thread to a different thread. Because the threads have links to other threads within the same forum, including these links might lead to duplicates later in the process. When the crawler was crawling 'inside a thread', i.e., not on the main page, it was allowed to collect all links to different pages of that thread. This is necessary, since some threads have an extensive amount of replies and the website then displays only seven posts per thread-page.

Thirdly, all links that were collected, were transformed in such a way that the crawler was able to check easily whether a link has been previously collected. For instance, some or the URLs have a long tail of numbers and letters, which do not influence the path of the URL. These tails were cut off, since they are redundant for this task. If it had been collected before, the link would be discarded. Otherwise, it was added to the queue of the crawler. This way, duplicates are prevented. The crawler will continue until there are no links left in the queue.

These measures were taken to make the web crawler more efficient. The process of web scraping can be time consuming, especially with bigger data sources. Therefore, it is useful to counteract duplicates and errors before they occur. Not only will duplicates lead to unwarranted frequency counts, duplicates also lead to more work cleaning up the collected data at a later stage. Errors during the crawl can interrupt the program while it runs. This can happen at any moment. Given that the process can take multiple days (and nights), it is less convenient for the programmer to continuously check the progress and restart the program after every error. Certain errors, such as internal server errors, can be accounted for quite easily, by reattempting that same URL twice.

One of the reason that web crawling is such a time consuming process, regardless of the size of a website, is that it is a burden on the server that hosts the crawled website. Repeatedly downloading can influence the performance of a website, which often leads to the person downloading being banished from the website. For that reason, most website allow crawling and scraping, but with a certain delay. That is, every download is followed (or preceded) by a short pause, in order

not to overload the server. Websites often specify a delay in the *robot.txt*. In this document, information is saved on whether one is allowed to crawl and scrape the website and at what pace. For the UNIX forum, the delay was unspecified, so a delay of 5 seconds was used. Additionally, it is possible to crawl making concurrent requests. Simplified, this means crawling (and possibly scraping) through more pages at the same time. Given the relatively small size of the AEU forum, I did not deem this necessary.

3.2.2 Scraper

As mentioned above, all links that are collected by the crawler, are processed by the scraper. Scraping refers to the process of collecting data from a page. Here, only the posts were collected from the forum. More specifically, for every posts two things were stored: the timestamp, i.e., the moment that a post was placed, and the text of a post. No other information, such as usernames, was stored. Since only the timestamp and the text in the posts are relevant to this thesis, all other information was disregarded.[2] Evidently, the most important data are the textual data. The timestamp, on the other hand, was collected mostly in other to illustrate the span of activity on the forum and show the activity on the forum over the years (cf. section 3.4).

The way the scraper works, is as follows: it downloads a HTML-page and extracts data that is connected to certain HTML-tags. This was done using BeautifulSoup4 (BS4) (Richardson, 2007). This Python library was chosen over other possible extraction methods, because it is relatively easy to use and fast enough for the task at hand. Using BS4's functions and classes, the HTML-page can be search through, using HTML-tags and their attributes. Because these are not standardized, Google Chrome's inbuilt *Inspect*-function was used to find the appropriate tags. An example is given in (1). Here the tag is 'div', the attribute is 'class', and the second line is the text that will be extracted when BS4 is order to look for this tag. Finally, the tag is closed by the '/div' tag on the third line. HTML enables multiple embedded tags, so one must be concise in defining which tags to extract. For example, some of the *parent* tags are used for layout. Extracting these would slow down the process and they would possibly make the data processing more difficult.

(1)
"The text of a post."

Many of the posts still contain unwanted information. Firstly, since the general topic of the forum is software development, many of the posts include chunks of code. While very helpful to other users, it is not natural language and thus not useful for this thesis. Secondly, references to previous posts, or *quotes*, are also redundant information. Because the entire forum is scraped, *quotes* will refer to sentences that are already in the corpus. This would unfairly boost the occurrences of certain data, thus skewing the distribution. BS4 was used to exclude these parts of the

[2]Of course, a username or an entire user profile could give information on language background or an indication of different language use (influenced by factors such as age or gender). However, there are few ways to verify whether this information is accurate and it will thus be excluded from the analysis.

posts, leaving only the text. That text was stored with its timestamp. The text from all posts with the timestamps forms the Advanced and Expert Users-corpus (AEU corpus).

3.3 Pre-processing

The data still contains noise, even after picking an appropriate source, constraining the crawler, carefully picking the HTML-tags, and excluding redundant code and *quotes*. The raw data needs to be pre-processed in order to be useful for analysis. This starts for instance with removing newlines and removing superfluous word spacing. Further, the pre-processing happens in a series of steps, which is explained below.

Segmenting

The first step that needs to be taken is segmenting the posts into sentences. Because this thesis looks at the use of preposition within a sentence, it is not necessary to keep the surrounding sentences as context. This way, a large part of the collected data can be excluded from the analysis. By using a pre-trained version of Punkt (Kiss and Strunk, 2006), which is facilitated by NLTK (Bird et al., 2009), the posts are divided into sentences. Punkt is preferred over built-in Python functions that simply use a full stop to split text into sentences. Using Punkt is necessary, especially for the UNIX forum, because the full stop is also used for other purposes, such as in (2).

(2) 'I need to rename all the .txt files present in current directory to .dat files respectively in UNIX.'

Filter for code

The next step is taken to counteract problems like (3). Although the blocks of code that were tagged by the HTML-code were accounted for, many users still write short chunks of code or a directory path in their text, as in (3) and (4). This can be problematic for the use of an automatic parser or part-of-speech tagger, which are often pre-trained on natural language and do not work well on data containing lines of code. All sentences containing more than six characters that were not alphanumerical or punctuation were deleted. The sentences containing directory paths, like in (4), were not deleted, since these 'paths' mostly indicate a certain location, which is relevant for this thesis.

(3) 'I have seen this posting here: perl -e '@d=localtime ((stat(shift))[9]);'

(4) 'Change -I/usr/src/linux/include (or similar) to /usr/include/linux/modversions.h:4:2:'

Search for prepositions

Next, the newly created segments are searched for the prepositions relevant to this thesis, using regular expressions such as in (5)[3]. This was done in order to create a dataset that contained only sentences with the relevant prepositions, all other data were disregarded for the analysis. This results in a dataset of 51,877 sentences containing one or more of the relevant prepositions.

(5) '\b[Aa][Gg][Aa][Ii][Nn][Ss][Tt]\b'

Account for infinitival 'to'

Because the sentences were searched using regular expressions, the data still include all sentences that contain infinitival 'to', like in (6). For this reason, a part-of-speech tagger from the spaCy library (Honnibal and Johnson, 2015) was used to separate these sentences from the relevant data. This leaves 12,210 sentences.

(6) 'I don't know really which flavour of linux to use.'

Dependency Parser

At this stage, it would be useful to use a pre-trained dependency parser to find dependency relations within a sentence and extract only the relevant relations. However, this was not feasible for this thesis. An attempt was made using the Stanford Dependency parser (De Marneffe et al., 2006). Even after clean up, the data are still too noisy to get a decent performance from a parser. The presence of spelling errors, the directory paths, and many domain specific terms, among other factors, make the dependency parser perform poorly. Furthermore, since PP-attachment is a well known problem for dependency parsers in general, the output from the parser would be unreliable to consider for analysis. The poor performance would mean that the parses of all 12,210 sentences had to be manually checked. For that reason, a dependency parser was deemed not useful.

Part-of-speech tagger

Alternatively, one could use the part-of-speech tagger to extract all verbs and prepositions in a sentence. However, this is unhelpful for sentences containing multiple verbs. Furthermore, some of the verbs, such as 'access', have an ambiguous form. In other words, without context, these words can either be verbs or nouns. As mentioned above, the data are not fit for the dependency parser. Likewise, it lowers the accuracy of a part-of-speech tagger too, which can lead to overlooking relevant data with these ambiguous verbs. Finally, in order to get a more complete impression or the distribution of the prepositions, one must not only look at the verbs, but also

[3]The regular expressions were created this way in order to account for any combination of lowercase and uppercase use.

at the Figure and Ground of the spatial expression. A part-of-speech tagger does not annotate relations between the parts-of-speech, which makes it impossible to distinguish what is the direct object and what is the prepositional object.

Verb list

Because the use of a dependency parser is not feasible for this thesis, the dataset needs to be downsized to a size that is small enough to go through manually. Therefore, I manually made a list of potentially interesting verbs. These verbs were picked during the initial data exploration. When a verb seemed to be used differently, or already occurred with one of the prepositions, it was added to the list. These verbs were searched for in the AEU corpus. Some of these verbs turned out to be irrelevant, these were cast aside. Then, extra verbs were added to the list, again by hand-picking verbs during the data exploration. This way, the current verb list, which is presented in Table 3.1, was put together.

It shows that some of the verbs are abbreviations that are specific to the software domain (such as 'FTP' and 'SSH'). Other verbs in Table 3.1 are common English words (such as 'copy' and 'mount'). Other verbs already seem highly domain specific, such as hard-code and synchronise. I will return to the frequencies in Table 3.1 in section 3.4.2.

The sentences were searched for these verbs in present and past tense, and with perfective and continuous aspect (loop, loops, looped, looping). This leaves 3,528 sentences containing both one of the verbs listed in Table 3.1 and one of the prepositions in that order, because, as chapter 2 described, only one particular sentence structure was considered.

Verb	Count	Verb	Count	Verb	Count	Verb	Count
connect	300	transfer	49	execute	25	program	5
write	183	install	44	authenticate	22	reboot	5
copy	133	pipe	39	link	22	hard-code	4
log	131	boot	39	upload	17	sync	4
login	96	append	36	type	16	iterate	3
convert	90	access	32	merge	13	sftp	3
redirect	78	cd	29	download	12	format	3
run	75	mount	28	compile	11	duplicate	3
load	65	loop	27	plug	10	synchronise	3
ssh	51	create	27	tar	9		
ftp	49	telnet	26	echo	8		

Table 3.1: The verblist included their counts in the manually selected subset.

Manual assessment

The remaining 3,528 sentences were filtered manually, which resulted in a total of 1,825 sentences. The most important criterion the sentences had to be checked for, was PP-attachment. In many of the sentences, the PP was not attached to the verb from the verblist in Table 3.1, like in (7a). Here, the PP 'across systems' is attached to the verb 'share', therefore this sentence is not useful for analysis. Similarly, some sentences were ambiguous, like (7b). When it was unclear whether the PP ('across the net') was attached to the verb ('syncing') or to the direct object ('some very large filesystems'), the sentence was omitted. Further, for some sentences it is unclear to which verb the PP is attached, like in (7c). These sentences were omitted as well. Lastly, the preposition is not always used with a spatial meaning, like in (7d). These occurrences were also omitted.

(7) a. 'NIS and YP (yellow pages) were created to **share** values across systems.'
b. 'I'm syncing some very large filesystems across the net with rsync'
c. 'the system understandably gets a little 'confused' and will mount/boot to a RAW device'
d. 'His command will re-create the tar file over and over again with the subsequent file.'

For all the remaining sentences, each sentence was annotated as such: 'everything left of the verb' (Left), verb, direct object, and different types of *path*, in order to make possible patterns easier to recognize. Every part of the sentence was put in the corresponding column. The paths were annotated in such a way that the preposition and the prepositional object were in separate columns. Table 3.2 shows the annotation for one of the sentences. As the Table shows, some of the sentences also include the preposition 'from', which express a *source*. This was annotated as well. When the sentence did not contain, for instance, a direct object, the corresponding cell remained empty.

Left	Verb	DO	RouteP	PO	SourceP	PO	GoalP	PO
I can't	copy	these files	via	/tmp	from	my home	to	the mount point

Table 3.2: The annotation of the sentence 'I can't copy these files via /tmp from my home to the mount point'.

3.4 Corpus and subset

Using the data collection method described above, the Advanced and Expert Users corpus (AEU corpus) was constructed containing all posts on the AEU forum, until 02-01-2019. Before analysing the data that are relevant, I will first show that the AEU corpus is useful, as it reflects natural data. In subsection 3.4.1, the global statistics of the corpus will be presented. In subsection 3.4.2, I will go into the manually selected subset.

3.4.1 AEU Corpus

This section shows that the AEU corpus does not show a skewed distribution over the years, that it superficially seems to resemble natural language, and that it uses prepositional phrases with a similar frequency to a subset of the British National Corpus (BNC), a corpus designed to represent British English.

Firstly, during the scraping of the AEU forum, both the timestamps and text of all posts were collected. Table 3.3 shows the amount of posts per year, the amount of sentences per year, and the amount of sentences per post. Then it shows the same factors for a subset of the data, which consists of only sentences that contain one of the verbs in Table 3.1 and one of the prepositions, excluding infinitival 'to'. This subset is henceforth called the PrepSet, sentences from the PrepSet will be called PrepSet sentences. The Table also shows the ratio of PrepSet sentences over all sentences, in percentages.

Year	Total posts	Sentences	Sent./post	Prep. sent.	Prep. sent./post	Percentage
2001	889	3744	4.21	361	0.41	9.64 %
2002	2335	10422	4.46	959	0.41	9.20 %
2003	1481	6604	4.46	630	0.43	9.54 %
2004	1052	4300	4.09	409	0.39	9.51 %
2005	2123	8035	3.78	714	0.34	8.89 %
2006	3030	9961	3.29	829	0.27	8.32 %
2007	3473	11960	3.44	972	0.28	8.13 %
2008	5365	18298	3.41	1560	0.29	8.53 %
2009	3438	13366	3.89	1165	0.34	8.72 %
2010	2886	10216	3.54	869	0.30	8.51 %
2011	3139	11742	3.74	1014	0.32	8.64 %
2012	2279	8787	3.86	760	0.33	8.65 %
2013	1625	6084	3.74	585	0.36	9.62 %
2014	1208	5042	4.17	465	0.38	9.22 %
2015	758	3093	4.08	239	0.32	7.73 %
2016	562	2357	4.19	213	0.38	9.04 %
2017	462	2036	4.41	215	0.47	10.56 %
2018	408	1719	4.21	169	0.41	9.83 %
2019	187	877	4.69	82	0.44	9.35 %
Total	36700	138643		12210		

Table 3.3: The amount of posts, sentences, and PrepSet sentences per year. Column four and six shows the amount of sentences per post and the amount of PrepSet sentences per post, respectively, and column seven shows the percentage of sentences in the PrepSet over amount in the entire corpus.

The first column 'Total posts' shows that the amount of posts increased since the first posts

in 2001, reaching its peak in 2008. From 2008 on, the user activity decreased again. Similarly, the amount of sentences increases and decreases along the same line, with 2002 as the exception. However, the column 'Sent./post' shows that the amount of sentences per post takes the opposite distribution, where the amount is higher than 4 sentences per post in the first 4 years and the last 6 years, in between it is lower than 4 sentences per post. This indicates that, even though there are more posts in the middle years, the posts by average were longer in the first and last years of the AEU corpus. The same pattern is seen in the columns 'Prep. sent.' and 'Prep. sent./post', showing the amount of PrepSet sentences and the amount of PrepSet sentences per post. The amount of PrepSet sentences increases until 2008, with again 2002 being the exception, and then gradually decreases. The amount of PrepSet sentences per post again shows the opposite distribution.

The amount of sentences per post ranges from 3.29 (in 2006) to 4.69 (in 2019), this shows no great fluctuations. Similarly, the amount of PrepSet sentences per post ranges from 0.27 (in 2006) to 0.47 (in 2017). The last column, called 'Percentage', shows that the amount of PrepSet sentences over the amount of sentences ranges from 7.73% (2015) to 10.56% (2017). These ratios do not show great differences and are quite evenly distributed. This indicates that the AEU corpus shows a more or less even distribution of PrepSet sentences over the years, meaning that it is a healthy dataset to use for analysis.

Secondly, Table 3.4 shows the amount of sentences and words of all the posts in the AEU corpus, and it shows the size of the Prepset. It also shows that the PrepSet makes up 8.81% of the sentences in the entire dataset, and the amount of words makes up 11.66%. By average, the sentence length in the AEU corpus is 12 words per sentence, whereas the sentence length in the PrepSet is 17 words per sentence. This shows that the sentences are sufficiently long to resemble natural language. Additionally, the average is quite high, indicating that the users or the AEU forum elaborate extensively. This was expected, according to the characterization of the forum and its users given in section 3.1.2.

	Sentences	Words
AEU corpus	138,643	1,709,664
PrepSet	12,210	204,309
Percentage	8.81 %	11.66 %

Table 3.4: Amount of sentences and words for the AEU corpus and for the PrepSet.

Thirdly, Table 3.5 shows the counts for the individual prepositions in the AEU corpus. On the right, the frequencies of the prepositions in the BNC are given, based on Leech et al. (2014). They automatically extracted prepositions from a 1 million word subset of the BNC. Because their subset of the BNC is smaller than the dataset used here, Table 3.5 also shows the frequencies for the subcorpus per 1 million words. This makes the number of occurrences in both corpora more comparable.

This comparison shows that the overall frequency of prepositions is lower in the PrepSet. However, the individual prepositions seem to pattern similarly. This holds for all prepositions

except ‘via’ and ‘against’. ‘Via’ occurs far more in the PrepSet than in the BNC subset. ‘Against’ on the other hand occurs far more often in the BNC subset than in the PrepSet. Note that ‘to’ was omitted from Table 3.5, because the pos-tagger may not tag occurrences of ‘to’ accurately. The cause of the high frequency of ‘via’ in the PrepSet remains unknown at this point. However, given that the distribution in the PrepSet and the BNC pattern similarly, this indicates that the data resembles natural language regarding the frequency of prepositional phrases.

PrepSet			Leech et al. (2014)	
Preposition	Count	Count/mil.	Preposition	Count
into	2540	1486	into	1634
through	1154	675	through	743
over	896	524	over	735
via	717	419	against	562
across	228	133	across	217
against	203	119	onto	62
onto	150	88	via	45
	5888	3444		3998

Table 3.5: Amount of sentences per preposition in the PrepSet (excluding ‘to’) and Leech et al. (2014).

Tables 3.3, 3.4, and 3.5 demonstrate that the AEU corpus does not show a skewed distribution over the years, that it superficially seems to resemble natural language, and that it uses prepositional phrases with a similar frequency to the BNC subset. The PrepSet shows the same patterns in the distribution over the years and, although having somewhat longer sentences by average, also seems to make up a reasonable portion of the entire dataset. Therefore, I assume that the data in the dataset and in the PrepSet resemble natural language and are suitable for linguistic analysis.

3.4.2 Manually selected subset

As mentioned in section 3.3, the pre-processing left around 3,500 sentences. These had to be filtered manually, leaving a total of 1,927 *goal-paths* or *routes*, divided over 1,825 sentences, since some sentences contain more prepositions. These sentences included at least one *path*. This subset is henceforth referred to as the Manually Selected Subset (MSS).

The sentences in the MSS are occurrences were the verb and preposition are used with a meaning specific to the language domain of software development. Table 3.6 shows the distribution of the prepositions in the MSS. It shows that ‘to’ and ‘into’ are highly frequent, ‘through’ and ‘via’ are less frequent, and ‘over’, ‘onto’, ‘against’, and ‘across’ are the least frequent. Table 3.6 also shows with how many different verbs each preposition occurs with. Similarly, Table 3.1 shows the frequencies for the verbs.

Preposition	Count	Verbs
to	876	35
into	536	37
through	195	27
via	180	20
over	53	18
onto	40	13
against	34	6
across	13	6

Table 3.6: Frequency of prepositions in the MSS.

In (8) some examples are shown of the sentences taken from the MSS. Firstly, example (8a) contains a preposition expressing a *goal*, namely 'against'. Additionally, (8b) shows another preposition expressing *goal* ('into'), but also 'from', which expresses a *source path*. Secondly, (8c) and (8d) show examples with a preposition associated with *routes*, 'over' and 'via', where (8d) also includes a *source path*. Lastly, (8e) shows both the prepositions 'via' and the preposition 'onto'. Appendix A shows additional examples, with different verbs and prepositions.

(8) a. 'I am looking to have unix authenticate against active directory'
b. 'copy the file from /lost+found directory into the correct path'
c. 'I want to use a shell-script to transfer data over sftp'
d. 'I inherit a solaris7 system with /home mounted via a share from the nfs server'
e. 'it can be mounted via nfs onto the servers'

The 42 verbs combined with the 8 prepositions that are examined in this thesis, combined into 162 verb-preposition pairs in the MSS, out of 336 possible combinations. For the analysis, in chapter 4, most verb-preposition pairs were considered. Only pairs that had a very low frequency were disregarded, because the occurrence of the pair might be coincidental. The verb-preposition pairs that occurred more than 10 times are shown in Appendix B, to give an impression of the frequencies. It shows the amount of times that a pair has occurred, as well as the amount of times that the verb occurs in the MSS. The Table shows that some verbs occur primarily with one or two prepositions, such as 'copy', while others are more divided amongst several different prepositions, such as 'run' with 'through', 'via', 'against', and 'to'.

Chapter 4

Analysis

4.1 Goal-paths

For *goal-paths*, the prepositions describe a path, where the Ground functions as endpoint. Depending on the used preposition, the endpoint can be the Ground (‘to’), inside the Ground (‘into) or in contact with the Ground (‘onto’, ‘against’). In this section, I will discuss the use of the prepositions ‘to’, ‘into’, ‘onto’, and ‘against’ individually and then generalize whether the spatial semantic notions apply to *goal-paths* in *cyberspace*. In each subsection, the frequencies are given for the verbs that are mentioned in the subsection. Horizontal lines in the tables group together verbs that are related.

4.1.1 To

By far the most frequently occurring preposition ‘to’ is used with 35 different verbs. Jackendoff (1983) approach to the preposition ‘to’ is the TO-function with a *thing* as argument (instead of a PLACE as argument). As mentioned earlier, the function does not lay any restrictions on what can be an argument and what cannot, as long as it is a *thing*. While this works for many of the sentences in the subcorpus, the examples in (1) show that this approach is also problematic. In (1a), the Figure gets downloaded to a location at a certain file path. This path refers to a certain location on (most probably) someone’s computer. Similarly, example (1b) states a directory, which is again a location. Admittedly, it is possible that the directory is considered a *thing*, rather than a *place*. Lastly, example (1c) explicitly says ‘place’ in the prepositional object. These, and other, examples show that Jackendoff (1983)’s approach does not cover the full extent of data.

(1) a. ‘They get downloaded to /var/spool/up2date’
    b. ‘You cannot link a file to a directory’
    c. ‘It will likely not write to the same place’

An alternative approach comes from Beavers (2008), who described that ‘to’ can only occur with durative events. In other words, the events described with ‘to’ have to be gradable. However,

this is already problematic for the most frequently occurring verb 'connect', but also for 'append', 'log', and other verbs. The occurrence with 'to', implies that 'connect' is a durative, gradable event, according to Beavers (2008). The example in (2) shows the Beavers (2008)'s test, mentioned in chapter 2, to distinguish between gradable and non-gradable events.

(2) 'The computer connected to the network (instantly).'

Beavers (2008) would predict that the use of 'instantly' in (2) is incoherent. However, it is not. Moreover, there is a bigger issue. This test does not seem to make sense for this particular domain. While it works for human movement, where an event takes time because of physical limitations, it does not work for computers, because their processing time for simple tasks can be so short that they are perceived by humans as instantaneous.

One could postulate that most computer processes do take processing time, but that the processing time can be too short for humans to notice. Some processes can be made more explicit, by applying them to larger files. For instance, the process of 'downloading a file' is mostly depended on the transfer speed and the size of the file. Simplified, with a connection speed of 100 bytes per second, a file of the size 1 byte takes 0.01 seconds, appearing to be instantaneous. The same process with a file of size 1,000 bytes, would take 10 seconds, appearing to be durative. Many computer processes only appear to be instant, because they are applied to small tasks, which makes the Beavers (2008)'s test not usable for these data.

There is another problem for Beavers (2008)'s approach that emerges with the verb 'connect', but also for verbs like 'append'. The problem is that these verbs do not have a 'middle state', which makes them non-gradable. In other words, something is connected, or not connected, but it cannot be halfway connected. So even without the test, the approach made by Beavers (2008) does not hold for these data. Temporarily, there is a phase where the computer is 'connecting', but practically the computer is only able to use the connection once the connection is made. Since 'to' can only occur with durative events and connecting in this domain is not durative, Beavers (2008)'s approach does not work for these data.

A third approach is more pragmatically motivated. It is possible that the speaker considers the prepositional object underspecified. In other words, because the user cannot perceive the spatial properties of an object, they use the preposition 'to' to specify the path, but remain neutral to other properties of the object. In the cases where the spatial properties are perceivable, the speaker will choose the appropriate prepositions. This approach also gives a possible explanation for the variation in prepositional use, namely that a verb occurs with various prepositions instead of just one. Different speakers have different experience with the prepositional objects and thus different perception of the spatial properties of the prepositional objects, such as whether the object has a surface or not. Therefore, some speakers would prefer 'to' within a certain prepositional object, while others would use the more specified 'into' or 'onto'. Moreover, this also implies that certain objects should be referred to with either 'to' and 'into' or 'to' and 'onto'. I will return to the variation and the preferences for either 'to' and 'into' or 'to' and 'onto' in subsection 4.1.5.

One important issue with verbs like 'connect', such as 'link', 'ssh', and 'telnet', is that it does not necessarily refer to some sort of movement, but rather to the establishment of a connection.

In English, this does not make a difference in the form of the preposition. However, in other languages, like Dutch, verbs like transitive 'connect' would use a different preposition, than verbs that describe a movement, compare (3)[1] and (4).[2] However, (5) shows that this does not hold for 'telnet' in Dutch.[3]

(3) Koppel de interface aan een netwerk.
connect the interface p.to a network.
'Connect the interface to a network.'

(4) Bill brengt zijn kind naar de speeltuin.
Bill brings his kid p.to the playground.
'Bill brings his kid to the playground.'

(5) Probeer [...] te telnet -ten naar uw webserver.
Try [...] inf.to telnet PL p.to your web server.
'Try to telnet to your web server.'

Verb	DO	Source	Count
connect	18	9	247
link	2	0	18
dowload	3	2	7
ftp	28	11	40
transfer	21	10	27
write	37	0	130
log	4	0	19
append	18	0	28
telnet	0	6	16
ssh	0	5	16

Table 4.1: The verb frequencies for 'to', including the frequencies for direct object (DO) and *source-path*.

4.1.2 Into

The preposition 'into' is probably the most interesting, while it is highly frequent, but also several interesting patterns. The preposition describes a *path* that ends in a Ground that is able to contain the Figure (Landau and Jackendoff, 1993, Jackendoff, 2012), which means that the prepositional

[1] p.to = prepositional 'to', inf.to = infinitival 'to', PL = plural.
[2] Intransitive connect would be translated with *verbinding maken met* 'to make a connection with'.
[3] Retrieved from https://forum.ubuntu-nl.org/

object in some way reflects this property. One verb that makes for a clear example is 'redirect' or 'pipe'. The Figure of this verb is almost exclusively 'output' of some code or algorithm, which then serves as input for the Ground (mostly a file or another algorithm), as in example (6a). Similarly in (6b), 'load' takes mostly 'data' or 'files' as a Figure and moves that into a Ground, which is mostly a database or a data structure (such as an array or table).

(6) a. 'this will redirect the output into a logfile.'
b. 'I would still load the data into the database.'

Secondly, some verbs turned out to describe a change of state, rather than a change of location, namely 'boot', 'format', and 'convert'. As the examples in (7) show, these verbs change the Figure into the Ground, while not necessarily translocating. The Ground in these sentences is often a certain mode (for 'boot') or a format (for 'convert' and 'format'). On the other hand, there are verbs that are in between state and location, such as 'cd' (Change Directory) and 'merge'. The verb 'cd' is used consistently without a direct object and with 'directory' as Ground, since it changes 'directory' into another one. It is unclear whether the Figure changes location or state. Similarly, 'merge' is used when two objects are put together, however it is unclear whether the content of two objects are put together into a new object, as in (8b), or one object is added to the other, as in (8c). The latter form of the verb 'merge' is also described by Levin (1993), however only with the preposition 'with'. One thing that stands out that 'convert' almost always occurs with a direct object, whereas the related verbs 'boot', 'format', 'CD', and 'merge' do not.

These verbs can be accounted for by the Thematic Relations Hypothesis (Gruber, 1965, Jackendoff, 1983). Concretely, this hypothesis imposes that other semantic fields make use of *paths*. So, while these sentences do not have a spatial reference, there is still a *path*, in this case to express *identification* (Gruber, 1965). Jackendoff (1983)

(7) a. 'It keeps booting into single user mode.'
b. 'One of sprintf's jobs is to convert binary into ASCII.'
c. 'I'm having trouble formatting my data into mllp format.'

(8) a. 'need to do above + cd into a dir'
b. 'I am looking to merge a group of lines into single line.'
c. 'They are using many log files and from time to time they merge them all into one'

Thirdly, there are verbs that describe making a remote connection, such as 'ssh' and 'telnet'. Additionally, there are verbs that transfer entities over a connection, such as 'ftp' and 'transfer'. While these two types verbs are related, the difference between them is reflected in the data, as the former two verbs never take a direct object and the latter two do in most occurrences. Next to that, 'ssh' and 'telnet' occur relatively more often with 'into' than 'ftp' and 'transfer', which occur relatively more often with 'to'. All four verbs take the same types of prepositional objects, which are a server, (operating) system, or computer 'box'. The use of 'into' with 'ssh' and 'telnet'

Verb	DO	Source	Count
redirect	13	0	15
pipe	10	1	15
load	41	2	52
boot	3	3	31
format	2	0	2
convert	35	1	43
CD	0	0	17
merge	10	0	12
ssh	0	5	33
telnet	0	1	10
ftp	2	0	7
transfer	4	0	6
log	4	3	101
login	0	1	15

Table 4.2: The verb frequencies for 'into', including the frequencies for direct object (DO) and *source-path*.

is therefore probably best explained by the secure connection, which allows only the user to gain access into a certain environment.

Lastly, there are other verbs that express accessing an entity, such as 'access', 'log', and 'login'. However, there are several issues here. Firstly, the distribution of 'access' is very different then the distribution of 'log' and 'login'. It is only seen with 'into' when it is used as a noun, rather than a verb, as in example (9a). Therefore, it is not relevant for this section.

Secondly, the verb 'login' is derived from the noun 'login', whereas 'log in' is the appropriate verb, according to Cambridge Dictionary. Companies like Microsoft and Apple have specific guidelines for this as well, so in software forums for other operation systems, this problem might be non-existend, even though these guidelines are focused on official documentation. Nevertheless, users use 'login' as a verb, even in combination with the preposition 'into', leading to sentences such as (9b), whereas 'log in into' almost never occurs in the data. For that reason, I choose to keep 'log in' and 'login' as two separate verbs, rather than merging them.

Thirdly, the verb 'log' is used in two ways. On the one hand, there is the sense of keeping records and, on the other hand, there is the sense of accessing system. While the latter sense is probably derived from the former, they are distinct in two ways. One, the first sense of log is used more with 'to' and takes prepositional objects that are files or 'history logs'. The occurrence of 'log' with 'to' patterns nicely with 'write', which carries a similar meaning in this domain and also occurs predominantly with 'to'. The second sense of log is used more with 'into' and takes prepositional objects that are a computer system or server. Because the environment is only accessible by using a login, 'logging in' allows the user to enter a container. The different senses

are shown in examples (9c) and (9d).

(9) a. 'That would allow folks access into your server.'
b. 'When I login into cde.'
c. 'It should not be logged into the history.'
d. 'How many users are logged into this box?'

4.1.3 Onto

The preposition 'onto' occurs almost exclusively with prepositional objects that are a system or a server, or a physical storage, such as a USB or DVD. This happens with verbs like 'copy' and 'mount'. All verb-preposition pairs with 'onto' are relatively infrequent. Like with 'into', both 'log' and 'login' are used with 'onto', as in examples (10a) and (10b). Interestingly, only one of these occurrences happened after 2007. This is relevant, because, as mentioned above, there are writing style guides published by companies like Apple and Microsoft, with specific guidelines on this. Up until 2012, Microsoft was suggesting to only use 'sign in' or 'log on' (*Microsoft Manual of Style* 2012, p.329), while Apple in 2003 already suggested the opposite, preferring 'log in to' (*Apple Publications Style Guide* 2003, p.87). Since, to my knowledge, there are no official guidelines on Unix or related systems, the occurrences of 'log onto' might be related to users that also work on Windows computers.

(10) a. 'How many people are logged onto an AIX system?'
b. 'I use netterm to remotely login onto a Linux sever'

As explained in chapter 2, 'on' and 'onto' are used when the Ground has a surface on which the Figure can be placed (Talmy, 1983, Landau and Jackendoff, 1993, Jackendoff, 2012). However, the use in these data does not seem to be based on any visible surface. One possible explanation could be that it is derived from the physical aspects of the computer, where something is shown on the screen, the visible surface of the computer, or written onto a disk, where the surface is manipulated to represent certain data. This explanation would account for the relatively low frequency, because it implies that 'on' and 'onto' are more hardware related, which is not a prominent topic on the AEU forum.

Verb	DO	Source	Count
log	0	0	6
login	1	0	3
copy	7	0	8
mount	6	0	7

Table 4.3: The verb frequencies for 'onto', including the frequencies for direct object (DO) and *source-path*.

4.1.4 Against

Interestingly, the preposition 'against' is used in almost all occurrences of the verb 'authenticate', like in example (11a). Additionally, it is mostly used with the prepositional object 'Active Directory', which is a service used to connect multiple devices. It is most likely that 'authenticate against' is derived from 'verify/check against', which is also used outside this domain.

(11) a. 'I am looking to have Unix authenticate against Active Directory.'
b. 'You cannot dynamically link a Solaris shared library against a Linux library.'
c. 'You wish use a for loop script to run it against each server'

On the other hand, with 'link', as in (11b), the use of 'against' implies making several connections between the Figure and Ground. This does imply contact between the Figure and Ground, which is in line with Jackendoff and Landau (1991). Similarly, 'run against' also implies contact, which is shown in (11c). Interestingly, the direct object for 'run against' is almost always a Unix command.

Chapter 2 showed that 'against' is static, which imposes that it can only occur with a *path* when it is used with a motion verb. As described in chapter 2, Lakoff (1966) has a series of tests to test whether a verb is stative or not, here repeated in (12). Not all these constructions were resembled in the data. However, (13) shows that the verbs that occur with 'against' can be used with progressive aspect, whereas stative verbs cannot. Further, outside this domain, these verbs also show grammatical sentences with the constructions in (12b) and (12c). In (14) and (15) these tests were done for the verbs outside the *cyber* domain.

(12) a. I am running/*knowing the answer.
b. What I did was run/*know the answer.
c. I persuaded John to run/*know the answer.

(13) a. I'm authenticating against windows server 2012 r2.
b. it says i am linking against different.
c. I'm writing a shell script that will be running against several servers.

(14) a. What I did was authenticate the painting.
b. What I did was link her to the crime.
c. What I did was run to the door.

(15) a. I persuaded John to authenticate the painting.
b. I persuaded John to link her to the crime.
c. I persuaded John to run to the door.

Given that the verbs occur with progressive aspect in the data and given that the verbs outside this domain are also non-stative, I assume that these verbs are not stative. One can also reason

that they are not positional verbs, as they do not denote a 'posture' and they do not denote a spatial disposition with respect to the ground (Evans, 2014). Therefore, the verbs are *dynamic*. This is in line with the Denis et al. (2003), as the *static* 'against' only occurs with motion verbs.

Verb	DO	Source	Count
authenticate	2	0	16
link	1	0	2
run	11	0	13

Table 4.4: The verb frequencies for 'against', including the frequencies for direct object (DO) and *source-path*.

4.1.5 Goal-paths revisited

Most data including *goal-paths* can be accounted for Jackendoff (1983). One important assumption must be made, in order to cover the overlap between the prepositions 'to' on the one hand, and 'into' and 'onto' on the other hand. This assumption is that many prepositional objects are underspecified for certain spatial properties, which makes speakers use the more *neutral* preposition 'to'. Here, *neutral* refers not making demands of the preposition object. In other words, a more *neutral* preposition has less restrictions than a less *neutral preposition*. In that sense, 'to' is more *neutral* than 'into' and 'onto'. This also allows variation between speakers, because one speaker can have more experience with or a different perception of certain objects. The variation is shown in (16) and (17). The examples in (16) show sentence pairs with 'to' and 'into' with a very similar meaning. Here the verb and prepositional object is the same, but the prepositions are different.

(16) a. '[it] is ssh'd to server b'
b. 'and I was able to ssh into server f'
c. 'multi-line commands are converted to single line'
d. 'The memo field converted into single line'
e. 'I was able to cd into the directory'
f. 'and you can't cd to the directory'

The data for 'onto' are more sparse, but the example pairs in (17) are again are very similar. The last pair, in (17e) and (17f), has a similar direct object, which is a sequence of characters (italicized in the example), that is added to the Ground ('a file').

(17) a. 'having placed the order the file will be downloaded onto the unix server'
b. 'if it is a script, is it downloaded to your unix server already'
c. 'read-only-filesystem means it cannot write to that partition in any way'

d. ‘write the files from where you deleted the file onto the same partition’

e. ‘i suggested appending *the tail end of xxx.log.0* onto xxx.log’

f. ‘since you would not want multiple processes appending *lines* to the same file’

I mentioned that this approach implied that there would still be a distinction between the prepositional objects that occur with ‘into’ and those that occur with ‘onto’. This is borne out partially, since ‘onto’ is almost only used with some form of system or storage unit (either physical or not), whereas ‘into’ selects for a wider variety of objects, but also non-physical forms of systems and storage units. While many prepositional objects occurred only with ‘into’ (and not with ‘onto’), only physical storage units, such as ‘floppy’ or ‘disk’, seem to be more exclusive to ‘onto’.

Additionally, many sentences also contain a *source-path*, which is indicated by the preposition ‘from’. Evidently, these *source-paths* are not obligatory. The presence of a *source-path* clearly shows that there is a *path* present in these sentences. This is crucial, because the presence of both a *source* and *goal* strengthens the claim that speakers in this domain use spatial relations similar to the use in Standard English. There are 24 verbs with a *goal-path* that also occur with a *source-path*, however never consistently.

One last note must be placed with the verb ‘copy’. The verb-preposition pair ‘copy into’ already exists outside *cyberspace*. This is shown in example (18).[4] However, it is more likely that the verb-preposition pair in this domain is derived from ‘copy + paste’, where ‘paste’ is omitted.

(18) ‘this she copied into the front of each of the eighteen notebooks’

Pragmatically, this omission makes sense, since only ‘copying’ without ‘pasting’ is useless. Additionally, the important part of this process is making a duplicate. Therefore, when someone just uses ‘copy’, they are perfectly informative, since the ‘paste’ is implied. Using only ‘paste’ would not suffice to carry across the same meaning. Although both alternations (with and without ‘paste’) originally occurred in the corpus, only instances where copy was used on its own were maintained. This was done, because it is difficult to say with which verb the preposition forms a pair, if both verbs are present.

4.2 Routes

The term *route* is used for a spatial relation where the Figure follows a *path* and at a certain point is nearby or inside the Ground. Like in section 4.1, I will first give an analysis on the individual prepositions (‘over’, ‘through’, ‘via’, and ‘across’), and then make a generalization for this type of *paths* in subsection 4.2.5.

[4]Retrieved from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

4.2.1 Over

The preposition 'over' does not occur more frequently with one verb in particular. However, it does select prepositional objects are primarily networks, servers or similar connections between two systems. The preposition is used to clarify remote connection, when used with verbs like 'connect', 'log', 'login', 'install', 'mount', and 'transfer', as in (19a) and (19b). Additionally, verbs like 'loop' and 'iterate' take any kind of prepositional object that is iterable, i.e. consist of multiple elements a program can go through, like in (20a) and (20b).

(19) a. 'I also have another windows xp machine that connects over wireless connection.'

b. 'I have a printer installed over ADSL.'

(20) a. 'They loop over the lines in a files.'

b. 'you can use a for loop and use wget and iterate over the files.'

Another pattern that emerges is the use of 'over' as a particle. As described in subsection 4.1.2, 'ssh' and 'telnet' behave differently from 'transfer' and 'ftp'. This behaviour is also reflected in using 'over' as a particle. While both 'transfer' and 'ftp' show several co-occurrences with the particle 'over', as in (21), 'ssh' and 'telnet' do not. The particle 'over' can be used as a diagnostics test to distinguish between verb classes. However, more relevant here, the verb 'copy' also uses the particle 'over'. This is particularly interesting, since 'copy and paste over' does not occur in the entire dataset. This implies that either 'copy into' does not derive from 'copy+paste into', or that there are two different senses of 'copy'.

(21) a. 'I ftp'd the file over to another server.

b. 'She tells me that i can just transfer it over'

Verb	DO	Source	Count
connect	0	1	8
log	2	0	2
login	0	0	4
install	5	0	7
mount	2	0	3
transfer	3	0	3
loop	1	0	4

Table 4.5: The verb frequencies for 'over', including the frequencies for direct object (DO) and *source-path*.

4.2.2 Through

There are several different ways how 'through' is used in the data. Firstly, when used with 'pipe', it behaves like Jackendoff (1983)'s VIA + IN, where the object translocates via the interior of the ground. Example (22a) shows this, where 'this' refers back to the output of a previous process. This is very similar to the use of 'transfer' in (22b). However, with the verb 'loop', the preposition behaves very similar to 'over'. This is similar to the example about 'going through the slides' in chapter 2.

(22) a. 'Is there a way of piping this through another filter?'
b. 'when I am trying to transfer a file through a program (where I call the script).'
c. 'Try looping through each list element.'

(23) a. 'However when I execute this command through PHP command (...)'
b. 'When the same code is run through crontab.'

The prepositional objects of 'through' are often programs, commands, scripts or networks, depending on the verb. Like in (23), the verbs 'run' and 'execute', for instance, take commands and code as the prepositional object, with a direct object that often implies some form of data.

The difference between the instrumental use and the spatial use of 'through' could be made clear by replacing the preposition with 'via' (for spatial) or with 'with' (for instrumental). While this replacement makes a clear difference between the two, in software development, it is often unclear what role the Ground plays. Like with the verb 'connect' in (24). Here, in (24a), 'through' can be replaced by 'via', making it more spatial. Likewise, in (24b) 'through' can be replaced by the preposition 'with', making this sentence more instrumental. Furthermore, the two roles might not be mutually exclusive.

(24) a. 'one of our requirement was to connect through sftp connection.'
b. 'What if host A now needs to connect through a script?'

Verb	DO	Source	Count
pipe	7	0	11
execute	11	0	15
run	15	0	29
loop	0	0	21

Table 4.6: The verb frequencies for 'through', including the frequencies for direct object (DO) and *source-path*.

4.2.3 Via

There are four verbs that occur alongside 'via' the most, which are 'connect', 'login', 'run' and 'access'. These four verbs also occur frequently with 'through'. The prepositional objects for 'via' are also very similar to the prepositional objects that are used with 'through'.

This is can be explained by the same approach as with 'to', in subsection 4.1.1. First, let me return to chapter 2, where I explained that 'via' can be seen as VIA + AT, in line with Jackendoff (1983), where AT expresses that the Figure is proximate or interior to the Ground. Secondly, 'through' can be seen as VIA + IN. The approach taken for 'to' and 'into'/'onto' in section 4.1.5 can be applied here as well. When the prepositional object is underspecified for spatial properties, the speaker uses a more neutral preposition. The main difference between 'through' and 'via', is the difference between IN and AT. Here, AT would be the more neutral preposition, because is has fewer restrictions on the prepositional object.

This is initially not reflected in the data. There are some verbs, like 'access', 'copy', 'install', and 'connect' are used more often with 'via' than with 'through'. Verbs like 'execute', 'run', and 'pipe' are used more with 'through'. However, these verbs almost exclusively occur with commands or code as their prepositional object. It is unclear why there is a preference for 'through' with these prepositional objects (commands/code). Further research might give more insight into this issue. The other verbs do not show a great difference in frequencies between the two prepositions, so it is impossible to show a preference for these verbs. Taken into account that commands and code are used more with 'through', the approach holds for the other verbs where there is no preference or a preference for 'via'.

Verb	DO	Source	Count
connect	7	2	63
access	17	1	19
login		1	18
pipe *	1	0	1
execute	3	0	5
run	13	0	17

Table 4.7: The verb frequencies for 'via', including the frequencies for direct object (DO) and *source-path*. *Pipe is not considered in this section because its frequency is too low, however it is displayed here for comparison with 'through'.

4.2.4 Across

The least occurring preposition in the data is 'across'. Like with 'against', I established that 'across' is a *static* preposition. The verbs that occurred with 'across' ('copy', 'create', 'write') are used with progressive aspect in the data, although not with the preposition 'against'. Furthermore, the verbs are non-stative outside this domain, given the tests by Lakoff (1966), and not positional (Evans,

2014)

As mentioned in chapter 2, Xu and Badler (2000) classified two different trajectories for 'across', namely inside and outside the ground. The data suggests that 'across' is used in this domain with the latter reading. The preposition is used to clarify that an object is being moved from one side of a connection to the other side, like in (25). Like with 'over', it occurs when a speaker describes sending an entity to a remote location.

(25) a. 'You have to be careful not to just write everything across the network'
b. 'This environment would not allow you to copy data across a network.'

Verb	DO	Source	Count
copy	3	0	5
create	2	0	2
write	1	0	3

Table 4.8: The verb frequencies for 'across', including the frequencies for direct object (DO) and *source-path*.

4.2.5 Routes revisited

Chapter 2 described that the prepositions associated with *routes* are ambiguous. This is partially reflected in the data. The prepositions 'over' and 'across' both only appear with one reading. 'Through' appears in three senses: as VIA + IN, similar 'over' (but only for the verb 'loop'), and as instrument. These three senses are all reflected outside this domain. 'Via' seems to function like described in Jackendoff (1983). However, there is a great overlap between 'through' and 'via'. Again, the assumption must be made that the prepositional object is underspecified, which accounts for the overlap between 'via' and 'through'.

Additionally, like *goal-paths*, *routes* also occur occasionally with *source-paths*. Again, this is crucial, because it shows that there is an underlying path present. Furthermore, 102 sentences use both a *route* and a *goal-path*, 17 of them even encode the *source*, *route*, and *goal*, such as the examples in (26). This occurs mostly with 'connect' and 'transfer'. In 4 sentences, two *routes* were specified, as in the examples, as in (27). Double *goal-paths* can only happen when the second goal extends the first goal, but these were not accounted for, as it goes beyond the scope of this thesis.

(26) a. 'copy large files via the nfs mount point from the application server to the ftps nas storage.'
b. 'When structures are transferred via network from Compaq to HP.'

(27) a. '(my bios supports usb boot as) i have installed debian over the usb through netinstall-iso image.

b. ‘it’s about connecting windows pc machine through hyper terminal to unix machine via dial up’

The preposition ‘against’, as argued in chapter 2 is a *static* preposition, in contrast to ‘over’, ‘through’, and ‘via’. *Static* prepositions can only denote a *path* when used with a motion verb. The verbs that occurred with ‘across’ (‘create’, ‘write’, ‘copy’) were not positional and non-stative, therefore the data are in line with Denis et al. (2003).

4.3 Conclusion

In this section, I have shown that the use of prepositions that express both *goals* and *routes* can be accounted for from the perspective of spatial semantics. More specifically, they can be accounted for by an approach that is based on Jackendoff (1983)’s functions for *paths*. This approach assumes that speakers prefer a more *neutral* preposition when the spatial properties of the prepositional object are unclear or underspecified.

The prepositions examined in this chapter occurred with ‘from’, which expresses a *source-path*. As I argued above, this is crucial, because it shows that there is an underlying *path* present. What does stand out is that ‘against’ and ‘across’ are never used together with a *source-path*, which might be related to ‘against’ and ‘across’ being *static* prepositions. However, more data needs to be gathered to further research this.

The prepositions ‘against’ and ‘across’ are *static* prepositions and are therefore only able to denote a *path* when used with a motion verb. I have argued in this section that this constraint is not violated. However, for these prepositions there is little data available to make robust claims. To make stronger claims, additional research must, which can be done by looking at just these prepositions in other forums, or their occurrences with different verbs.

Chapter 5

Conclusion

In this thesis, I have looked at two types of prepositions. On the one hand, the prepositions 'to', 'into', 'onto', and 'against', which all express a *goal-path*. On the other hand, there are the prepositions 'over', 'through', 'via', and 'across', which express a *route*. These prepositions are used in an 'atypical' way in software development forums, as the prepositions occur with verbs they would not occur with in Standard English.

Data collection

In order to research these prepositions, I have collected data from a question answering forum on software development. In particular, the AEU forum, a subforum of the UNIX forum, was deemed valuable. The UNIX forum in general has strict policies, active enforcement of these policies, and subcategorizes different topics into smaller subforums. Of these subforums, the AEU forum was most appropriate, because of its target group (expert users), size, and accessibility to non-users.

Data were collected from this forum using web scraping. I extracted all posts from the website. A total of 37,600 posts were scraped, which were then segmentized into sentences. Due to the domain of the data, a dependency parser gave unreliable analyses for the sentences in the dataset. Therefore, a list of verbs was constructed to downsize the dataset to a size that was manageable for manual assessment. The data were then filtered, leaving only the sentences that were useful for analysis.

There are two main problems with this method. Firstly, the list of verbs was hand-picked. The list is constructed based on data exploration and casting aside the irrelevant verbs. This is problematic, because the selection is based only on the intuitions of the researcher. While this is merely a pilot study, it would have been beneficial to base the verb list on a categorization of some sort. For instance, balance it for *dynamic* and *stative* verbs, or only use verbs that have a counterpart outside the *cyber* domain. Another approach would be basing the verb list on Levin (1993)'s famous verb classification. This might not be helpful for this language domain, because many domain-specific verbs (like 'to SSH') are not covered by Levin (1993). However, further research using on verbs that are covered in Levin (1993) is now possible and will further show

how this domain relates to Standard English.

The second issue with this method is the manual assessment, which leaves room for mistakes. Ideally, every sentence would be evaluated by a second researcher. However, there was no room to do this, which is problematic for two reasons. On the one hand, there could still be some noise in the dataset that should not be there. For this thesis, that would mean the frequencies described in chapter 3 are too high. For the analysis, this would not make a great difference, since the analysis is based on the clearest examples in the data. On the other hand, the manual assessment is problematic, because data that supports the analysis might have been deleted accidentally. This negatively influences the frequencies described in chapter 3, but also influences the analysis. Some verbs or verb-preposition pairs were not considered in the analysis, because they occurred only once in the dataset. Since the frequencies for most of the verb-preposition pairs is not that high, accidentally deleting one sentence can exclude a verb-preposition pair from the analysis. While the issues mentioned here might no be completely resolved by having a second assessor going over all the sentences, it does give a more trustworthy dataset and thus a more valid thesis.

There is a third issue, that has influenced that statistics and the analysis in the thesis, namely the sentence structure. Only one particular structure was searched for, namely sentences where the verb is followed by the preposition. This way, some data containing both the preposition and the verb is overlooked. For example, sentences like 'I am sure that through this connection nothing was transferred' would not be included in the dataset that was used for analysis. However, this is only a minor issue, given that there still was enough data for an analysis.

Web scraping is a valuable method to collect data. While it is time consuming, it is an efficient way for an individual researcher to collect a substantial amount of data. However, chapter 3 showed that using this method results in noisy data, even when precautionary steps (such as excluding *quotes*) were taken. Further research will benefit from developing web scraping methods for linguistic purposes. Additionally, developing web scraping further for data concerning software development can make the data suitable to use with a dependency parser.

Cyberspatial prepositions

In this thesis, I have tried to find an answer to the question to what extent spatial prepositions within the language domain of software development are used similar to their use in Standard English. More specifically, I have tried to make clear to what extent both the use of prepositions associated with *goal-paths* and *routes* in *cyberspace* is similar to the use in Standard English, with regards to spatial semantics. The sentences containing verb-preposition pairs were analysed in chapter 4. This chapter showed that Jackendoff (1983)'s approach does not cover the full extent of the data. Therefore, I propose that the speakers use 'to' instead of 'into' or 'onto' when the prepositional object is underspecified for certain spatial properties. Here, underspecified means that the speaker does not know whether the prepositional object can function as a container or supporting surface, and therefore chooses to use a different, more neutral preposition.

This also applies for 'via' and 'through', although less strongly. Most verbs occur more frequently with 'via' or equally frequent with both. Only 'execute', 'run', and 'pipe' occur more

frequently with 'through'. Upon closer inspection, it seems that these verbs select prepositional objects that are almost exclusively commands or code. It is unclear why speakers have a preference for 'through' when the prepositional object is a command or code, but it seems to be the only clear exception. Further research needs to be done to determine the origin of this exception. Apart from that, the data can be accounted for using Jackendoff (1983)'s theory on spatial prepositions and extending it with my proposal.

The meaning of the preposition does not change, as was expected in 1. Given that language changes gradually (Anttila, 1989, Hopper and Traugott, 2003), it was to be expected that the meaning of verbs changed and the meaning of prepositions did not, since prepositions are a functional category, and verbs are not. I also expected variation in the use of prepositions. Furthermore, given that the frequencies for most verb-pairs were quite low, it is difficult to tell whether this variation is consistent. Additionally, because the user names were not collected, it is not possible to link the posts to a speaker. So, it is impossible to tell whether this variation is between speakers of within speakers.

The selection of the prepositions highly influences the outcome of the thesis. For instance, in line with the research of Denis et al. (2003), 'across' and 'against' are *static* prepositions, whereas the other six prepositions are *dynamic*. In the end, this did not make great differences for the analyses. In general, it is important to consider what prepositions are usable for comparison. For instance, some prepositions, like 'at' and 'of', are highly frequent, which can lead to *bleaching* and polysemy. This makes it more difficult to compare prepositions, the focus should then be put more on the individual senses of one preposition. Prepositions with a very low frequency can have too little occurrences to compare at all. Studying these prepositions takes a different approach as well. More data needs to be collected to study prepositions with low frequencies, possibly from different sources.

Furthermore, this thesis considered only *paths*. Other types of prepositions need to be considered too. The analysis relies on the prepositional being 'underspecified for spatial properties', meaning that the speaker does not know whether the prepositional object can function as a container or supporting surface. It is highly necessary to further research this claim by studying the uses of 'in' and 'on', in order to see if this claim holds.

Another suggestion for follow-up research is directly related to Jackendoff (1983). Firstly, this thesis looked at the preposition 'from' only superficially, as only sentences that contained a *source-path* in addition to a *goal-path* or *route* were considered. Next to *bounded paths* and *routes*, Jackendoff (1983) also describes *directions*, that are expressing by prepositions such as 'toward'. Both *source-paths* and *directions* can be interesting for further research.

More generally, it can be valuable to compare the use in this forum to the use in, say, another forum on software development. Since this thesis is a pilot study, still much work is needed to give a more complete view on the use of language in this specific language domain, let alone how it relates to other domains.

A general issue with online data is that many users do not speak English as their first language. This is an important issue, because prepositions are problematic to second language learners of English, see for instance Dalgish (1985) and Bitchener et al. (2005). It is a very difficult problem

to solve. The UNIX forum does not require a user to register their native language or country of origin. Different forums might, but it is difficult to check whether this information is truthful. One solution could be to analyse just one preposition and take into account the influence of second language interference on that particular preposition, but even then the information about the first language needs to be present. While this thesis is a pilot study, it was not possible to take this into account. However, follow-up research should seriously consider this issue.

The current research sheds light on the perception of spatial relations within a virtual environment, but also on the conceptualization of something non-spatial as something spatial. Next to that, it raised many more questions. The data collected in this thesis are very promising, as they raise issues that are specific to this language domain. On the one hand, more research is needed to explore this language domain further. On the other hand, other language domains, such as related online communities, or software documentation, become relevant, as they can be compared to the data gathered in this thesis.

Bibliography

Aitchison, J. & Lewis, D. M. (2003). Polysemy and bleaching. In B. Nerlich, Z. Todd, V. Herman, & D. D. Clark (Eds.), *Polysemy: Flexible patterns of meaning in mind and language* (pp. 253–265). Walter de Gruyter.

Anttila, R. (1989). *Historical and comparative linguistics*. John Benjamins Publishing.

Apple publications style guide. (2003). Cuptertino, CA: Apple.

Beavers, J. (2008). Scalar complexity and the structure of events. In J. Dölling, T. Heyde-Zybatow, & M. Schäfer (Eds.), *Event structures in linguistic form and interpretation* (pp. 245–265).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on esl student writing. *Journal of second language writing*, *14*(3), 191–205.

Crystal, D. (2011). *Internet linguistics: A student guide*. New York, NY: Routledge.

Dalgish, G. (1985). Computer-assisted esl research and courseware development. *Computers and Composition*, *2*(4), 45–62.

De Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of lrec* (Vol. 6, *2006*, pp. 449–454). Genoa Italy.

De Souza, C. S. & Preece, J. (2004). A framework for analyzing and understanding online communities. *Interacting with computers*, *16*(3), 579–610.

Denis, P., Kuhn, J., & Wechsler, S. (2003). V-pp goal motion complexes in english: An hpsg account. In P. S. Dizier (Ed.), *Proceedings of the acl-sigsem workshop: The linguistic dimensions of prepositions and their use in computational linguistics formalisms and applications* (pp. 121–132).

Dodge, M. & Kitchin, R. (2003). *Mapping cyberspace*. New York, NY: Routledge.

Evans, N. (2014). Positional verbs in nen. *Oceanic Linguistics*, *53*(2), 225–255.

Fenk-Oczlon, G., Fenk, A., & Faber, P. (2010). Frequency effects on the emergence of polysemy and homophony. *International Journal of Information Technologies and Knowledge*, *4*(2), 103–109.

Fillmore, C. F. (1968). The case for case. In E. Bach & R. T. Harms (Eds.), *Universals in linguistic theory* (pp. 1–88). New York, NY: Holt, Rinehart and Winston.

Gehrke, B. (2007). Putting path in place. In E. Puig-Waldmüller (Ed.), *Proceedings of sinn und bedeutung* (Vol. 11, pp. 244–260).

Gruber, J. S. (1965). *Studies in lexical relations.* (Doctoral dissertation, Massachusetts Institute of Technology).

Hawkins, B. W. (1984). *The semantics of english spatial prepositions* (Doctoral dissertation, University of California, San Diego).

Holtz, P., Kronberger, N., & Wagner, W. (2012). Analyzing internet forums: A practical guide. *Journal of Media Psychology*, *24*(2), 55–66.

Honnibal, M. & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In L. M‘arquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373–1378).

Hopper, P. J. & Traugott, E. C. (2003). *Grammaticalization*. Cambridge University Press.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT press.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic inquiry*, *18*(3), 369–411.

Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT press.

Jackendoff, R. (2012). Language as a source of evidence for theories of spatial representation. *Perception*, *41*(9), 1128–1152.

Jackendoff, R. & Landau, B. (1991). Spatial language and spatial cognition. In D. Napoli & J. Kegl (Eds.), *Bridges between psychology and linguistics: A swarthmore festschrift for lila gleitman* (pp. 145–169). Psychology Press.

Kiss, T. & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, *32*(4), 485–525.

Lakoff, G. (1966). Stative adjectives and verbs in english. In A. Oettinger (Ed.), *Mathematic linguistics and authomatic translation, report nsf-17*. Cambridge, MA: The computation laboratory of Harvard University.

Landau, B. & Jackendoff, R. (1993). ‘‘what” and “where” in spatial language and spatial cognition. *Behavioral and brain sciences*, *16*(2), 217–265.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford, CA: Stanford university press.

Lawson, R. (2015). *Web scraping with python*. Birmingham, UK: Packt Publishing.

Leech, G., Rayson, P., & Wilson, A. (2014). *Word frequencies in written and spoken english: Based on the british national corpus*. London: Longman.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Levinson, S. C. (1996). Frames of reference and molyneux’s question: Crosslinguistic evidence. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (Chap. 4, pp. 109–169). Cambridge, MA: The MIT Press.

Mandler, J. M. (1992). How to build a baby: II. conceptual primitives. *Psychological review*, *99*(4), 587–604.

Microsoft manual of style. (2012). Redmond, WA: Microsoft Press.

Miller, G. A. & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.

Nguyen, D. & Rosé, C. P. (2011). Language use as a reflection of socialization in online communities. In *Proceedings of the workshop on languages in social media* (pp. 76–85). Association for Computational Linguistics.

O'Keefe, J. (1996). The spatial prepositions in english, vector grammar, and the cognitive map theory. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (Chap. 7, pp. 277–316). Cambridge, MA: The MIT Press.

Pütz, M. & Dirven, R. (2011). *The construal of space in language and thought*. Walter de Gruyter.

Rhee, S. (2002). Semantic changes of english preposition 'against': A grammaticalization perspective. *Language Research*, *38*(2), 563–583.

Richardson, L. (2007). Beautiful soup documentation.

Svorou, S. (1994). *The grammar of space*. Amsterdam: John Benjamins.

Talmy, L. (1983). How language structures space. In H. L. Pick & L. P. Acredolo (Eds.), *Spatial orientation: Theory, research, and application* (Chap. 11, pp. 225–282). Springer.

Talmy, L. (1996). Fictive motion in language and "ception". In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (Chap. 6, pp. 211–276). Cambridge, MA: The MIT Press.

Vasardani, M., Stirling, L. F., & Winter, S. (2017). The preposition at from a spatial language, cognition, and information systems perspective. *Semantics and pragmatics*, *10*(3).

Xu, Y. D. & Badler, N. I. (2000). Algorithms for generating motion trajectories described by prepositions. In *Proceedings computer animation 2000* (pp. 30–35). IEEE.

Zlatev, J. (2003). *Holistic spatial semantics of thai* (G. B. Palmer & E. H. Casad, Eds.). Berlin and New York: Mouton de Gruyter.

Zlatev, J. (2007). Spatial semantics. In H. Cuyckens & D. Geeraerts (Eds.), *The oxford handbook of cognitive linguistics* (Chap. 13, pp. 318–350). Oxford: Oxford University Press.

Zwarts, J. (1997). Vectors as relative positions: A compositional semantics of modified pps. *Journal of semantics*, *14*, 57–86.

Appendix A

Example sentences

Verb	Preposition	Sentence
authenticate	against	I am looking to have unix authenticate against active directory
link	against	you cannot dynamically link a solaris shared library against linux libraries
run	against	what happens when you run e2fsck against the filesystem?
access	into	I can remotely access into the machine
boot	into	it all failed to boot into single user mode
converted	into	hp-ux can be converted into a trusted system
copy	into	and then move or copy the file from /lost+found directory into the correct path
log	into	I have different groups of users log into our servers
mount	into	”mount_smb” which allows you to mount a windows folder into a unix system
telnet	into	I want to be able to telnet into an sco openserver 5 system
writing	into	the file creation process is still writing records into the file
appending	onto	this prevents ssh from appending onto known_hosts
copy	onto	I want to open an excel file and copy one column of the file onto another text file
transfer	onto	could I possibly transfer the files from disk 1 onto a floppy?
connected	to	make sure your fallback sshd is connected to your ssh-session
converting	to	is the process for converting an imap user to pop3
downloading	to	why aren’t you downloading to the target directory
ftp	to	so the issue happends when we ftp the file to unix alone
pipe	to	immediately after that pipe the output to a grep command
redirect	to	so you can just redirect it to /dev/null
ssh	to	when I ssh from my cluster to another cluster
written	to	so if a command aborts, it gets written to that log file

Table A.1: Example sentences that contain a *goal-path*, taken from the AEU corpus. The first column shows the verb, the second shows the preposition, the third colum shows the example sentence.

Verb	Preposition	Sentence
copies	across	scp is behaves like the cp command but copies across servers
execute	across	run several unix commands and need to execute across different servers
ftp'ing	across	this is very useful explorer for ftp'ing across operating system
written	across	only data that has changed gets written across the network
connect	over	to login, you connect over port 21
copying	over	mounting the drives and copying over network
install	over	I never did figure out how to boot or install windows over a network
iterate	over	you can use a for loop and use wget and iterate over files
transfer	over	I want to use a shell-script to transfer data over sftp
accessing	through	now that I am accessing remote servers and accounts through ftp
connect	through	I want to connect ubuntu linux through putty
created	through	how are the files created through a script
login	through	note that if this user login through telnet
login	through	if so you have to login through some one else user id
uploaded	through	the path to which these get uploaded through http
accessed	via	note that if the file is being accessed via nfs
access	via	when I am trying to access the console via root login
boot	via	my users are unable to boot via grub
download	via	if I download a file via sftp
executing	via	check the exit status of a remote command executing via rsh?
mounted	via	I inherit a solaris7 system with /home mounted via a share from the nfs server
loads	via	the boot loader then loads its configuration file via tftp from next-server

Table A.2: Example sentences that contain a *route*, taken from the AEU corpus. The first column shows the verb, the second shows the preposition, the third colum shows the example sentence.

Verb	Sentences
connect	ideally, you would connect the computer via a lan to the aix box
copy	permissions are such that I can't copy these files via /tmp from my home to the mount point
ftp	ftp the same large files via the nas mount point from the application server to the ftps nas storage
link	I am supposed to link the ibm storage device via emulex host bus adapter to the dell poweredge
mount	because of this, writing across a nfs mount to a fifo
mounted	it can be mounted via nfs onto the servers
pipe	don't try an experiment where you just pipe a file through uuencode to sendmail -t
pipe	of course you can simply pipe syslog via a named pipe into a filter
transfer	when I am trying to transfer a file through a program from one directory to another directory

Table A.3: Example sentences that contain both a *goal-path* and a *route*, taken from the AEU corpus. The first column shows the verb, the second shows the preposition, the third colum shows the example sentence.

Appendix B

Verb-preposition pairs with frequency

verb	prep	pair count	verb count	verb	prep	pair count	verb count
connect	to	247	300	ssh	into	33	51
connect	via	63	300	ssh	to	16	51
connect	through	34	300	ftp	to	40	49
write	to	130	183	transfer	to	27	49
write	into	43	183	transfer	through	12	49
copy	to	78	133	install	into	15	44
copy	into	39	133	boot	into	31	39
log	into	101	131	pipe	into	15	39
log	to	19	131	pipe	to	14	39
login	to	48	96	pipe	through	11	39
login	through	18	96	append	to	28	36
login	via	18	96	access	via	19	32
login	into	15	96	cd	into	17	29
convert	to	47	90	cd	to	12	29
convert	into	43	90	loop	through	21	27
redirect	to	61	78	telnet	to	16	26
redirect	into	15	78	execute	through	15	25
run	through	29	75	link	to	18	22
run	via	17	75	authenticate	against	16	22
run	against	13	75	type	into	14	16
run	to	12	75	merge	into	12	13
load	into	52	65				

Table B.1: The frequencies for all verb-preposition pairs that occurred more than 10 times in the AEU corpus. Next to the frequency of the verb-preposition pair, also the overall frequency of the verb in the AEU corpus is given.