Chapter 1: Introduction

1. GENERAL CONCEPTS

This thesis explores mutual intelligibility in the Slavic language area – how to measure it, how to predict it and how to increase it. It consists of two main parts: in part one, we measured the intelligibility levels among six Slavic languages as well as the factors we believed could predict them and tried to create a model of intelligibility for the Slavic language family. Part two consists of two separate studies whose aim is to explore specific issues which arose during the development of the intelligibility model. The first deals with improving the level of intelligibility through teaching and the second explores two Slavic languages bound by a very specific and a somewhat troublesome political relationship – Serbian and Croatian.

The present thesis extended the work on Slavic languages substantially. Firstly, several ordered pairs of languages were tested using the same, uniform methodology, making the results commensurable for the first time. Secondly, linguistic distances between six Slavic languages were measured on several levels (lexicon, orthography, phonology, morphology and syntax), using uniform methodology and relying on computational methods. Thirdly, we developed a model which can predict the level of intelligibility between Slavic languages to a substantial degree. Fourthly, we showed that a very short teaching intervention targeting similarities and differences between the two related languages (in this case, Czech and Croatian) can significantly improve the level of intelligibility. Finally, we demonstrated that despite the high linguistic similarity, speakers of Serbian and Croatian hold negative attitudes to each others’ varieties. To our knowledge, all of the above work has been carried out on Slavic languages for the very first time.

Before we set out to explore intelligibility in the Slavic language area, we must first define what intelligibility is. The term *intelligibility* often refers to the quality of being understood under adverse conditions, for instance if the speech is time-compressed (Heiman, Leo, Leighbody, & Bowler, 1986), if multiple people are speaking (Bronkhorst, 2000) or if a foreign accent is involved (Munro, 2008). In this context, intelligibility is a property of the signal.

Smith (1992) introduced the Smith Framework for Intelligibility, which refers to different varieties of English. The framework distinguishes between three levels of participants’ apprehension of speech events. The first one, *intelligibility*, refer to the actual recognition of words and utterances, either in the written or in the spoken modality. The next level is called *comprehensibility* and it
pertains to the understanding of meaning of those words or utterances, while the third and the most complex level, interpretability, refers to the recognition of intent or purpose behind the input. So for instance, if person A asks person B: “Do I know you?” understanding the speech signal and recognizing the utterance would be intelligibility; understanding the meaning of the question and conceivably being prepared to provide the yes or no answer would be comprehensibility; while interpretability would refer to the intention behind the utterance, which might be a genuine question, a conversation starter, an attempt at flirting etc.

Intelligibility in Smith’s framework is in fact almost inextricably linked to comprehensibility: we could talk about pure intelligibility only in the case a person was receiving messages in a language they do not understand at all, so they would be able to see/hear and possibly discern the words, but not understand their meaning. An example of this would be a person of faith listening to a church service in a language they do not speak, for instance Hebrew, Old Church Slavonic or Latin. They could know parts of what is said by heart, parse what is said into words and could even recite them without ever understanding the meaning.

In most cases though, intelligibility comes with at least a bit of comprehensibility i.e. recognizing an utterance involves decoding at least a portion of the meaning. A speaker of language A, who reads or listens to a related language B and is able to understand to an extent what is written/said, is engaged in both intelligibility and comprehensibility in Smith’s framework. The level of interpretability is not essential for basic communicative purposes such as this one and even native speakers of the same language may have trouble interpreting intentions behind utterances.

We believe that in the case of related languages, it is unnecessary if not impossible to tease intelligibility and comprehensibility apart. But if a single term is to be used, we opt for intelligibility, since in line with previous research, the term implies the act of understanding under unfavorable conditions (an unknown, albeit a similar language being the unfavorable condition in this case). Therefore, throughout the thesis, the term intelligibility will refer to the understanding of the meaning of individual words or texts presented in a related, but unfamiliar language.

If intelligibility of a closely related language is defined as the ability to understand a language by virtue of it sharing some similarities to another language the participant speaks, then for measuring intelligibility effectively, we would have to make sure that, first, the participants do not have any knowledge of the test language i.e. that they have never learnt it and second, they have never been exposed to the test language for a prolonged period of time.
In reality, however, language matters are rarely that straightforward: it is practically impossible to find a Dutch or Swedish native speaker who does not have any knowledge of English or who has never been exposed to it. If we were to measure the influence of linguistic and extra-linguistic factors in such a setting, we would come to the fairly obvious conclusion that the best predictor of intelligibility of English by Dutch or Swedish participants is their previous knowledge of the language and/or their exposure to it. Therefore it is necessary to differentiate between inherent intelligibility, which is based purely on the similarities between the related languages in question (for instance Slovak speakers reading or listening to Bulgarian for the first time in their lives) and acquired intelligibility, which refers to intelligibility of a language acquired over time through formal education or other forms of exposure. Acquired intelligibility without any inherent intelligibility occurs in the case of completely unrelated languages, such as Russian and Estonian (Bahtina, ten Thije, & Wijnen, 2013). But for many languages belonging to the same family, their mutual intelligibility is in fact a mixture of these two types: for instance, in the case of Slovak and Polish speakers living in a border area – their languages are similar to an extent by virtue of belonging to the same language family (Slavic) and the same sub-family (West Slavic), but the level of intelligibility in this case is also related to the amount of exposure, which represents acquired intelligibility. For this reason, when it comes to intelligibility of closely related languages, it is necessary to distinguish between linguistic factors influencing intelligibility, which are related to the concept of inherent intelligibility and extra-linguistic factors, such as language exposure, which are related to the concept of acquired intelligibility.

If the level of mutual intelligibility between two languages is high enough, the speakers of those languages might be able to communicate by each using their native language. This type of communication is far from new – it has been used in some parts of Europe in the late Middle Ages (Braunmüller, 2007). Haugen (1966b) coined the term semicommunication to describe this type of interaction, but the term was later abandoned as it implies incomplete, inferior or only half-successful communication (Bahtina & ten Thije, 2012). Today receptive multilingualism (RM) is the more commonly used term for communication whereby interlocutors communicate by speaking their native language and understanding the language of their interlocutor, essentially performing code-switching at turn level (Zeevaert, 2004; ten Thije & Zeevaert, 2007; Gooskens & Heeringa, 2014).

Rehbein, ten Thije, and Verschik (2012) introduced the term Lingua receptiva (LaRa), which refers to “the ensemble of those linguistic, mental, interactional as well as intercultural competencies which are creatively activated when interlocutors listen to linguistic actions in their ‘passive’ language or variety” (p. 249). In LaRa, the focus is on the communicative process itself, which brings
about the need for distinguishing between the competencies of the speaker and the competencies of the hearer. Another distinction between LaRa and receptive multilingualism is that receptive multilingualism has generally been used for communication between native speakers of related languages whereas “LaRa is neither restricted to constellations between typologically related languages nor to the L1-L2 formula” (Bahtina, 2013). Therefore, when describing the actual type of communication between two speakers who use their closely related native languages, LaRa and receptive multilingualism may be considered synonymous.

Other related concepts include plurilingual communication and intercomprehension. Plurilingual communication refers to the multilingual communication on the level of the whole society (Lüdi, 2007). Intercomprehension is used in a similar way as receptive multilingualism, since Doyé (2005) defines it as “a form of communication in which each person uses his or her native language and understands that of the other” (p. 7). The term intercomprehension is most commonly used in the Romance language area (Blanche-Benveniste, 1997; Degache, 2006; Escudé & Janin, 2010, García Castanyer & Vilaginés Serra, 2010), although it is sometimes found in the literature on Scandinavian languages as well (Delsing, 2007).

In order for receptive multilingualism to be possible between speakers of different languages which belong to the same language family, a necessary but not a sufficient condition is a reasonable degree of mutual intelligibility between the languages in question. The type of intelligibility in this case can be either inherent only or a combination of both inherent and acquired, but a high level of intelligibility does not necessarily translate into successful RM communication. The success of communication can depend on a number of other factors including, but not limited to: the complexity of the subject matter, the motivation of the speakers as well as their ability to draw inferences. But if the level of mutual intelligibility of the languages in question is low to begin with, RM communication is bound to fail. Measuring intelligibility of closely related languages provides us with important insights into the possibilities of successful RM communication. In addition, finding out how important the differences in language structure are to the ability to understand a related language might help us identify the cases where 1) not much language education is necessary, since the intelligibility level is high enough to ensure a meaningful interaction; 2) targeted language education, potentially the one focused on pointing out the similarities and differences between the language pairs may increase intelligibility levels significantly and 3) extensive language education is necessary, since the language pairs in question are so different in structure that intelligibility levels are relatively low.
2. The MICReLa project

In 2007, the High Level Group on Multilingualism (HLMG) reported on the lack of knowledge about the potential for using receptive multilingualism in the context of closely related languages in Europe (High Level Group on Multilingualism, 2007). The MICReLa project was established with the main aim of answering the questions about the level of intelligibility between closely related languages in Europe, and the factors influencing it. The focus of the project is on the three largest language families in Europe: Germanic, Romance and Slavic, with one PhD student in charge of each language area. This thesis represents a collection of studies on the intelligibility in the Slavic language area, the thesis on the Germanic language area has already been published (Swarte, 2016), while the thesis dealing with the Romance languages is currently being written (Voigt, in preparation).

The MICReLa project started in 2011. So far it has yielded a number of publications dealing with the material and the results of the main experiment conducted online (Heeringa, Golubović, Gooskens, Schüppert, Swarte, & Voigt, 2013; Swarte & Hilton, 2013; Heeringa, Swarte, Schüppert, & Gooskens, 2014; Golubović & Gooskens, 2015; van Heuven, Gooskens, & van Rezooijen, 2015; Heeringa, Swarte, Schüppert, & Gooskens, accepted). It is also worth noting the publications of MICReLa team members which describe the results of numerous side-projects conducted in the same time frame (Voigt & Schüppert, 2013; Golubović & Sokolić, 2013; Bergsma, Swarte, & Gooskens, 2014). In addition, many other articles are currently in preparation. More information can be found on the project homepage: [http://www.let.rug.nl/gooskens/project/](http://www.let.rug.nl/gooskens/project/)

The main decisions regarding the experimental procedures and the testing material were reached through consensus of all the MICReLa team members. Each PhD student was in charge of creating versions of the test material for her own language families, setting up the experiment, collecting and analyzing data and finally, discussing and writing up the results. We made sure that we used comparable testing material and the same experimental protocol for all three language families, while each PhD student had the freedom to filter and analyze the data in accordance with the specific language situation and her particular research questions. The findings stemming from the MICReLa project are described in chapters 1-5 of the thesis, while chapters 6 and 7 are the author's own side projects which have been developed and conducted independently in the same time frame.

We employed four basic criteria in choosing the languages of the study.

1. The languages should belong to one of the three largest language families in Europe: Germanic, Romance or Slavic.
2. All languages had to be official languages of the European Union.¹
3. In case there are several standards of the same language, we would work with the one spoken in Europe (this would apply to Spanish or English, for instance).
4. In case a language was spoken in more than one country in Europe, we would opt for the biggest one in terms of speakers (which meant German from Germany was used, rather than German from Austria or Switzerland).

**Figure 1.1:** A map showing the geographic distribution of Slavic languages
(https://en.wikipedia.org/wiki/Slavic_languages)

---

¹ As of 2013, there are 24 official EU languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish.
These criteria left us with a total of 16 languages: five Germanic (German, Dutch, English, Swedish and Danish), five Romance (French, Spanish, Portuguese, Italian and Romanian) and six Slavic (Czech, Slovak, Polish, Croatian, Slovenian and Bulgarian). The focus was only on the standard languages, as including different dialects of each language would result in a much more complicated study design and would bring about additional problems concerning dialect demarcations and dialect choice.

This thesis deals with the Slavic language family, so in the next section, we shall present more information on the Slavic languages in general and the six languages of our studies in particular. A map showing where Slavic languages are spoken in Europe can be seen in Figure 1.1.

When choosing participants, we decided to focus on young educated adults. The age range chosen was 18-30 years of age, with university students as our primary target group. This group was the most likely to be mobile and therefore most likely to get in touch with other languages in the future. Secondly, focusing on young participants would enable us to speculate about the future potential of using receptive multilingualism in the Slavic language area, which might offer numerous practical benefits that could ultimately be translated into economic benefits, such as business collaboration with reduced or non-existent translation costs. Finally, working with university students offered practical advantages: no parental or school consent was required for their participation and we had access to large pools of participants through their universities.

3. THE SLAVIC LANGUAGE FAMILY

With a total of about 315 million speakers, the Slavic language family is the second largest language family in Europe (the largest is the Germanic language family with about 500 million speakers). Slavic languages are traditionally divided into three sub-families, West Slavic, South Slavic and East Slavic, which reflect the division of Proto-Slavic into dialects. The sub-families also feature divisions into further branches. A complete overview of the Slavic language family on the basis of Sussex and Cubberley (2006) is provided in Figure 1.2.

According to De Bray (1969) the main typological characteristics Slavic languages are: varying degrees of palatalization; few true diphthongs, agglomeration of consonants (but no aspirated consonants), a high degree of inflection in nouns, pronouns, adjectives and verbs, highly developed verbal aspect and elastic and variable word order.

Comrie and Corbett (2003) also note a relatively large number of palatal and palatalized consonants which form minimal pairs in many Slavic languages. For instance in Russian, just the
palatalization of the /t/ sound results in different words: /mat/ ‘checkmate’ and /matʃ/ ‘mother’. Another important characteristic is the rich inflectional morphology of the fusional type, meaning that one affix can mark a number of different grammatical categories. In Czech word mladěmu, the suffix –ému simultaneously marks the dative case, singular number and masculine gender. Most Slavic languages also exhibit a great amount of morho-phonemic alternations, for instance Croatian dječak ‘boy’ but dječaci ‘boys’.

Figure 1.2: The Slavic language family tree. The six languages we focus on are marked with grey.

The aspectual system in Slavic languages is very developed, with perfect-imperfect opposition as a characteristic of practically every verb form. E.g. in Polish the imperfective verb pomagać means to ‘be helping for a period of time either in duration or across several instances’, while pomóc, the perfective counterpart of pomagać, means ‘to finish (one instance of) helping’. Comrie and Corbett (2003) also mention the free word order of major clause constituents, which is determined mostly by pragmatic factors, such as topic or focus. While in English the sentence *Alex saw she would be ungrammatical, in most Slavic languages this word order would be perfectly acceptable and would carry the meaning It was Alex she saw, not David.

Extensive agreement system, e.g. in case, number and gender within the noun phrase is another characteristic of almost all Slavic languages. For instance in Croatian Marija je donela knjigu – ‘Maria brought the book’ (lit. Maria is brought the book), the auxiliary verb je shows agreement in
person and number (in this case 3rd person singular) while the participle *donela* agrees with the subject in number and gender (singular and feminine).

Hapslmath (2001) also lists multiple fronting of interrogative words as a feature that is most clearly demonstrated in the Slavic languages. Consider the following examples:

Bulgarian

\[
\text{Koj kogo e vidjal?} \\
\text{who whom saw} \\
\text{‘Who saw whom?’}
\]

Polish

\[
\text{Co komu Monika dala?} \\
\text{what to whom Monica gave} \\
\text{‘What did Monica give to whom?’}
\]

Fronting of interrogative words is actually a strong requirement in most Slavic languages and failure to do so, results in an ungrammatical sentence as in the following examples (from Haspelmath, 2001):

\[
*\text{Koj e vidjal kogo?} \text{ (Bulgarian)} \\
*\text{Co Monika komu dala?} \text{ (Polish)}
\]

Double negation is another way in which Slavic languages differ from many other European languages. We shall illustrate this with a Czech phrase *nikdo to neví* (‘nobody knows that’), where the negation is both the subject and the predicate, which means that the literal English translation would be ‘nobody doesn’t know that’ (Horálek, 1992).

On the basis of the aforedescribed structural similarities, it appears that Slavic speakers would probably not have too much trouble navigating the complex systems of declensions, conjugations and relatively free word order in a related Slavic language, as it is likely that their native languages exhibit very similar characteristics. However, Slavic languages have been developing more or less independently for centuries, which brought about a number of changes and ultimately differences between individual languages.

One of the most obvious distinctions among West, South and East Slavic languages concerns their scripts. Dale (1980) notes that the rivalry between Christian Orthodox and Catholic churches is neatly reflected in the scripts that Slavic languages use even today: the languages of the areas where the population was originally converted to the Roman Catholic church are written in local versions of the Latin alphabet (Czech, Slovak, Polish, Sorbian, Slovenian, Croatian), whereas the languages of the regions where orthodox missionaries used to operate are now written in Cyrillic (Russian, Belarusian, Ukrainian, Bulgarian, Macedonian). Therefore, East Slavic languages are written exclusively in Cyrillic, West Slavic languages are written exclusively in Latin, whereas in
the South Slavic group some languages only use the Latin script (Croatian, Slovenian), some only use the Cyrillic script (Macedonian, Bulgarian), while Serbian actually uses both Cyrillic and Latin.

When it comes to the linguistic features separating the three sub-families, we shall name just a few main distinctions taken from Sussex and Cubberley (2006). In East Slavic languages, Proto-Slavic initial *je- changed into e and then o (Proto-Slavic jezero, Russian and Ukrainian ozero, Belorussian vozera 'lake'). Word stress is strong, i.e. there is a high contrast in energy between stressed and unstressed syllables and in East Slavic languages this leads to a reduction of length and quality of unstressed vowels. The word stress in East Slavic languages is also free, which means that it can occur on any syllable in a word and is mobile, meaning that it can “move” depending on the word form (the same noun can have stress on different syllables in the nominative and the genitive case).

Proto-Slavic had two types of very short vowels called ‘yers’. Front yer was short front closed vowel and was written as i or й. Back yer was reconstructed as short back closed vowel and was written as Ü or ъ. Since these vowels were produced in a central area of the mouth, this made them relatively weak, similar to the position of schwa in many modern languages. The yers then turned into different vowels in different Slavic languages. West Slavic languages are characterized by the change of yers into e (Proto-Slavic sьnъ ‘dream’ dьnъ ‘day’ => Czech sen, den; Slovak sen, deň; Polish sen, dzień). In West Slavic languages word stress is fixed on the initial syllable, with the exception of Polish, where the change occurred at a later stage.

Languages belonging to the South Slavic language family are the most heterogeneous group, since Eastern South Slavic languages (Macedonian and Bulgarian) differ in terms of their morphology and syntax from all the other Slavic languages. Most linguistic features that all South Slavic languages share can be found on the phonological level, such as the hardening of palatals and dental affricates or the change from Proto-Slavic high central vowel y into i (Proto-Slavic syn ‘son’) => Croatian sin; Slovenian sin; Bulgarian син transliterated: sin). But it should be noted that on other linguistic levels, Bulgarian and Macedonian contain many characteristics which separate them not only from the other South Slavic languages, but from the Slavic language family in general. Here we shall offer a basic list, on the basis of Sussex and Cubberley (2006), which is by no means exhaustive:

1. The loss of almost all case forms of the declensions
2. The loss of the infinitive, which was replaced by subordinate clauses introduced by da (in order to)
3. Suffixal definite article
4. 3rd person pronoun toj, while in all other Slavic languages this pronoun has the form on
5. The renarrative forms (a set of inflectional forms marking the events which the speaker has not witnessed and which she cannot vouch for as a fact)

4. THESIS OVERVIEW

As previously mentioned, the aim of Part 1 is to create a model of intelligibility of Slavic languages using both linguistic and extra-linguistic factors. The four main research questions posed here are:

1. What distances can be established among Czech, Slovak, Polish, Croatian, Slovene and Bulgarian on the level of lexicon, phonology, morphology, orthography and syntax? (Chapter 2)
2. What is the level of cross-language intelligibility among Czech, Slovak, Polish, Croatian, Slovene and Bulgarian? (Chapter 3)
3. How well can linguistic distances predict cross-language intelligibility among Slavic languages? (Chapter 4)
4. How well can both linguistic and extra-linguistic factors predict intelligibility among Slavic languages? (Chapter 5)

In Chapter 2 we describe the methods for measuring lexical, orthographic, phonological, morphological and syntactic distances. The main aims are establishing the distances among the six Slavic languages, measuring how well different linguistic levels correlate with each other and finding out which levels reflect the traditional historical representation of the Slavic language tree best.

Chapter 3 deals with the question what levels of intelligibility can be established among Czech, Slovak, Polish, Croatian, Slovenian and Bulgarian. In order to answer that question, we employed three different methods of measuring intelligibility: the word translation task, which measures intelligibility on the word level; the cloze test, which measures intelligibility on the level of sentences and discourse and the picture task which measures very global topical intelligibility. All three methods were used in the written and the spoken modality. This chapter presents the intelligibility results and compares the three methods.

Chapter 4 represents an attempt to model intelligibility among the six Slavic languages of our study by using linguistic factors only. The assumption here is that linguistic factors (lexical, orthographic, phonological, morphological and syntactic distances) are overall better predictors of intelligibility than extra-linguistic factors (language attitudes and language exposure). Language attitudes have already been shown to be relatively weak predictors of intelligibility and there is not much exposure to other Slavic languages in our sample. Therefore the research question is how well we can model
intelligibility just using linguistic factors. In this case, intelligibility is represented as the results on the written and spoken cloze test.

Chapter 5 describes a more comprehensive approach to modelling intelligibility. We ran a series of stepwise regression analyses using the results of both the cloze test and the translation task as dependent variables and linguistic and extra-linguistic factors as independent variables with the purpose of drawing more general conclusions on the importance of these factors for mutual intelligibility in the Slavic language area.

Part 2 of the thesis represents two independent studies. The research question of the first one came as a consequence of realizing that the intelligibility levels for most language combinations in the Slavic language area were relatively low, so we asked if it was possible to increase it with a short targeted teaching instruction. In Chapter 6 we present the results of a study whose aim was to test the effectiveness of a 4.5 hour instruction to native speakers of Czech on how to understand Croatian better. The effect of the instruction was measured by comparing the results of an experimental group (who received the instruction) and a control group (who did not) on a pretest and posttest of intelligibility. In this case, rather than having our participant read or listen to material in the target language, we measured the success of their interaction. This was done through a spot-the-difference task performed in dyads where one speaker was Czech and the other was Croatian.

In Chapter 7 we zoom in on the Balkans to take a closer look at Serbian and Croatian. The two languages were a part of the common standard language, Serbo-Croatian, just a few decades ago and now they are codified separately. They are still highly mutually intelligible, to such an extent that it is still somewhat unusual to label them as completely separate languages. One of the factors influencing intelligibility are language attitudes. Due to the complex and violent historical developments in the region, the attitudes of Serbian and Croatian speakers to each others’ languages are bound to be negative. In Chapter 7 we test that hypothesis using the matched guise technique and looking into several different age groups. The participants in the oldest group remember the war years vividly, while the ones in the youngest group were not even born at that time.

Chapter 8 is reserved for a general discussion of the findings, limitations of the studies and potential future directions.
Mutual intelligibility
in the Slavic language area

Modelling intelligibility
of Slavic languages