The University Of Sheffield.

Morphological Doublets in Croatian: A multi-methodological analysis

By:
Dario Lečić

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts and Humanities
Department of Russian and Slavonic Studies

20 January, 2017

Acknowledgments

Many a PhD student past and present will agree that doing a PhD is a time-consuming process with lots of ups and downs, motivational issues and even a number of nervous breakdowns. Having experienced all of these, I can only say that they were right. However, having reached the end of the tunnel, I have to admit that the feeling is great.

I would like to use this opportunity to thank all the people who made this possible and who have helped me during these four years spent researching the intricate world of morphological doublets in Croatian.

First of all, I would like to express my immense gratitude to my primary supervisor, Professor Neil Bermel from the Department or Russian and Slavonic Studies at the University of Sheffield for offering his guidance from day one. Our regular supervisory meetings as well as numerous e-mail exchanges have been eye-opening and I would not have been able to do this without you. I hope this dissertation will justify all the effort you have put into me as your PhD student.

Even though the jurisdiction of the second supervisor as defined by the University of Sheffield officially stretches mostly to matters of the Doctoral Development Programme, my second supervisor, Dr Dagmar Divjak, nevertheless played a major role in this research as well, primarily in matters of statistics. I appreciate all your efforts for trying to make me understand R, it is my own fault that it did not work.

I would also like to thank Dr Mateusz-Milan Stanojević from the Faculty of Humanities and Social Sciences in Zagreb for spending hours discussing my work with me during my visits to Croatia and reading my scribbles, even though he was not involved with my research in any official function. Thanks for always believing in me and pushing me forward, Mat. I do not know what you wrote in that letter of recommendation four years ago, but it must have been gold.

The majority of this work was created in Office 1.85e of the Jessop West building. I would like to thank all the people that I shared the office with in the past four years, who have made that office a pleasant place to work in. My special thanks

go to Nina and Jarek for actually making me talk to them and tolerating my office pranks.

Every once in a while, a PhD researcher needs to step out of his cubicle, forget about his work and socialise, otherwise he would go crazy at some point. I was fortunate enough to be a member of the Sheffield University Volleyball Club, whose members provided me with the necessary social life and a chance to let off some steam. Damien, Peter, Adam, Laura, Yvett, Melina, Dahan and all Club members past and present, thank you for being amazing friends. Go Black & Gold!

And finally, I would like to thank all the participants in my numerous questionnaire studies. Some of you I know, some of you I do not, but you all made this possible as well. I apologise for pestering you, but it was all for the sake of science. I was especially proud when several of my participants approached me later by e-mail enquiring about the results. I guess this shows the questionnaire wasn't as boring as I was afraid it would be and that people did feel intrigued by the topic I was researching.

Summary

The term *morphological doubletism* refers to a situation in language when there are two (or more) morphemes available for a single cell in an inflectional paradigm of a lexeme. Slavonic languages, with their rich inflectional systems, show particularly high levels of doubletism. In the present dissertation we analyse examples of doubletism in Croatian nominal paradigms.

As shown by the dissertation's subtitle, "a multi-methodological analysis", we compare and contrast evidence obtained by various methods. First we conduct a corpus study to determine the frequency distributions of the doublet pairs in present-day Croatian. This analysis has shown that the distribution of the doublet pairs is not determined by any intra- or extra-linguistic factor, but that it is not completely random either. These distributions are later used in several additional studies, the purpose of which is to answer the question of how such forms are processed in speakers' mental grammars. One of the analyses is a computational one, in which we try to reproduce a grammar of a Croatian speaker by using two memory-based models (AM and TiMBL). The models were highly successful in producing the desired output without resorting to any rules or generalizations. We also report the results of three questionnaire studies, all of which show that native speakers are extremely sensitive to the language input they receive, in line with usage-based theories of language, as well as that mental grammars are gradient. The speakers' ratings and production rates closely matched the proportions of the doublet pairs in the corpus. Furthermore, speakers distinguish between several levels of domination of one ending over another. When the domination of one form is weak, speakers resort to a different decision criterion, namely they look at the dominant ending of phonologically similar words.

Table of contents

Chapter 1. Introduction - Can something be said 'both ways'?

The author of this work has spent many an hour discussing with friends and colleagues what the correct inflected form of certain Croatian words (one of them being the author's own name) in a particular case, tense etc. was and trying to determine what the form might depend on. Each discussant had their own view on the matter and at the end of the day, we had to agree to disagree. However, it turns out that both participants in the discussion, regardless of the word being discussed, were actually right. How is that possible? In this dissertation we analyse this kind of situation, when there are several word forms expressing the same set of morphosyntactic features in the same context in one language, Croatian.

1.1. Definition of research topic and terminology

Corbett (2009) defines a canonical inflectional system as one in which there is "a unique mapping from form to function and from function to form." In a canonical inflectional paradigm each cell would be realised by a different form, a principle that harks all the way back to Humboldt's Universal (Vennemann, 1969), later expressed in the Principle of Contrast, which stated that "every difference in form must necessarily mark a difference in meaning" (Clark, 1987, p. 158). However, Corbett himself admits that canonical instances are unlikely to be frequent; in fact, "they are more likely to be rare, and may even be non-existent" (Corbett G. G., 2010, p. 142). Across the world's languages we can find deviations from the canonical situation, even more frequently than instances of canonicity. For instance, there are numerous examples of the phenomenon of *syncretism*, when one form spreads over several cells (e.g. the English form *are* is used in 2nd person singular and all three plural persons). On the opposite end of the scale, we can find two or more forms available for a single cell. When such forms are mutually exclusive and their use defined by a phonological, syntactic or some other criterion, we speak of *allomorphy*. A well-known example of allomorphy from English is the variation in the indefinite article (*a car* vs. *an elephant*). An even more complex example is the Dutch diminutive suffix, which has as many as five allomorphs (-je, -etje, -pje, -kje, -tje), whose distribution is

phonologically conditioned. However, closely connected to this is the situation "whereby a paradigmatic cell is being filled by two or more synonymous forms which realize the *same* set of morpho-syntactic properties" (Thornton (2011, p. 359), italics mine). In other words, such forms should be completely interchangeable and their distribution unconditioned. This is the situation of *morphological doubletism*, which is the main topic of this work.

This phenomenon is known by a variety of names. Thornton (2011), (2012), for instance, uses the term *overabundance.* Bermel & Knittl (2012) call such forms *competing forms*, Bauer (2014) refers to *variable outputs*, whereas Baayen, et al. (2013) join several instances of free variation in Russian under the label *rival forms*. Marković (2012) calls this phenomenon *morphological synonymy*. As we know, (lexical) synonymy can be expressed in the formula [different lexemes ↔ same meaning]. By analogy, morphological synonymy is a situation where different morphemes express the same grammatical meaning.[1] As pointed out by Edmonds and Hirst (2002, p. 106), "synonymy has often been thought of as a 'non-problem': either there are synonyms, but they are completely identical in meaning and hence easy to deal with, or there are no synonyms, in which case each word can be handled like any other."

Baayen, et al. (2013, p. 254) introduce an additional dimension to the traditional form-meaning relationship – that of *environment* (i.e. syntactic, morphological or phonological context). In such a three-dimensional space morphological doublets or synonyms would be high on the form axis, but their values on the meaning and environment axes would be close to zero. In that same space, the phenomenon which is conceptually closest to doubletism is allomorphy, in which the meaning is also the same but the environments are different.

In Poplack's (2001, p. 408) view, when grammarians are faced with such a variation, they typically respond to it by attempting to factor out the variability, "either by (1) ignoring it, (2) condemning the offending variant, or (3) attempting to redress the form-function asymmetry, typically by assigning to each form a preferred 'reading' or function. (...) This makes it possible to attribute the variability to such

[1] Using the same terminology, syncretism can also be called *morphological homonymy*, as one morpheme is used for expressing several grammatical meanings.

unobservables as *speaker intent*, and thereby explain it away." This work will not assume any of the above positions. What we will try to do is determine the distribution of doublets in language and examine their status in native speakers' mental grammars. Unlike some linguists, whose views will be presented below, we do not consider doublets a problematic feature of language, hence we will not try to invent principles for the sole purpose of introducing a functional distinction between them by any means necessary.

In Cappellaro's (2013, p. 210) view, the phenomenon of morphological doubletism is not simply an artefact of grammar books and dictionaries. Examples of morphological doubletism do appear in a number of Indo-European languages, but their number is not very large. In English, for instance, we find some verbs (not more than 20) that vary between the strong and the weak formation of the past tense (*dive > dived / dove, leap > leaped / leapt, shine > shined / shone*; see Haber (1976) for more examples). Furthermore, some nouns of Latin origin can form the plural either as they do in Latin or they can adapt to the English pattern (e.g. *cactus > cacti / cactuses, formula > formulae / formulas* etc.). Even though German plurals are often cited as a struggle for both linguists and learners of the language, with as many as eight possible endings and low predictability, remarkably few nouns are found with doublet plural forms (e.g. *Mund* 'mouth' > *Münder / Münde, Park* 'park' > *Parks / Parke, Herzog* 'duke' > *Herzoge / Herzöge*; see Morth and Dressler (2014) for more examples). Another cell where doubletism is known to appear in German is the genitive singular with examples such as *Brot* 'bread' > *Brotes / Brots, Hund* 'dog' > *Hundes / Hunds* etc. (see Fehringer (2004) for additional examples). Doubletism in Italian is mostly located in the verbal system – a few verbs can have two present (*dovere* 'must' > *devo / debbo, (po)sedere* 'possess' > *(po)siedo / (po)seggo*), two imperative (*andare* 'to go' > *vai / va', dare* 'to give' > *dai / da'*) or two past participle forms (*vedere* 'to see' > *visto / veduto, perdere* 'to lose' > *perso / perduto,* see Thornton (2011) for additional examples). Similarly, some Spanish verbs can have two forms in the imperfect subjunctive (*ser* 'to be' > *fuere / fuese, cantar* 'to sing' > *cantara / cantase,* see Guzmán Naranjo (2016) for additional examples).

It would sound logical to assume that the richer the morphological system of a language, the more instances of doubletism it would tend to show. Since Slavonic languages inherited from Proto-Indo-European (PIE) a rich morphological system

characterised by fusional endings, based on the above assumption, we would expect morphological doubletism to be quite widespread in the Slavonic world. Without going into much detail about their history, we present Janda's (2014) account of the emergence of doubletism across the Slavonic languages. Namely, the number of declension classes in modern Slavonic languages is much lower than their number in Proto-Slavonic (as reconstructed from its oldest written variety, known as Old Church Slavonic). After the disappearance of some inflectional patterns,

> neither the nouns nor the desinences that belonged to these patterns were entirely lost; for the most part they were reclassified or reconceptualised (…) The move from a moribund paradigm to a productive one was a gradual process, and nouns often brought some morphological 'baggage' along with them. Desinences from the 'old' paradigm could tag along as alternative endings for the immigrant noun in its 'new' paradigm. In many instances the 'old' desinences that were imported along with their nouns were adopted and became productive in the 'new' paradigms, often serving new distinctive purposes (Janda, 2014, p. 1577).

One widespread instance of doubletism in the Slavonic world (in terms of encompassing a large number of lexemes)[2] is the variation in the locative singular of masculine nouns in Czech, the two options being the conservative ending -ě and the innovative -u (*hrad* 'castle' > *hradě / hradu* etc.). Another widespread doubletism in Czech, although involving a much smaller number of lexemes, also appears in the genitive singular, with -a and -u as the two variants (*sýr* 'cheese' > *sýra / sýru*).[3] The same case, genitive singular, is also the *locus* of doubletism in Polish (*żurnal* 'fashion magazine' > *żurnala / żurnalu*).

Croatian is a South Slavonic language and just like other Slavonic languages it has a rich morphological system – seven cases, two numbers, three genders, four nominal declension classes etc. When it comes to matters of doubletism, it does not

[2] Cummins' (1995) analysis identifies over 1,200 lexemes showing (or having the potential for) this variation.

[3] Bermel & Knittl's (2012) corpus analysis identifies 112 nouns attested with both endings.

lag behind other members of the Slavonic family. Even though Cummins (1995, p. 242) argues that "no Slavic language has the remarkable declensional variability of Literary Czech," we would dare to say that standard written Croatian exhibits high levels of doubletism matching or exceeding that found in other Slavonic languages, which reflects its turbulent diachronic development as well as lack of agreement on issues of grammar by linguistic authorities. Each declension class contains at least two cells in which doubletism appears; on top of that, it is also widespread in the verbal system (which will not be dealt with in this work).[4] A detailed overview of the inflection classes and instances of doubletism within them is given in Chapter 4, which, we believe, will give credit to the above claim.

As far as we are aware, there have been very few analyses of this phenomenon, not only in Croatian, but in other Indo-European languages as well. One of the main reasons for this is that the traditional view in linguistics was that "to understand how a language is truly composed and how it works, we must clear the ground of all that is irrelevant to theoretical purposes and, therefore, first and foremost, of variation" (Berruto, 2004, p. 295).

Amongst the rare works on this topic we came across were Kottum (1981) on variation in Polish genitives, Bermel (1993) and Cummins (1995) on Czech locatives, all written in the period when generative theories were still predominant in linguistic science. These and similar works primarily try to find a reliable criterion for the differentiation of the two forms (phonological, syntactic, semantic, sociolinguistic etc.) but mostly fail because they pursue what we believe to be an initially flawed assumption – that a functional distinction is necessary in the first place as that would be in accordance with the belief that underlay most of the early generative work (later expressed in the Principle of Contrast) – that variation has to be motivated.

Cognitive linguistics challenged dichotomies that were taken as given by proponents of generative syntax (and structuralism before that), among others the distinctions between grammaticality and ungrammaticality, category membership and non-membership, and finally grammar and usage (i.e. competence and performance).

[4] We will not deal with derivational morphology in this work either, even though this part of Croatian shows a remarkable degree of doubletism as well.

All of these are now viewed as continua, rather than dichotomies.[5] Language structure in Cognitive Linguistics is viewed as a dynamic concept (fluid and temporary), which is in contrast to it traditionally being conceptualized as a building (i.e. stable and fixed).[6] Dynamicity of language is reflected in the fact that language is constantly changing and that no two speakers share the same system. This is so because no two speakers have the same experience of language – they receive different input. For instance, people growing up in different countries of the English-speaking world will naturally be exposed to different experiences; however, this will not stand in the way of their mutual comprehension. Does such a view of language in which the underlying system differs from one individual to another have any detrimental consequences for the system itself? Can we call such a language a system at all? As Langacker (2010, p. 125) explains, "the fact that every rock is different does not mean that there is no such a thing as a rock". Rather, variability can be considered to cause problems for a system only if it is shown to be completely non-systematic (and it is this kind of variability that the Principle of Contrast also tried to eliminate). The basic idea behind the majority of William Labov's work was to show that variability is also systematic. The only way to determine the amount of systematicity in a linguistic system is empirically. This is what we attempt to do in the present dissertation.

Even though the use of empirical methods has been present in certain areas of linguistics, such as phonology and language acquisition as early as 1950s, it started penetrating other linguistic areas, most notably morphology and syntax, only in the 1990s. This was facilitated by the technological revolution, which provided linguists with many useful tools for collecting and analyzing morphosyntactic data. Prior to that, it was common belief that "linguistics does not need empirical data, and that it gets on faster and more efficiently if it bypasses painstaking observation of natural usage and relies instead on speakers' intuitive 'knowledge' of their language" (Sampson, 2005, p. 16). However, as argued by the same author, "surely we would think it strange if, say, physicists based their laws of motion on the fact that they

[5] For this reason, Newmeyer (2003, p. 682) calls all usage-based approaches, of which Cognitive Linguistics is a part, anti-Saussurean as it was Ferdinand de Saussure who introduced most of these dichotomies.

[6] Many usage-based linguists (e.g. Bybee (2001)) refer to language as a Complex Adaptive System.

'knew' without experiment that balls of different weights released from a tower simultaneously would hit the ground at the same time" (Sampson, 2007, p. 12).

1.2. Research questions

R. Ellis (1999, p. 467) argues that there are two logically possible starting points in any study of free variation: 1) we can assume that the variation is systematic until shown otherwise or 2) we can assume it to be free until it is shown to be systematic. The author references many scholars who assumed the former position. We believe, on the other hand, that the latter view is more appropriate as the null hypothesis. Hence this is also the starting point of our work. Kilgariff (2005, p. 264) defines four possibilities of how two phenomena can be associated: they can be Random, Arbitrary, Motivated, or Predictable. The author strongly argues that language is never random, hence any null hypothesis that assumes so will necessarily not be confirmed. We will try to determine into which of these four categories the Croatian doublets we retrieve can be placed. As argued by Bybee (2010, p. 97), "lexemes do not occur in corpora by pure chance. Every lexeme was chosen by a speaker in a particular context for a particular reason." We believe that the choice of doublets can be, at least partly, accounted for by using reasons which we describe below, the most important being the frequency of the form in the input, thus making them, in Kilgariff's terms, Motivated.

Cognitive linguists are interested in retrieving the linguistic system existing in the mind of an individual speaker, which is often termed *mental (lexico-)grammar* (see Langacker (1987), Halliday (1991a), Divjak (2006) etc.).[7] The emphasis here is on the word *individual speaker*, which is in direct opposition to the generative concept of an *ideal speaker/listener*. One of the most pressing questions cognitive linguists try to answer is how brain and the mind interact in acquiring, comprehending and producing language. This work will go in the same direction. We are interested in discovering how linguistic (more specifically morphological) knowledge is encoded in the brain,

[7] This term is another example of how cognitive linguists re-define traditional linguistic dichotomies. Whereas generative linguists viewed grammar and lexicon as separate domains of language, cognitive linguists view them as forming a continuum. For instance, Halliday (1992, p. 63) believes that grammar and vocabulary are "the same thing seen by two different observers."

i.e. what types of linguistic units are stored in our mental grammars and how these units are organized. Current models all assume that individual lexical items are stored as units and form the basis of our mental lexicons. However, when it comes to morphology, the question arises in what form the various inflected forms of those lexemes are stored. Decompositional (or *morpheme-based*) theories argue that the root and all the inflectional morphemes are stored separately from each other and later connected in actual language production. This merging takes place using either: a) simple concatenation (*Item-and-Arrangement* model), where the meaning of the word is a combination of meanings of the constituent morphemes (e.g. *bake* 'the act of baking' + *-ed* 'past tense' → *baked*), or b) by means of operations on lexemes (*Item-and-Process* model), where the semantics is a function of the *Word Formation Rule* (WFR) mapped directly from input to output (e.g. /X/ → /Xed/). Lexical (or *word-based*) theories, on the other hand, do not recognize any morphological units smaller than words; rather, "roots, stems, and exponents are abstractions over a set of full forms" (Blevins, 2006, p. 532). All inflected forms are stored individually in lexical memory and are related to each other by various types of connections – in *Word-and-Paradigm* theory, words are analysed in relation to their inflectional paradigm in such a way that "lexemes form one tree, and each lexeme then comes with its own separated disjoint tree for its inflected variants" (Milin, Kuperman, Kostić, & Baayen, 2009, p. 233). In the *Network Model* (Bybee (1985)), words form lexical connections to other words with the same semantic or phonological parts (e.g. the word *cats* is connected with *cat* as a member of its paradigm, but also with all the words that form the plural with -s). These connections in the Network Model create schemas, into which new words can be fitted. Furthermore, each word has its lexical strength, which increases every time it is experienced in language.[8] The *Satellite-Entry Hypothesis*, developed by Lukatela, et al. (1980), seems particularly apt for describing languages with an elaborate case system, such as the Slavonic family.[9] In this system, nouns are represented in a satellite-like fashion, such that the nominative form is the 'access' form and other case forms are at an equal distance from the nominative form.

[8] Note that this is the only model of all that have been mentioned so far that uses frequency as a built-in factor.

[9] As a matter of fact, this model was originally tested on Serbian, both by Lukatela, et al. (1980) and Feldman & Fowler (1987).

The authors, however, argue that it is possible that different areas of the lexicon (e.g. verbs vs. nouns) have different organizational structure. Finally, generative linguists traditionally argue for a *Dual-Mechanism* approach, in which regular and irregular forms are accessed using different mechanisms, the former being derived by rule – hence part of grammar, and the latter lexically stored – hence belonging to the lexicon (see e.g. Ullman (1999)).

Bauer (2014, p. 97) explains how variable outputs would be handled by the theories above.

> If this is done simply by allowing two or more rules to specify different outcomes from the same input, we are leaving a lot to the interpretation of the rules. If it is done with some kind of formula we cannot necessarily predict the outcome on any given occasion. (…) Using stored exemplars to predict new forms begins with a denial that there is a single input form. Rather, multiple factors may be important (including the frequency of the bases involved, degrees of phonological similarity with other bases, semantic content, pragmatic value).

The results presented in this dissertation give credit to lexical theories, similar to Fehringer (2004, p. 324), who concludes that it is necessary to regard all doublets as being lexically listed as whole words within a network of connections, "with information on the preference of doublet signalled in some way." Our goal, then, is to determine the way in which this preference is signalled. We argue that a piece of frequency information appears in the mental grammars alongside the lexeme itself and its inflectional variants (as per word-based theories). This is in line with emergentist theories of language, which claim that "knowledge of language includes not just knowledge of syntactic, morphological, and phonological categories, but also knowledge of the frequency and probability of use of these categories in speakers' experience" (Tily, et al., 2009, p. 149).[10] Most recently, Taylor (2012) argues that language users keep track of the utterances they encounter, thereby compiling a "mental corpus" of constructions at various levels of abstraction.

[10] Halliday (1991a, p. 32) illustrates this feature with a simple example. The author believes that the meaning of the concept 'negative' is not coded in our mental grammars simply as 'not positive', but rather as 'not positive, with the odds 9:1'.

A multitude of works in cognitive science have shown that humans are very good at evaluating frequencies of things surrounding them.[11] For instance, if someone asked you how many movies you had seen in the previous year, you would be able to provide a fairly accurate answer. Furthermore, Grant, Hake, & Hornseth (1951) provide experimental evidence showing that people are successful in internalizing patterns of statistical variation. In their experiment, subjects were presented with a light which blinked at random, but with various probabilities. Subjects had to guess whether the light would next be on or off and, after sufficient experience, their guesses closely approximated the underlying probability of the light being on. However, as pointed out by Hasher & Zacks (1984, p. 1374), "despite empirical evidence regarding the quality of stored frequency information, most people have no awareness of having this information and so have little confidence in their potential accuracy in tasks based on frequency knowledge."

It would not be surprising if language was also susceptible to perceptions of frequency. Halliday (1991a, p. 35) argues that "those who know a language can make an informed guess about the relative frequency of its words. (…) This is just an aspect of knowing the language." For instance, most English speakers would accurately recognise that *go* is more frequent than *walk* and *walk* more frequent than *stroll,* or that active was more frequent than passive and so forth.[12] Saffran's (2003) results showed that human learners, both adults and infants, can use statistics to find word boundaries. Additionally, Bod (1998), Arnon & Snider (2010), Gurevich, Johnson, & Goldberg (2010) have all provided evidence that speakers encode the frequencies of multi-word units. This information is extracted from thousands and thousands of hours of exposure to one's native language. More importantly, as argued by Bybee (2010, p. 10), "no types of data are excluded from consideration as they all represent performance." What makes the storage of such a massive amount of information possible is the large memory capacity of our brain. As pointed out by Householder

[11] In fact, Hasher & Zacks (1984, p. 1378) call the encoding of frequency "a fundamental cognitive process for which we are attuned early in life."

[12] However, Baayen, Milin, & Ramscar (2016) argue out that they would be unable to provide estimates of pure frequency in such a case. Similar to this, McEnery & Wilson believe that "human beings have only the vaguest notion of the frequency of a construct or a word" (2001, p. 15) and that they are "poor sources of quantitative data" (2001, p. 134).

(1966, p. 100), "our brains, unlike most computers, have no need for economizing with storage space."

Bybee (2010, p. 122) also argues that "quantitative distributions matter and are part of the grammar."[13] The question we ask in this dissertation is what kind of quantitative information is actually stored in the brain (which number, put simply)? Some linguists would argue in absolute terms, that our brains are like counters and that every single *token* is imprinted in memory. For instance, Hoey (2005, p. 11) believes that our mind has "a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context." N. Ellis (2002, p. 146), on the other hand, is not convinced that we spend our time consciously counting the units of language, but rather that our knowledge of the underlying distributions and their most usual settings is implicit.[14]

Bybee (1995) and her collaborators (Bybee & Newman, 1995) have all put forward evidence that it is *type frequency*, rather than token frequency, that has the greatest influence on people's intuitions (about novel words). In this *Exemplar Model*, only distinct items are noted in memory and every subsequent encounter with the same item does not change anything in the representation. Tomasello (2003, p. 327), on the other hand, suggests that the process whereby language structure emerges from language use depends on both type and token frequency, in such a way that type frequency registers items, whereas token frequency entrenches them. Moscoso del Prado Martin, Kostić, & Baayen (2004) have provided evidence that *family frequencies* (the number of different morphological forms that exist for a given word or morpheme) work even better than type frequencies alone.

Saffran (2003) argues that human beings are capable of much more intricate forms of statistical learning other than simple counts, which has led researchers to explore other frequency measures, counting both individual words and constructions.

---

[13] In a slightly different manner, Dąbrowska (2004, p. 28) believes that "information about frequency is stored separately from linguistic knowledge proper, as some kind of an appendix to the grammar."

[14] Similar to this, Hasher & Zacks (1984, p. 1374) argue that "when we ask people to estimate the extent of their knowledge of the frequencies of such things as individual syllables or occupations, they generally express surprise at the question itself; they think they lack the information necessary to answer. (…) It is only when subjects begin to respond to specific test items that they realize they actually do have considerable knowledge about the number of occurrences of specific events."

Erker & Guy (2012) examine how various frequency counts contribute to the omission of the subject pronoun in Spanish – besides *raw* (i.e. token) *frequency*, they also distinguish between *log frequency* (calculated as the base-10 logarithm of the raw frequency) and *discrete frequency* (a binary category with 'frequent' and 'infrequent' as its values; frequent items were those constituting at least 1 percent of the corpus). The respective correlations bring opposite results: the more frequently a form occurs (in absolute terms), the more likely it is to be used with a pronoun, whereas higher log frequency is associated with lower pronoun usage.

Other authors have raised arguments in favour of using frequencies of word combinations instead of individual word frequencies (e.g. Stefanowitsch & Gries (2003) on collostructional strength, Gries & Stefanowitsch (2004) on distinctive collexeme analysis, Krug (1998) on String Frequency etc). Divjak (2008) contrasts several measures of relative frequency[15] to predict speakers' behaviour, such as *attraction* (the degree to which a lexico-grammatical pattern attracts a verb relative to all other verbs competing on the paradigmatic axis), *reliance* (the degree to which a verb relies on a lexico-grammatical pattern relative to the occurrence of the same verb in other patterns), *collostructional strength* (co-occurrence frequencies of verbs and constructions) etc. Divjak concludes that raw frequency alone is not sufficient, but rather that relative frequencies are also needed (2008, p. 231). Various studies (e.g. (Hay, 2001)) have demonstrated the effect of relative frequency on the transparency of morphologically complex forms – complex forms that are more frequent than their bases are perceived as less decomposable than forms that are less frequent than their bases. Raymond & Brown's (2012) analysis of initial fricative reduction in Spanish has shown that once a variety of contextual frequency measures is taken into consideration, the non-contextual measures lose explanatory value.

Another group of authors (see Moscoso del Prado Martin, Kostić, & Baayen (2004), Milin, et al. (2009) etc.) use measures inherited from information theory, such as *relative entropy* (measure of the information gained when one revises one's beliefs from the prior probability distribution Q to the posterior probability distribution P or,

[15] Relative frequency is defined in one of two ways: as a normalised value whereby data from different corpora can be compared to each other (e.g. incidences per million tokens) or the difference between the frequency of a form and another related frequency (such as its base, a construction it is a part of etc.)

put more simply, how different the exemplar is from the prototype), *information residual* (the amount of information carried by a word as estimated from its relative frequency) etc.[16] Stefanowitsch (2005) argues that the only way for observed frequencies to become relevant facts for scientific analysis is if they are evaluated against their *expected frequencies of occurrence* (the author calls this *the expected frequency epiphany*). And finally, various studies advance the argument that (*conditional) probabilities* are a more appropriate metric than frequencies. For instance, Jusczyk, Houston, & Newsome (1999) have shown that infants have a remarkable ability to compute transitional probabilities for phonemes (given the phoneme X, what is the likelihood that the next phoneme will be Y?).[17]

Although it appears to represent a deceptively simple concept, frequency of occurrence in language, as pointed out by Baayen, Milin, & Ramscar (2016, p. 2), "actually turns out to be a remarkably complex construct that is entangled with a large set of highly collinear lexical random variables, and that has been interpreted in many different ways." With the diversity of evidence presented above it is reasonable to assume that perhaps different kinds of frequency predict different kinds of behaviour of speakers. In this work we examine yet another measure of frequency that has not been researched by linguists as often as the ones above – proportions. The use of proportions in the decision-making process, both in humans and other animals, has been demonstrated in numerous experiments in domains other than language. For instance, in Longo's (1964) experiment, a cockroach, when shocked 30% of the time in one arm of a T-maze and 70% of the time in the other, would pick the arm where shocks are less likely 70% of the time. Similar to this, in one of Estes' (1976) experiments on category membership subjects were told 70% of the time a stimulus belongs to category A and 30% to category B. In a later assignment task, they assigned A to 70% of the new items and B to 30% (the categories and items were arbitrary, so

[16] For instance, Milin, et al. (2009, p. 215) believe information theory "offers exactly the right tools for studying the processing consequences of paradigmatic relations. The use of these tools does not imply that we think the mental lexicon is organized in terms of optimally coded bit streams."

[17] See articles in Gries & Divjak (2012) and Divjak & Gries (2012) for additional state-of-the-art investigations of various frequency effects. However, also note authors who diminish the role of frequency, e.g. Baayen (2010, p. 437), who says that frequency of occurrence, "when understood in the sense of repeated experience, plays only a minor role in lexical processing. (…) [Rather], the word frequency effect is an epiphenomenon of learning to link form to lexical meaning."

they could not use any inherent properties). In this work we conduct several experiments where proportions are used as the primary frequency measure. The purpose of these experiments will be to examine whether speakers keep track of relative proportions of the two forms in competition just like they keep track of other frequency measures mentioned in the discussion above. Similar views have been expressed by e.g. Haber (1976, p. 231) and more recently by Fehringer (2011, p. 103).

Linguists have traditionally assumed that the continuous spectrum of acceptability that constantly emerges from intuition studies is caused by extra-grammatical (performance) factors, such as plausibility, working memory limitations, ambiguities etc. (see Sprouse (2007)). What we attempt to show is that it is due instead to frequency considerations. In other words, doublet forms that appear in the language with a 50: 50 distribution will not have the same status in our mental grammars as those that appear with a 70: 30 distribution nor as those with a 99: 1 distribution. On the other hand, all items that appear with a distribution of e.g. 95: 5 should show the same behaviour in native speakers, regardless of whether this proportion reflects a cumulative frequency of 20 or 20,000. However, we will not argue that frequencies are the only thing that shapes our mental grammar. As Roeper (2011, p. 24) points out, "notions like similarity are where the real mental talents are hidden." We also discover effects which we call *family effects*, but which in literature are also referred to as *analogical* or *gang* effects. We believe that a 'hybrid' model that combines frequency information and similarity relations best describes the processes taking place in the morphological processing of doubletism.

We will attempt to get to this model by comparing and contrasting data collected from a variety of sources in order to see whether the same conclusions about speakers' mental grammars can be drawn from all of them. Whereas observable data were not traditionally part of the generative programme, which relied more on introspective methods, usage-based linguists incorporate observations of external data into their inferences about mental grammars. In this view, production data can serve as "a window to the mind" (Gilquin, 2010). Large computer corpora are used as a source of production data such that analyses of them can yield hypotheses about our questions about human cognition, which can then be further tested through additional corpus or experimental investigation. This work offers new insights into both of these methods, expanding the correlation between corpus and intuition data with predictions

of several computational models which claim to replicate the behaviour of real speakers. Previous work using such a multi-methodological approach has already determined that "the frequency with which features appear in the corpus could mirror the frequency with which native speakers produce them in ordinary speech or writing situations" (Bermel & Knittl, 2012, p. 241), thus disproving the generative position that there is no connection between usage patterns in corpora (which reflect conventionalized language use) and entrenched linguistic knowledge.

1.3. Dissertation overview

In the next chapter we touch upon the long-standing debate between generative and cognitive linguistics. We are especially interested in determining how doubletism is treated in each of these traditions. Whereas generative linguistics in general had a negative approach to doublets, considering them irrelevant for linguistic analysis, cognitive linguists study them to a much greater extent; however, they mostly study alternative syntactic constructions – there has been very little research on morphological doublets. The final section of Chapter 2 also examines how doublets have previously been treated by Croatian linguists.

Chapter 3 discusses various ways in which linguistic data have traditionally been collected, with a special focus on intuition judgments and corpus data, both with their advantages and weaknesses. Instead of making a strict distinction between the two, we argue for a pluralistic approach to methodology, in which combining various types of evidence can give us more information about mental grammars than any single one.

The following three chapters analyse specific examples of Croatian doublets, each using a different methodological approach. Chapter 4 presents raw corpus data. We identify cells in the inflectional paradigms of Croatian nouns, adjectives, and pronouns which exhibit doubletism. For each of those, we try to determine whether the variation is conditioned or free. In Chapter 5 we analyse two of the cells from the previous chapter by means of two computational, exemplar-based models (Analogical Model – AM and Tilburg Memory-Based Learner – TiMBL). This approach is used as an alternative to traditional rule-based approaches for describing inflection. Finally, in Chapter 6 we report the results of three questionnaire studies conducted over the

course of two years among native speakers of Croatian. Two of these are acceptability studies, whereas the final one is a forced-choice study. The common theme running through all three studies is that the distributions of doublet forms in language (represented in the corpus) substantially affect our judgments and selection of the respective forms.

Chapter 2. What can doublets tell us about mental grammars?

Two major linguistic schools of thought have been competing in the past 50 years, each providing their own theories of the nature of human language. Generative and cognitive linguistics both share a mentalist approach to language, in that they are both interested in how language is represented in the mind. However, as nicely summarised by Gilquin (2010, p. 89), "the human mind is a black box. We know what goes into the box and what comes out, but we do not know what happens inside." Since it is impossible to penetrate this black box directly, it has to be done using indirect probing methods. In the next two chapters, we analyse generativism and cognitivism in more detail. First, in this chapter we show how both linguistic schools see language structure, with a special emphasis on the place of doubletism within it. In the following chapter, we present the methods both theories use in trying to penetrate the 'black box'.

2.1. Introduction to linguistic theories

Generative linguistics makes a firm distinction between language *competence* (unconscious knowledge a speaker has about his language) and language *performance* (how that language is used), terms roughly equivalent to classical structuralist terms *langue* and *parole*. The first of these elements is, in the generative view, innate to all humans – infants are born with a knowledge of what human languages are like. As described by Dąbrowska (2004, p. 58), this is "a mental module specifying the universal properties of human languages and the parameters along which they vary", called Universal Grammar. The only knowledge children need to acquire are the details of the language they are learning, for which Chomsky postulated a specific *Language Acquisition Device* (LAD), "a specific mental organ (a structure in the brain) which is dedicated to extracting from haphazard and often degenerate speech the generalizations required for the child to construct the necessary rules of phonology and grammar" (Trask, 1999, p. 94). As evidence for the existence of LAD, generativists have pointed to numerous studies showing that all children go through similar stages of development. For instance, English children have been shown to acquire grammatical morphemes in roughly the same order: first they acquire the -ing

marker, then the prepositions *on* and *in*, then the plural marker (-s), followed by the possessive marker (-'s) etc. Furthermore, this innate 'language faculty' is viewed by generativists as a distinct part of our genetic system, independent from other cognitive capacities, a hypothesis supported by the nature of certain genetically based disabilities, which disrupt language while affecting little else (e.g Broca's aphasia), or which leave language largely unaffected while disrupting most other cognitive abilities (e.g Williams' syndrome).[18] However, cognitive linguists and neurolinguists have provided even more evidence demonstrating that language acquisition is not as uniform as assumed by generativists and that children who hear less language develop it more slowly (see references for both sides of the argument in Dąbrowska (2004, pp. 29-39)). Furthermore, Dąbrowska (2004, p. 55) claims that dissociations between language and cognition are partial – some cognitive defects will have a profound effect on language performance, while others will be of little or no consequence.

Since in a traditional generative account language is acquired, not learned, input is secondary for the development of language, hence language structure can be described independently of it.[19] One of the main arguments used to argue against the importance of input in language acquisition has been the concept of the *poverty of the stimulus*. In Chomsky's (1957, p. 15) view, speakers' grammatical competence surpasses the information available in the input – humans are able to produce and understand any number of utterances despite having had only finite and haphazard exposure to language. Furthermore, input can inform learners only about well-formed sentences and cannot give any information about ungrammaticality.[20]

---

[18] Trask (1999, p. 95) points out that in recent years, Chomsky himself has seemingly abandoned his arguments for the LAD in favour of an even stronger and deeply controversial claim, the so-called *parameter-setting model*: "he now believes that so much information about the nature of human language is already present in our brains at birth that all the child has to do is to 'set a few switches' to the correct values for the language being acquired."

[19] However, note later generative approaches, such as Yang (2004), who argues for a division of labour between endowment and learning. The author does acknowledge statistical learning as a factor in acquisition, but thinks it should not be given the primary role.

[20] This is known as the *no negative evidence problem* in linguistics. Children, the argument goes, are only ever presented with positive evidence – they only hear sentences that are grammatical and never those that are ungrammatical. The fact that they also avoid ungrammatical sentences in production shows that they have an innate ability to distinguish between the two regardless of the input (see e.g. Stefanowitsch (2008) for usage-based approaches to the same problem).

Cognitive linguistics, on the other hand, argues that language needs to be viewed as part of the whole human cognitive system, a product of both general cognitive abilities and an interaction with our environment rather than as an innate mental module. This is what makes cognitive linguistics fundamentally interdisciplinary in nature. In general, cognitive linguistics can be subsumed under the wider term *usage-based* approaches to language. Usage-based theory developed out of American functionalism – the first usage-based linguists were typologists (e.g. Joseph Greenberg and others), who were using frequency effects to explain cross-linguistic patterns. As the name suggests, usage-based theories of language claim that language structure emerges from language use (hence the alternative name *emergentist* theories of language) – as nicely put by Bates & MacWhinney (1988, p. 147), "language is a new machine built out of old parts" (where 'old parts' refers to previous instances of language use). In this view, there is no distinction between knowledge of language and use of language – rather, the knowledge of language is the knowledge of how language is used (as opposed to the knowledge of language structure, as argued by generativists). Furthermore, there is no specific language acquisition device; rather, "the cognitive mechanism underlying acquisition is simply a processor – the same processor that is responsible for interpreting and forming sentences in real time during actual language use" (O'Grady, 2008, p. 456). Dąbrowska's (2004: 213) description of this processor is the following: "each usage event leaves a trace in the processing system. Every time a unit is accessed, its representation is strengthened (entrenched), so units which are accessed more frequently become easier to activate."[21] However, as pointed out by N. Ellis (2012, p. 7), it is not only frequency of exposure that leads to entrenchment, but also recency – "the more recently we have experienced something, the stronger our memory of it, and the more fluently it is accessed."[22] In sum, language seen through the lens of usage-based theories is not acquired, but rather invented on the basis of input received.

[21] This mechanism is by no means particular to language. Repetition of virtually any task results in routinization. For instance, an analogous example of entrenchment (provided by Roeper (2011)) can be found in human anatomy, where human muscles grow fatter with use.

[22] As a matter of fact, Langacker (1987, p. 59) proposes that "extended periods of disuse have a negative impact on entrenchment."

The question we try to answer, then, is whether this input is incorporated into the mental grammars and in what form. Traditional generative approaches claim that usage and frequency are excluded from grammar per se.[23] For instance, Newmeyer (2003, pp. 697-8) will agree with cognitivists that language users and hence their grammars are sensitive to frequency. "But from the fact that Y is sensitive to X, it does not follow that X is part of the same system that characterizes Y." In other words, even though speakers are sensitive to quantitative information, their grammars nevertheless still consist only of categorical information. In a similar vein, Roeper (2011, p. 23) claims that learning inherently involves adding information but that the addition of information is quite separate from counting information. "Whereas learning necessarily changes the mental representation of an item, counting the instances of X does not change X itself." Furthermore, Roeper believes that all the previous studies that have shown frequency to be operative was because they were counting the wrong things or, more specifically, at the wrong level of magnification. This work attempts to oppose this view by showing that information about frequency is indeed stored in the mental grammars, in the form of relative proportions.

The two linguistic theories also share opposing views on the matter of grammaticality. In a traditional generative view, grammaticality is almost always defined in categorical terms – our mental grammar distinguishes only two kinds of strings: those that are possible sentences of our language (i.e. *grammatical*) and those that are not (i.e. *ungrammatical*). However, N. Ellis (2002, p. 162) points out that grammatical categorization is no different from other types of categorizations – and authors such as Rosch (1978), Taylor (1989) have all argued that humans categorize in a gradient rather than a binary manner[24] – hence the question whether something is grammatical or not should also be approached in a gradient manner. Featherston (2005b, p. 190) believes that "we must have a model of well-formedness as a continuum, on which there is not only good and bad, but also good *and* better." Later

[23] Generativists also demote the role of memory. According to Roeper (2011, p. 43), "the properties of memory are so different in different domains (vision, language, social) that one can challenge the idea that the concept represents something coherent."

[24] The most impressive and often cited non-linguistic example of people imposing gradience on a concept traditionally seen as dichotomous comes from Armstrong, Gleitman, & Gleitman (1983), whose participants had no problems rating odd and even numbers as more or less typical of their respective categories.

generative approaches do allow for the grammaticality/ungrammaticality distinction to be gradient rather than categorical, using labels such as '**', '?', '??', '*?' etc. alongside the traditional marker of ungrammaticality – '*' (see e.g. Andrews (1990)).[25] Sorace & Keller (2005) argue that even though grammatical constraints are violable, not every violation is fatal, which, in turn, leads to gradience in grammaticality.

Sampson (2007), however, argues for a complete abandonment of the grammatical/ungrammatical distinction. The author brings forward quite a radical proposal – that nothing in language is ungrammatical. Applying the logic of the famous maxim attributed to William Butler Yeats 'There are no strangers, only friends I haven't met yet', Sampson comes to the conclusion that there are no ungrammatical examples, only those we have not come across yet – once we see them, we can no longer regard them as wrong. Geoffrey K. Pullum, in a reply to the above article by Sampson, offers a completely opposite view – that "*almost all* strings, whether of words or lexical categories, are *un*grammatical" (Pullum (2007, p. 41), italics original). The author explains his logic in probabilistic terms – the probability of getting a grammatical sentence using N randomly selected word tokens goes down as N goes up.[26]

Having presented the basic premises of the two major linguistic theories, we move on to reviewing previous works written within those two traditions that deal specifically with doubletism and related matters.

[25] Adli (2005) notes that Chomsky originally also assumed a graded nature of grammaticality, in which the term 'grammatical' is necessarily absolute, but 'ungrammaticality' has various degrees, depending on which principle was violated. However, as pointed out by Adli (2005, p. 6), Chomsky soon abandoned this view in favour of a distinction between grammaticality and acceptability.

[26] Korecky-Kröll, et al. (2012, p. 37) avoid the terms *grammatical* and *ungrammatical* by introducing a distinction between *actual*, *potential* and *illegal* forms. Potential forms are "those that do not exist in the language, but contain a pattern that is attested in the language." The authors believe such a distinction "provides a vantage point from which we can better understand the nature of native speaker knowledge in this domain of morphology."

2.2. Previous treatments of doubletism in linguistic theory

2.2.1. Doubletism in the generative tradition

In Berruto's (2004, p. 293) words, for some linguists "variation represents an element of disturbance, something that seems to obscure the true perception of things." Instead, what is deemed a worthy object of linguistic study is "that which is constant, invariable (...) and independent from the speaker's actuation." Generative linguistics in its original conception (later developments are explored below) is focused on the ideal speaker/listener of a homogeneous speech community and their competence rather than performance – in competence there is no room for variation. Since generative linguistics is concerned with defining the limits of 'core grammar', i.e. the architecture of the human language, variation has been cast aside to the linguistic periphery (i.e. lexicon). As Haber (1975, p. 240) puts it, in generative linguistics variation is "an uncomfortable fact uncomfortably accepted."

Philosophers such as Quine (1951) argue that true synonymy is impossible, primarily because it is impossible to define. One of the earliest expressions of this view in linguistics was Vennemann's (1969) definition of the *Humboldt's Universal*, which can be summarised in the formula *one meaning – one form*. Drawing on this universal, members of the Geneva School of Linguistics (followers of de Saussure) argued that "there are no expressions in language that would be completely identical in their functions (…) If two words really have the same intellectual sense, they necessarily belong to two different speech plans" (Jakobson, 1932 [1974], p. 14), translation mine).[27] This view was later inherited by generative linguistics and expressed in the aforementioned *Principle of Contrast* (Clark, 1987). Other authors within the same tradition defined additional principles, all of which implied that absolute synonymy or absolute free variation cannot exist, i.e. even synonyms that do exist necessarily show some contrast, either in dialect, register, connotation, perspective etc. For instance, Aronoff's (1976) *Morphological Blocking Principle* applies in situations when an irregular form is already present in the lexicon and thus blocks the appearance of a regular form to express the same function (e.g. regularly

[27] Serbo-Croatian original: "U jeziku ne postoje izrazi koji bi po svojim funkcijama bili apsolutno istovetni. (…) Ako dvije riječi i imaju zbilja isti intelektualni smisao, one neizbježno pripadaju dvama različitim govornim planovima."

formed word *stealer* is blocked by the existence of *thief*). According to Carstairs' (1987, p. 28) *Inflectional Parsimony Principle*, for every combination of morphosyntactic properties to a given word class, each word in that class will have one *and only one* inflectional realization (italics mine). Giegerich (2001, p. 65) defines the *Blocking Effect* as "the non-occurrence of a morphologically complex form due to the existence of a simpler semantic rival." Finally, Goldberg defines the *Principle of No Synonymy* in syntax, which states that "if two constructions are syntactically distinct, they must be semantically or pragmatically distinct" (quoted in Klavan (2012, p. 28)).

Faced with instances of doubletism in Polish stems, Cameron-Faulkner & Carstairs-McCarthy (2000) argue that either the Principle of Contrast does not apply to inflectional affixes (but only to lexemes) or the synonymy of these forms is just apparent. Namely, if we re-define what constitutes the 'meaning' of an affix, then no two affixes will be synonymous. The authors argue that once we include information about the inflectional class (paradigm) into the meaning of every individual affix, the majority of forms stop being in competition and the Principle of Contrast can be successfully extended to inflection as well. However, even such a modification does not account for the situation we are faced with in Croatian. The respective doublets that we present in Chapter 4 occupy the same cell of the *same* inflectional class.

Bauer (2006), however, argues that it is not the coinage of synonyms that is prevented by the Blocking Principle, but rather their institutionalisation. In other words, if a new word form that was absolutely synonymous with an existing one was to be created, it most definitely would not be accepted by the language (it would be pre-empted). The only case in which such an innovation would be permitted is when there is a lexical gap, either momentary (when we cannot think of a particular word) or long-term (when one does not exist). Due to such beliefs, there has not been much work on free variation in morphology. If there was, the variation was mostly explained using sociolinguistic criteria. For instance, one of Nida's (1948, pp. 431-2) examples of 'fluctuation of forms' is the English variation *have shown* / *have showed*. The author concludes that both allomorphs may occur in the same person's speech, but the latter occurs in more colloquial sociolinguistic environments. The author classifies this example as an instance of 'overlap', i.e. forms which are in complementary

distribution except at certain points where there is a contrast resulting from fluctuation of forms.

William Labov (1969), (1978) was one of the first scholars to regard variability as a central aspect of language competence (he calls it *inherent variability*). Fluent speakers have many variants as part of their grammatical competence, and they employ them differently depending on the circumstances of the conversational interaction (such as social positions of the interlocutors, formality of the situation etc.). More importantly, Labov demonstrated that speakers alternate between the variants in a statistically regular way. He explains this regularity by introducing the principle of *Variable Rules*, which states that grammatical rules are not static (i.e. categorical) and that every description of a rule should also include the predicted relative frequencies (or probabilities) of its application. For instance, Labov demonstrates that the deletion of a word-final stop consonant (*t, d*) applies more often if the stop is followed by a consonant (*fast car*) than a vowel (*fast automobile*); also, it is less likely if the coronal is a separate inflectional morpheme (*miss-ed*) than if it is not (*mist*). Those degrees of likelihood are calculated from actual usage events. However, in Labov's view, observed frequencies of those events are never recorded in the heads of the speaker. This is because frequencies, proportions and any kind of statistical data are viewed as random variables (in the sense that they cannot be predicted with 100% accuracy) and it was common belief that human language processing capabilities could not possibly include a random component, hence non-deterministic variables were always reduced to performance. On the other hand, probabilities are fixed numbers and not subject to variation, hence a part of competence (even though they are not categorical). As mentioned in the previous section, this work will argue for the opposite view – that the frequencies are in fact mapped into speakers' mental grammars.

Some more recent theories and authors within the generative framework, drawing on the concept of Variable Rules, do, in fact, include both frequency and variation in their descriptions (see, for instance, Yang (2004), who uses token frequency in parameter setting; Featherston (2007), who argues that generative syntacticians need to pay more attention to data; articles in Kepser & Reis (2005); Optimality Theory below etc.). In the view of such approaches, free variation arises from the possibility that some features were left underspecified for particular lexical items or in terms of the (non)-execution of optional rules. Adger (2006) tries to predict

frequencies of *was/were* variants in a Scottish dialect within the Minimalist Program. The author's model makes predictions by combining only three features (±singular, ±participant, ±author) and no historical or functional factors. Adger's claim is that "the grammatical system produces variants as a result of the featural specifications of lexical items and their mode of combination in the syntax" (2006, p. 511).

Optimality Theory (OT) is a generative theory designed specifically for phonology. The starting point of OT is a constraint, which is defined as a "phonological state which is universally preferred (or in the case of negative constraints, universally avoided)" (Guy, 1997, p. 335). The constraints themselves are universal in OT, but their mutual ranking is language-specific. However, unlike in traditional generative phonology, where a rule that is violable is not regarded as a rule at all, constraints in OT are allowed to be violated. The choice of the final output starts with a list of possible candidates, which is derived from the available input. Each candidate is then evaluated in light of the ranked constraints. What determines the best output for a grammar is the least costly violation of the constraints (e.g. violating the smallest number of constraints or violating the lowest ranked one).[28] However, Standard Optimality Theory still assumes a binary notion of grammaticality – only one candidate is selected as optimal and all losing candidates are assumed to be ungrammatical (suboptimal). No predictions are made about the relative ungrammaticality of suboptimal candidates nor is there the possibility of having more than one winning candidate.

Antilla (1995) discusses variation in Finnish genitives within the framework of Optimality Theory. The OT grammar defined by Antilla accounts for free variation by either introducing partial rankings of constraints or eliminating the rankings altogether, which, in turn, simplifies the grammar and makes it easier to learn.[29] The quantitative predictions of this grammar are later shown to closely resemble the proportions of the respective variants in a corpus of Finnish. A more recent version of

[28] See Kager (1999) for a detailed overview of Optimality Theory.

[29] Partial ranking means that in a set of constraints {A, B, C}, the ranking of B vs. C is left undefined, so sometimes B will be ranked above C, sometimes the other way round. However, in versions of OT that use such variable ranking, the number of possible rankings for N variables is N! (N-factorial), which might be too large a number for a learner to handle.

OT, Stochastic Optimality Theory (Boersma (1997), instead of a fixed ordering of constraints, defines a continuous constraint ranking scale. Each constraint has a predefined range on the scale and for each individual analysis a constraint is assigned a random real-number value within its range. The ranges of constraints which are closer to each other may overlap, resulting in variable ranking between them: sometimes the ranking will be A ≫ B and sometimes it will be B ≫ A. Guy (1997) is more in favour of this kind of an approach than the one that uses partial or variable rankings as it brings OT more in line with Variable Rules, which the author thinks is a superior model.

Kroch's (1994) main argument is that doubletism is necessarily a temporary state and signals instability in the system, which the system tries to change. This change, according to the author, is slow, but it proceeds in a certain well-ordered way, namely the rate at which the newer option replaces the older one is the same in all contexts (*Constant Rate Effect*). By analysing data from *The Oxford English Dictionary* on English past tense doublets, the author determines that the life span of doublets is 300 years, after which one of them either disappears or they develop a distinction in meaning, register, or grammatical properties.[30] The author also believes that during the period of time when they co-exist, morphological doublets are organized in the brain as *competing grammars*, i.e. each variant is described in its own terms and is part of a separate system.

2.2.2. Doublets in the cognitive tradition

Modern linguistics has abandoned the traditional generative teachings in this particular area and has started looking at variation, perhaps not as central, but as an important component of language. Every language varies and the reason for this variation, according to Berruto (2004, p. 297), can be considered broadly functional – "it represents an element of adaptation of language in human communities." Furthermore, variation has internal linguistic importance as "in the absence of

[30] Note that Bauer (2006) also asserts that the 17th century English system of derivations was chaotic, whereas now, 300 years later, it is orderly – both Kroch and Bauer refer to the same time span.

variation, languages would no longer be the versatile, flexible semiotic instruments they are, malleably following the flow of time and history."

Even though cognitive linguists take a more inclusive view of variation, we still find signs of an implied belief that no variation is completely free. Cruse (1986, p. 270) claims that "natural languages abhor absolute synonyms just as nature abhors vacuum." In this view, free variation is only apparently free and it results from a failure of the researcher to find the crucial variable that determines the choice between the alternatives; once that variable has been found, it will be shown that the variation in question is non-random, as anything else in language. However, not all authors would agree with Cruse. Erker & Guy (2012, p. 533), for instance, consider the two Spanish utterances *Yo hablo* and *Hablo* ('I speak') completely equivalent as they have identical truth conditions and logical form. This work will pursue the claim that the examples of doubletism we encounter in Croatian can indeed be considered examples of absolute synonymy, on the level of morphology at least. Such morphological synonymy can be expressed in the formula [different inflectional morphemes ↔ same grammatical meaning].

Naturally, if no two synonyms are absolutely identical, as claimed by Cruse, then there must be something that makes them different. The biggest innovation cognitive linguists introduce into the study of synonymy, and variation in general, is that it is no longer viewed as dichotomous in nature, but rather as a continuous feature. Hence differences in both meaning and environment (as the third element of the form-meaning relationship, as per Baayen, et al. (2013)) are gradient. Cruse (2000, pp. 158-160) defines (near)-synonyms as words 1) whose semantic similarities are more salient than their differences, 2) which do not primarily contrast with each other, and 3) whose permissible differences must in general be either minor, backgrounded, or both.[31] These differences are traditionally found on several levels: conceptual (differences in designation), stylistic (differences in connotation, register, or situation), structural (differences in argument structure), pragmatic (differences in emotion, implication) etc. On top of this, Divjak (2006) delineates Russian INTEND verbs by looking at the constructions these verbs open up. Divjak & Gries (2006) have

[31] Near-synonyms (or plesionyms) are, according to Divjak (2006, p. 21), neither in free variation nor in complementary distribution.

shown that there is often more than one type of these factors in play at the same time, and that it is therefore worthwhile to observe all categories together and in unison rather than separately one by one. Taken together, all these variables form the *behavioral profile* of a lexeme.[32] The two above studies by Divjak and Divjak & Gries are also noteworthy in that they are looking at the whole network of Russian near-synonymous verbs, which is methodologically a more complex task than simply looking at a synonym pair (also see Arppe (2009) on Finnish THINK verbs). They conclude that the best way to represent the relationship between the near-synonyms is by means of a radial network, where smaller distance between the nodes is associated with more synonymity.[33] Similar to this, Edmonds & Hirst's (2002) model groups near-synonyms into clusters by taking account of differences in granularity at a sub-conceptual level.

The majority of research on variation in the cognitive framework was concentrated in syntax, more specifically English syntax, which abounds in alternations such as the ditransitive construction (Gries (2003), Stefanowitsch (2006), Bresnan (2007)), *'s*-genitive / *of*-genitive (Gries (2002), Rosenbach (2003)), phrasal verb particle placement (Gries (1999), Anderwald & Szmrecsanyi (2009)) etc. (see contributions in Guerrero Medina (2011) for further examples of alternations in English). Syntactic variation was, to a smaller extent, researched in other languages as well (e.g. Heylen (2005) on NP ordering in German; Bader & Häussler (2010) on the dative alternation in German; Grondelaers, Speelman, & Geeraerts (2008) on Dutch existential-*there* sentences; Divjak (2008), (2016) on Polish [V + that + present] / [V + infinitive] constructions; Sokolova, Lyashevskaya, & Janda (2012) on the locative alternation in Russian; Diesing, Filipović Đurđević, & Zec (2009) on clitic placement in Serbian, etc.). Traditionally, all of these alternative constructions were considered to have the same associated meaning; the only difference was assumed to be

[32] For instance, Divjak & Gries' (2006) behavioral profile of Russian TRY verbs includes as many as 87 variable levels.

[33] In Divjak's (2006) analysis using clustering techniques, for example, *dumat'* 'think to' and *xotet'* 'want to' were shown to be very close to each other in synonymity, whereas *namerevat'sja* 'intend to' was the most dissimilar to all members of the network.

pragmatic, such as a difference in focus or information flow.[34] However, all of the works cited above show that each of the respective alternatives is a construction in its own right with its own, clearly delineated, scope of meaning.[35] Furthermore, most works discover *alternation biases*, i.e. lexical preferences of individual verbs (or other parts of speech) for one of the two alternative constructions. They do so by means of advanced statistical methods applied to large datasets retrieved from a corpus (e.g. Fisher's exact test, logistic regression, cluster analysis, Random Forest etc.).[36] The advantage of these models is that they can calculate the predictive strength of a great number of variables (lexical, semantic, syntactic, register, sociolinguistic etc.) and their combinations. When explaining grammatical choice, it is difficult to focus on only one type of pattern as often there is more than one factor determining the use of a particular feature. For instance, Arppe (2009) tries out several models with differing combinations of variables – his largest model contains as many as 64 features. Naturally, this model has the greatest *Recall* rate (i.e. accuracy of prediction) compared to other models which use less information. However, even such a large model does not achieve full predictability. Arppe (2009, p. 12) admits that some of the inaccuracy in his model could be a reflection of genuine synonymy. As pointed out by Kapatsinski (2014), it is highly likely that there is a certain degree of variation that is simply not possible to model because it is not determined by context but is rather a case of *probability matching* (i.e. replicating the proportions from the input; cf. Longo's (1964) cockroach experiment mentioned in Section 1.2).

[34] In Langacker's (1987, p. 39) view, these alternative constructions present "the same scene through different images."

[35] To take just one example from above, Rosenbach (2003) shows that both the '*s*-genitive and *of*-genitive in English can serve the same functions (determiner or modifier), but that only the *of*-genitive can have an additional function, that of describing a property (e.g. *king of honour / *honour's king*). Furthermore, the author identifies the animacy of the possessors as one of the crucial variables – proper nouns are more likely to occur with the *'s*-genitive, as well as certain inanimates, such as geographical or temporal nouns (*today's*) and collectives (*committee's*). The analytic variant, on the other hand, will be used more in situations that involve a greater processing load, e.g. with abstract possessors.

[36] Note that not all linguists unanimously support the use of statistical testing in linguistics. Kilgariff (2005, p. 273) quotes Lord Rutherford, who said "if your experiment needs statistics, you ought to have done a better experiment." Kilgariff's main objection is the fact that a large enough dataset will always establish that the null hypothesis is not true.

A further feature of these statistical models is that they produce probabilistic output (i.e. odds ratios), even though their input is categorical. Bresnan, et al. (2007) use the probabilities from such a model to predict native speakers' behaviour, showing in the process that speakers also make probabilistic, rather than categorical, predictions. Theijssen, et al. (2013, pp. 229-30), however, point out several problems with regression models, which make the authors doubt their interpretability – first of all, the features in the model should all be independent of each other, but more often than not they are correlated. "Correlated features cause problems with the interpretation of the roles that the individual features play in the model. For example, correlations can cause coefficients to flip sign or loose statistical significance" (Theijssen, ten Bosch, Boves, Cranen, & van Halteren, 2013, p. 229). The authors' second objection is that the regression involves so many transformations of data (e.g. residualisation) that it becomes impossible to relate them to the original data by the end of the analysis. However, in spite of these pitfalls, regression models of various types (linear, logistic etc.) continue to be widely used in linguistic research.

Boyd and Goldberg (2011) examine another type of syntactic phenomenon in English, A-adjectives (adjectives beginning with a syllabic schwa). Some of those adjectives cannot be used attributively (e.g. *adult man* vs. **asleep man*); however, this restriction is not semantically or phonologically determined, but is rather idiosyncratic. The authors are not directly interested in identifying the linguistic criteria speakers might use to determine which adjectives go into which category; rather, their interest is in how individuals learn to avoid using adjectives such as *asleep*, *alive* and *afraid* in an attributive position. They argue that any situation of this kind relies on the principles of statistical pre-emption. This means that the more the speakers hear one construction used in a certain context (e.g. *asleep* used predicatively), the more they block (or pre-empt) the occurrence of its alternative (*asleep* used attributively) in that specific context. The concept of pre-emption is not new in linguistic theory – in fact, it is roughly equivalent to the Blocking Principle described in the previous section. The authors' innovation is that they consider this pre-emption statistical – only by collecting a sufficient number of tokens can one form pre-empt another. Hence this approach would rely on speakers recording facts about the actual use of linguistic expressions alongside traditional structural information.

Proper doubletism, then, would constitute a situation where pre-emption has failed to take place.

Divjak (2008) legitimately asks the question why some forms are maintained with very low frequency levels in language (compared to their more frequent alternatives) when there is no obvious motivation for their use. We can simply attribute such deviations to performance errors arising from memory lapse, short attention span and the like, or we can say that even such low frequent forms serve some kind of a 'higher' purpose. For instance, Kapatsinski (2014, p. 31) believes that the selection of the less likely pattern of behavior for production "can, perhaps, be justified on the grounds of the need for practice to maintain the pattern in one's repertoire (…) It might also be explained by the greater salience of rare events compared to common ones [since] the occurrences of the rarer linguistic pattern might be more surprising and therefore more noticeable."

R. Ellis (1999) presents an interesting view of free variation in *interlanguage* – the language of second language learners. The author tries to explain the very existence of variation (i.e. why learners register a form when another one is already present in their lexicon) by resorting to the concept of playfulness: "communicative force is not the driving force here, but rather the *expressive need*, i.e. a personal need to perform a particular function using a variety of formal means. Language learners, as human beings, value variety for its own sake; they instinctively use language forms as objects which can be *experimented* and *played with*" (Ellis (1999, p. 470), italics mine). Just like we buy a variety of sweaters (all of which perform the same function of keeping us warm) and choose to wear a different one every day to avoid sameness, so we acquire two or more linguistic forms with the same purpose in mind. This desire to experience novelty has been shown by psychologists to be one of the basic human instincts (termed *investigatory reflex* by I. P. Pavlov). Various maze trials involving animals (see references in Ellis (1999)) have shown that animals chose a different route through the maze every time one was made available and that choice seemed to be quite random.

In recent years, several authors made use of the concept of *persistence* to explain free variation, primarily in syntax (see Gries (2005), Szmrecsanyi (2005a)). The basic idea behind persistence is that speakers, in situations where they have a

choice of structures (such as the dative alternation), tend to re-use structures they (or their interlocutors) have already used in preceding discourse. In other words, speakers are disinclined to switch between two structures in continuous discourse – they prefer to abide with a given structure once they have used it. The motivation behind this phenomenon can be found in the concept of recency, as explained by N. Ellis (2012, p. 7): "the more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed." In Szmrecsanyi's (2005a, p. 139) view, "models omitting persistence would leave a substantial share of the observable variation unaccounted for, or even erroneously identify it as 'free' variation although it is clearly patterned." However, even if this effect is legitimate, situations such as these, when we use two identical structures in close proximity to one another, make up only a small proportion of our whole language production; we are still no closer to finding an explanation as to why we use one or the other construction in all other situations, e.g. the very first occurrence of a structure or a word form in discourse.

Lyn Haber's little known work (1975), (1976) analysed the few examples of morphological doubletism in English, e.g. the past tense of *leap > leapt / leaped, knit > knit / knitted* etc. Even though these works pre-date the period of Cognitive Linguistics by at least ten years,[37] there are numerous similarities between her approach and the arguments later used by cognitive linguists. Haber admits that free variation does very much exist and then proceeds to analyse it in light of her *Muzzy Theory*. Haber extensively refers to Labov's work on Variable Rules, which she believes is only suitable for *describing* variation; the Muzzy Theory, on the other hand, accounts for the existence or *emergence* of variation. In the author's opinion, the reason why doublets arise in the first place can be traced back to the process of language acquisition. "The speaker has a choice because, when he was a child, he also had a choice" (Haber, 1975, p. 241). During the acquisition process, children go through a sequence of generalizations and rules for particular forms, which get modified every time a child receives new input. "Rather than assuming that these early rules disappear, Muzzy Theory asserts that they will remain in the behavioural

[37] 1987 is often taken as the year when Cognitive Linguistics was born as two major works in the field were published that year – Ronald Langacker's *Foundations of Cognitive Grammar* and George Lakoff's *Women, Fire and Dangerous Things*.

repertoire of the speaker and will intrude in his adult language wherever his language permits choice" (Haber, 1975, p. 246).[38]

However, Muzzy Theory only tries to account for the reason behind the emergence of competing forms; Haber makes no attempt in trying to predict their actual usage – in fact, the author thinks it is impossible. Her experiments on English past tense doubletism showed no correlation between either the age or gender of speakers (or any other linguistic variables, for that matter) with the percentage of verbs rendered in either the strong or weak form (which was in direct opposition to Quirk (1970), who showed a correlation between the choice of the form and the aspect of the sentence). More importantly, Haber's experiments have shown that speakers are quite inconsistent when they have different strategies for getting to the end result at their disposition (such as the regular and irregular strategies). The same subject could not be relied on to use the same strategy in similar situations (e.g. 47% of her respondents used one strategy for *fit* but a different one for its minimal pair *knit,* 17% used them both in the irregular, and 37% used them both in the regular form), nor could groups of subjects who consistently used the same strategies be distinguished. The author observed a similar lack of systematicity in a small-scale experiment eliciting the plural of the word *mouse* from school children.

Based on this data, Haber (1976) comes to the conclusion that speakers are very much aware of distributional patterns of language and that each alternant is stored in memory as a separate lexical item alongside its own variation figure (e.g. I should use *knit* as the past tense of *to knit* 30% or *sped* as the past tense of *to speed* 80% of the time). This figure, according to Muzzy Theory, is derived from the *sequence* of the child's acquisition of the relevant rules and the *duration* of time over which those rules were applied. For instance, some children might learn the regular rule first and apply it to all verbs, and then learn the irregular exceptions one by one. Other children might learn the regular rule first and then abandon it completely for an irregular rule,

[38] This is in line with views expressed later by various usage-based scholars, who argue that our mental grammars are quite "messy" (Dąbrowska, 2004, p. 148), containing both low-level and higher-level schemas as well as exemplars from which those schemas are extracted. For instance, Bybee (2010, p. 15) argues that "once the generalization is made, the speaker does not necessarily have to throw away the examples upon which the generalization is based." Similar to this, N. Ellis (2009, p. 154) says that "the language calculator has no 'Clear' button."

which they apply across the board; only later would they learn to separate the two and re-apply the regular rule. Yet other children might learn the irregular rule first and so forth. Haber's claims were quite insightful and advanced for her time and it is surprising they have not been referred to more. This work will attempt to give additional credibility to these claims, using larger and more reliable corpus data and native speaker testing.

It took several decades for the interest in morpohological doubletism to be re-ignited. Fehringer (2004) analyses two cases of morphological doublets in Germanic languages: Dutch adjectives in -lijk/-elijk and German genitives in -s/-es. The author starts off by cross-checking the lists of doublets found in various dictionaries of the respective languages with actual usage, a method which we already labelled as methodologically most appropriate. She traces the historical development of the two forms and tries to determine whether their number is increasing or decreasing. Most linguists tend to agree with Kroch (1994) that morphological doubletism should only be a temporary state. For instance, Dąbrowska (2004, p. 227) believes that "systems in which there are several competing patterns are inherently unstable. (…) Once a particular variant gains a clear advantage over the others, whether by chance or as a result of a language fad, it will tend to spread and may eventually monopolise the grammar." Interestingly enough, Fehringer's data show competing patterns of behaviour: Dutch data show a tendency for doublets to be eliminated over centuries, so there are only a handful left in modern Dutch; however, German is showing no signs of this tendency as genitive doubletism is quite abundant in the modern language. The author, like other scholars before her, tries to find factors (morphological, prosodic, register-related etc.) that would explain a preference for one form over another. In cases where she fails to find one, the author concludes that forms are truly in random variation and that "this is not perceived as problematic for the speaker in any way" (2004, p. 317).[39] However, what the author also notes is that words showing a strong preference for one variant over the other tend to be those that are used very frequently. A similar pattern emerged in our studies as well.

[39] This view was inherited from Bybee's *Network Model* (2001), in which inflected words are stored as lexical wholes. In this model, the lexicon is not represented as merely a list of words but as a network of connections to related items, which makes storage more efficient. This way doublets are not treated as something problematic and uneconomical.

Thornton (2011) analyses overabundance in Italian (in her terminology, doublets are called cell-mates) in light of Corbett's (2009) canonical approach. As a reminder, Corbett defines a canonical paradigm as one in which all cells display the same stem, but each cell exhibits a different inflectional ending. In such a view, overabundance is obviously a deviation from canonical behaviour; however, Thornton believes that within overabundance itself there are canonical and non-canonical cases as well. Since in the author's model *no condition* > *condition* (where '>' means 'more canonical than'), cell-mates can be considered canonical only if they are completely interchangeable and occur with the same frequency. She identifies the Italian doublet pair *sepolto / seppellito* (past participle of the verb *seppellire* 'to bury') as canonical in this sense. Furthermore, even though previous studies (Chumakina, Hippisley, & Corbett, 2004) have shown that the maintenance of overabundant forms was tied to the existence of specific constructions that required a specific cell-mate, in the Italian examples this factor does not seem to be at work: both cell-mates appear in identical constructions. When it comes to the role of frequency in the maintenance and reduction of overabundance, Thornton's (2012) results show a similar pattern to Fehringer (2004) – overabundance tends to be better preserved in low-frequency items. For instance, *possedere* 'possess' was the least frequent verb in Thornton's dataset, but also the one showing the most variation. However, her results do not support the conclusion that highly frequent words show a strong preference for one of the forms. In her analysis, the verb *dovere* 'must' has the highest frequency considering both the cumulative frequencies of the cell-mates in each of the overabundant cells and the global frequency of the lexeme, but it does not have the highest ratio between the two cell-mates in its overabundant cells. *Sedere* 'sit' has a much lower frequency than *dovere*, but a higher ratio between the two cell-mates in each of its overabundant cells.

Like Fehringer above, Thornton believes that overabundance can maintain itself for centuries. In fact, the author believes that the only reason why overabundance was reduced in some Italian verbs was conscious normative action by the reformer of the Italian language, Alessandro Manzoni. Verbs that escaped Manzoni's interventions have not reduced overabundance over the years, nor do they show signs of doing so.

Cappellaro (2013) adds a diachronic and acquisitional dimension to Thornton's theory. The author believes that overabundance emerges only in later stages of acquisition as by that point the Principle of Contrast stops operating. Items that are prone to overabundance are not part of the core lexicon/grammar (which is acquired first) and are of low frequency.

Morth and Dressler (2014) distinguish between two types of doublets. Doublets of the first type exhibit no discernible difference in meaning within the same strictly defined speech community and these are the only proper instances of overabundance. The other type, apparent doublets, includes, among others, forms which are not part of the same language variant or lect (e.g. Standard German *Park-s* vs. Swiss German *Pärk-e*), diachronic doublets (forms which co-existed only at a certain stage of development of the language, such as *Uhu-s* vs. *Uhu-e* 'eagle owls'), onomastic doublets (*Palm-en* 'palm trees' vs. *Palm-es* 'members of the Palme family'), or plurals of figures (*David-s* 'people named David' vs. *David-e* 'statues of David'). By looking at the pragmatic level, the authors are able to identify an additional number of doublets as belonging to the second type. For instance, there are cases where one of the forms can have connotations of strangeness or pejoration. In addition, the plural ending of a loanword will depend on the way the word is pronounced, e.g. if *der Balkon* 'balcony' is pronounced the French way, it will tend to have -s, whereas if it is pronounced the German way, it takes the productive ending of masculine nouns, -e. All in all, the authors find evidence for a diachronic trend towards eliminating overabundance by paralleling distinctiveness in form with a distinctiveness in meaning.

Finally, Guzmán Naranjo (2016) approaches Spanish imperfect subjunctive doublets from a Construction Grammar perspective. In the author's approach, "both constructions, -se and -ra, instantiate the same grammatical core construction, retain the pragmatic value associated with it (which the author labels $PRAG_l$) but specify additional pragmatic information associated exclusively with the specific form in question (labelled $PRAG_j$ and $PRAG_k$)." In a subsequent statistical analysis of corpus data using the Naïve Discriminative Learner model (more on which in Chapter 5), the author reveals lexical preferences associated with both morphemes (e.g. all modal verbs and verbs ending in *-er* prefer the variant -ra, while -se does not show any clear preferences; furthermore, -ra is more likely to appear in sentences without overt

subjects than -se). At the same time, none of the traditional linguistic variables (e.g. PERSON, NUMBER etc.) helped distinguish between forms.

2.2.3. Doublets in Croatian philology

We have seen that the dominant opinion about doublets in traditional generative linguistics is that they are a feature of performance unnecessary for the understanding of competence. Since Croatian linguistics is, in its core, still predominantly traditionalist and prescriptivist, a great number of published grammars of Croatian and many acknowledged experts in the field start from a similar presumption that doublets are an undesirable feature of language.

We bring a selection of quotes from various normative manuals and scientific articles in which doubletism is seen as a negative feature of Croatian which should be eliminated by all means. For instance, August Kovačec, one of the participants in a roundtable discussion on the state of affairs of Croatian language norms organised by *Matica hrvatska* in 1998 (published as Samardžija (1999)) said that "many people argue that the more words a language has, the richer it is. That is simply not true! A language is rich if every form has its specific function; I would not call having fifty words at disposal to express the same thing richness, I would call it *prodigality*" (Samardžija (1999, p. 333), translation and italics mine).[40] On a similar note, the authors of *Hrvatski jezični savjetnik ('Croatian Language Advisor')* argue the following:

> The standard language cannot *bear* several signifiers for a single signified, several synonymous and equivalent linguistic units (…) The standard language needs variants to fulfil the needs of all its functional styles, but it does not need all variants in equal measure. Variants in orthography and grammar *threaten it*, whereas lexical (except in

[40] Croatian original: "Mnogi polaze od toga, što je više riječi, to je jezik bogatiji. To jednostavno nije istina! Jezik je bogatiji ako svaka riječ ima točno određenu svoju funkciju, a to da li ću ja imati na raspolaganju pedeset riječi za istu stvar, mislim da to nije bogatstvo nego *rasipnost*."

terminology) and stylistic ones are desirable (HJS (1999, p. 48) translation and italics mine).[41]

These quotes sound as if they were copied straight from the writings of the great Italian writer and language reformer Alessandro Manzoni. Some of Manzoni's views on language appear in Thornton (2012, p. 199): "having different ways of meaning many different things, that's the richness of languages; having more ways of meaning one and the same thing, that's not richness, but *overload*, it's not freedom, but *hindrance*; and such a hindrance that usage naturally and continually tends to get rid of it" (Thornton's translation, italics mine).

Babić (1962, p. 62) invokes the concept of *language economy* as an argument against allowing doublets:

> If two forms have the same meaning, and if they can be used equally in every context, then one of them is *deadweight* in the language. Imagine what would happen if we had two forms for every concept. We would have to learn everything twice, we would put twice the load on our brain, yet we would still speak only one language. It is clear that this would be a major defect of any language, so one of the synonyms should disappear from the language. If that does not happen, their meanings start to differ (translation and italics mine).[42]

The concept of language economy first appeared in the works of Prague School functionalists (where it was known as the 'one form – one function' principle) and American structuralists. More recently, Kiparsky (2005) proposed an economy constraint, which requires a meaning to be expressed by as few forms as possible. In such a conceptualization of economy, having no variation is more economical than

[41] Croatian original: "Standardni jezik ne '*trpi*' više označnika za isti označenik, više istoznačnih i istovrijednih jezičnih jedinica. (…) Standardnomu jeziku inačice su potrebne da bi zadovoljio potrebe svih svojih funkcionalnih stilova, ali mu nisu potrebne sve inačice podjednako**.** Pravopisne ga i gramatičke inačice *ugrožavaju*, a poželjne su leksičke (osim u nazivlju) i stilističke."

[42] Croatian original: "Ako dva oblika imaju isto značenje i ako se mogu podjednako upotrijebiti u svakom kontekstu, jedan je od njih *balast* u jeziku. Zamislite što bi se dogodilo kad bismo za svaki pojam imali dvije riječi. Morali bismo sve dvostruko učiti, dvostruko bismo opteretili svoj mozak, a znali bismo samo jedan jezik. Jasno je kako bi to bio velik nedostatak svakog jezika i da zbog toga jedna od istoznačnica s vremenom nestane iz jezika. Ako se to ne dogodi, počinju se razlikovati po značenju."

having variation in abundance as this minimizes the number of forms a speaker deals with. If we regarded economy simply in terms of number of forms, polysemy would be an example of economy *par excellence* (as pointed out by Klavan (2012, p. 18)) and the most economical language the one with only one word. However, common sense tells us that such a language would also be highly ambiguous and not very efficient.

However, as Hopper & Closs Traugott (1993, p. 72) point out, "there are great difficulties in defining the notion of economy in anything like rigorous terms. We know very little about what does and does not take 'effort' in producing or interpreting utterances, and still less about what would constitute economy of mental effort on either speaker's or hearer's behalf." Kapović (2012, pp. 29-30) wonders whether it is more economical for a language not to decline nouns, like in English, or to decline them, like Slavonic languages do. We can see a certain economy, or 'advantage', in both cases. In the former case, a language is definitely morphologically simpler, as we use a word in the same form regardless of its syntactic function. However, at the same time we are constrained by a fixed word order and obliged to use a greater number of words (e.g. prepositions) to signal grammatical relations. In the latter case, we are freer in the choice of word order. Similar to this, Croatian, like Latin, has a three-way distinction of demonstrative pronouns (*ovaj* – close to the speaker, *taj* – close to the the interlocutor, *onaj* – away from both), whereas English only has a two-way distinction (*this* vs. *that*). Which one of the two is a more economical conceptualisation?

As pointed out by Langacker (1991, p. 262f.), arguments based on economy generally carry limited weight in usage-based models. The reason is the model's emphasis on psychological reality. If empirical investigation should suggest that speakers prefer complex grammars with massive redundancy, the best linguistic description is one that reflects this state of affairs accurately – even if it is not very economical.

In usage-based linguistics, economy is not related to the number of forms, but rather to their complexity. The economy principle as described by Haspelmath (2008) defines how languages code their material. In this view, more frequent forms are coded so that they require less articulatory effort, i.e. with zero coding (as opposed to less

frequent forms that will have overt coding), or are shorter, or are coded in a more straightforward way (e.g. synthetic rather than analytic), and so forth. This reflects a well-known maxim by John W. Du Bois that "grammars code best what speakers do most" (quoted in Haspelmath (2008, p. 185)). In such a view of economy, as pointed out by Rosenbach (2003, p. 400), "given two structural alternatives, the option demanding less mental effort will be preferred to that option requiring more mental effort." All in all, we dismiss the above arguments for the elimination of doublets based on the concept of economy as unjustified.

Other Croatian linguists justify the elimination of doublets by invoking arguments from information theory: "It is not redundant to say that doublets, unlike synonyms, make communication harder because they increase entropy: the more of them there are, the smaller the predictability of information" (Silić (1998, p. 175), translation mine).[43] We feel this view is a grave simplification of information theory and the concepts of economy and redundancy. There are more parameters that co-affect predictability within a system other than the pure number of forms. More importantly, items serve their purpose in context not in isolation. Even less probable forms can serve an important function by significantly narrowing the possibility space (or entropy) of what is about to come. Hence the claim that the very existence of doublets makes the system less predictable is unrealistic.

The dominant view in Croatian linguistics still seems to be that linguistic forms are either 'correct' or 'incorrect' and that it is up to the linguist to define which form is which. Orthographic manuals ("*pravopisi"*) regularly top the lists of bestselling books in Croatia and speakers are reprimanded for not learning their language properly. However, modern linguistics has abandoned this view a long time ago – "all words and forms used in a language variety are 'correct'. Some of them may be standard and the others non-standard, but they are all, linguistically speaking, correct" (Kapović, 2011, p. 46). Native speakers cannot speak incorrectly. As argued by Lehmann (2007, p. 233), "if a person who does not suffer from pathological conditions

[43] Croatian original: "Nije naodmet reći da dublete, za razliku od sinonima, otežavaju komunikaciju jer povećavaju entropiju: što ih ima više, predvidljivost je informacije manja."

is speaking his mother tongue, whatever he says has to be taken as linguistically correct by the linguist."[44]

Furthermore, as pointed out by Anić (2009, p. 563), speakers expect to find unambiguous solutions in normative manuals. When they come across doublets or contradictory solutions in different manuals, they blame the 'sad condition' the language is in on the incompetence of the linguist – "they talk so much, yet they are not even able to say what is correct."[45] Vince (1977, p. 143) approaches this problem from the perspective of the language learner: "one loses faith if s/he encounters different solutions to the same problem in relevant authoritative manuals. If that happens often, the learner will not only be confused but also impeded in acquiring the norm" (translation mine).[46] In such a situation, the speaker eventually develops a fear of speaking Croatian (publicly at least) due to the danger of making a mistake and being labelled as not knowing his/her own language.

Due to all the reasons stated above, the majority of Croatian grammarians try to introduce some kind of a distinction between the two variants so as to avoid the awkward construction 'it can be said both ways'. However, more often than not, this distinguishing factor is not reflected in actual use. As we will see later in the corpus analysis, the majority of doublets are used completely interchangeably, even by the same speaker, in the same contexts.

Croatian linguists reared in the ideas of Neogrammarians share drastically different views. As Jakobson (1932 [1974], p. 9) explains, "a Neogrammarian can state: form B replaced form A. If you ask a Neogrammarian which form is better, they will not be able to answer you. Form A is older, form B is younger, but this information does not enable one to give a value judgment" (translation mine).[47] Josef Zubatý, a

[44] Also see the discussion in Sampson (2007), as described in Section 2.1.

[45] Croatian original: "Što toliko pričaju kad ne kažu kako je pravilno."

[46] Croatian original: "Dobro nam je poznato da se gubi pravo povjerenje ako se o istim pitanjima dobivaju iz mjerodavnih priručnika raznolika autoritativna rješenja. Ako se to ponovi više puta, učenik može biti ne samo zbunjen nego i ometan u prihvaćanju jezične norme."

[47] Serbo-Croatian original: "[Mladogramatičar] konstatuje: oblik B nadomjestio je oblik A. Ako mu postavite pitanje koji je oblik bolji, on nema šta da odgovori. Oblik A je stariji, oblik B je mlađi, ali iz toga ne proističe ništa za vrednovanje oblika."

Czech linguist, argues that "we should be happy that we are able to say one thing in one way on a certain occasion and in another way on a different occasion" (quoted in Jakobson (1932 [1974], p. 7), translation mine).[48] Some (Serbo-)Croatian linguists have adopted similar views, but we are under the impression that they are a minority. For instance, Anić (2009, p. 469) believes that in situations where the system is open for several options, the codification of a single form is "unnecessarily rigid" (Croatian: 'nepotrebno rigidna'). Bosnian linguist Milan Šipka believes the following:

> In a single language it is not possible, nor necessary, nor desirable to unify all varieties. It is least desirable to abolish synonyms and reduce them to a single word, a single expression. Because that would cause irreparable damage to the language itself and the multiple functions it has as an instrument of culture and civilization, and especially as a means of artistic expression" (Šipka (1987, p. 95), translation mine).[49]

Finally, Ljudevit Jonke (1965a, p. 199) argues that "the introduction of doublets, even triplets, means both a theoretical and a practical step forwards. It enables the utilization of the complete lexical treasure of our language. (…) The forced elimination [of these doublets] would cause irreparable damage to the language norm" (translation mine).[50]

We adopt the view of the latter group of linguists and we proceed with an analysis of Croatian doublets with a completely non-restrictive mindset. We now move on to the methodology of such an analysis.

[48] Serbo-Croatian original: "Opet moram ponoviti da (…) treba da se radujemo što smo u stanju da istu stvar izrazimo jedanput ovako, a drugi put drugačije."

[49] Bosnian original: "Moramo reći da u jednom jeziku sve raznolikosti nije ni moguće, ni potrebno, a ni poželjno ujednačavati. Ponajmanje je poželjno 'ukidanje' sinonima i njihovo svođenje samo na jednu riječ, na jedan izraz. Jer, time bi se nanijela golema šteta samom jeziku i njegovim mnogostrukim funkcijama koje ima kao instrument kulture i civilizacije, a pogotovo kao sredstvo umjetničkog izražavanja."

[50] Croatian original: "Uvođenje dvostrukosti, pa i trostrukosti, znači i teoretski i praktički znatan korak naprijed. Ono omogućuje iskorištavanje cjelokupnog rječničkog blaga našeg jezika. (…) Njihovo nasilno uklanjanje prouzrokovalo bi teško podnošljiv poremećaj norme književnog jezika."

Chapter 3. Types of linguistic data - diverging or converging evidence?

In this chapter we analyse the two most common types of data acquired by linguists, which we also use in our analyses in Chapters 4 through 6 – intuition data and corpus data. Wallace Chafe (1992, p. 84) observed that data can be 'artificial' or 'natural' and collected through processes that are either 'behavioural' or 'introspective'. In such a view of linguistic data, corpus data are considered natural, whereas any kind of experimental data (including intuition) are necessarily artificial because they are elicited in artificial settings.

3.1. Intuition judgments

Due to a lack of (or the inability to acquire) other types of data, linguists traditionally relied on introspective data in the form of intuition judgments. Examining intuition has in general been considered a feature of so-called *armchair linguistics*. However, it also largely features in generativist writing. For instance, in a recent survey of data points from articles that appeared in the generative-oriented journal *Linguistic Inquiry* from 2001 through 2010, Sprouse, Schütze, & Almeida (2013) estimated that 77% were derived from some kind of a judgment task.[51] Furthermore, the majority of those judgments were collected rather informally, which has brought a great deal of criticism on this method, especially concerning its lack of statistical analysability and replicability. Schütze & Sprouse (2014, p. 30) identify five features of such informally collected judgments: "(i) relatively few speakers (fewer than ten), (ii) linguists themselves as the participants, (iii) relatively impoverished response options (such as just 'acceptable,' 'unacceptable,' and perhaps 'marginal'), (iv) relatively few tokens of the structures of interest, and (v) relatively unsystematic data analysis." Furthermore, Wasow & Arnold (2004, p. 1483) object to generative grammarians providing judgments as if they held for all speakers, without checking to see how variable they are. However, Sprouse & Almeida (submitted, p. 34) suggest

[51] A similar analysis, with comparable results, has been performed by Sampson (2005), spanning more than 40-years' worth of articles from the journal *Language*.

that these concerns have been overstated – "not only do traditional methods produce low false positive rates, but they also seem to be well protected against false negatives."[52]

Geoffrey Pullum offers a somewhat sarcastic, yet not totally inaccurate, view of how introspective judgments have traditionally been collected. As we will see later, a similar method is widely used in Croatian linguistics as well:

> If you want some sequence of words to be grammatical (because it would back your hypothesis), the temptation is to just cite it as good, and probably you won't be challenged. If you are challenged, just say it's good for you, but other dialects may differ. If it doesn't sound so good, decorate the context a bit to enhance its plausibility and cite it as good anyway. Or if you need the same word sequence to be ungrammatical, fiddle with the context or the meter or some irrelevant lexical choices to make it sound a bit worse, and put an asterisk in front of it (Pullum, 2007, p. 36).

However, the method of examining intuition is being introduced more and more into cognitive linguistics as well, with notable differences. As argued by Wasow & Arnold (2004, p. 1483), "consulting the primary intuitions of native speakers is a type of psychological experiment. Hence, such data collection should be subject to the usual methodological expectations of experimental psychology." This means the intuitions need to be collected from a large representative sample of linguistically naïve speakers in a carefully designed experiment, the results of which are subject to appropriate statistical analysis. Gilquin & Gries (2009, p. 9) mention three obvious advantages of such experimental approaches: "a) they allow the study of phenomena that are too infrequent in corpora, b) they make it possible to systematically control for confounding variables, and c) depending on the nature of the experiment, they permit the studying of online processes."

[52] The term *false positive* (or Type I error) refers to cases when a statistical test indicates a relationship between two variables when in fact one does not exist; conversely, *false negative* (or Type II error) refers to rejecting a relationship when in fact there is one. Both errors are usually caused by too small a sample or wrong methodological design.

3.1.1. Issues with intuition judgments

As Trask (1999, p. 88) points out, we all have intuitions about our own language, "about what is normal, acceptable, unusual, strange or impossible, or about what a given form means and when we might use it, if at all. The issue is how much trust (if indeed any at all) we should place in speakers' intuitions in compiling our descriptions of language." Leonard Bloomfield specifically rejected reference to the speaker's internal states as beyond the reach of science (qtd. in Cowart (1997, p. 1)). Similarly, Stefanowitsch (2006, p. 73) is of the opinion that judging the acceptability of utterances is not a natural human activity (as opposed to interpreting utterances). Schütze (2005) objects to all kinds of artificial unfamiliar tasks (such as acceptability judgments) as they can always open up the possibility of a different kind of variation than the one we are interested in. For instance, even though speakers might not use a particular form themselves, they still might recognise it as a form used by others, hence deem it acceptable (a phenomenon Aarts (1991, p. 33) calls *currency*).

On the other hand, there are cognitive linguists who see value in such judgments. Antti Arppe (in Arppe, et al. (2010, p. 18)) believes judging acceptability is as natural as language production, albeit of a different quality. Wasow & Arnold (2004, p. 1484) argue that making judgments of 'well-formedness' is also a type of language use, "albeit an unusual one." Bader & Haussler (2010, pp. 321-2) also think rating a sentence as (un)grammatical is a natural task, especially for students. Rating one sentence as better or worse than another one, on the other hand, is a type of task not many people have encountered before. However, the authors believe that even such a task can be mastered quickly and without much effort. Finally, Featherston (2005b, p. 205) sees introspective judgments "the data type of choice for syntax" because they can provide judgment on all types of structures, not only on the 'best' ones – those that appear in the corpus.

Antilla (1995, p. 25) believes that 10,000 intuitions are even less reliable than intuitions of a single person and that they are especially unstable if the variants are close to each other in optimality (as is the case with morphological doublets). On the other hand, Culbertson & Gross (2009, p. 723) believe that "even if individual linguists were more reliable than individual non-linguists, the judgments of large numbers of

non-linguists might be more reliable than any individual (including individual linguists)."

Pullum (2007, p. 40) starts off his discussion of intuition judgments by saying that asking people questions can provide the researcher with 'meaningless junk' – "and collating large quantities of meaningless survey junk is not a path to truth." However, the author nevertheless concludes that they can be an important source of evidence, if properly designed and collected under optimal conditions. Schütze (1996) also only accepts the results of such studies if a number of possible confounding factors have been controlled for – both task-related (e.g. plausibility of content, sentence length and complexity, order of presentation, method of elicitation etc.) and subject-related (dialect variation, social class, response speed, even handedness of respondents[53]). However, Featherston (2009, p. 130) believes that many of those do not need to be controlled for, as long as extreme values are avoided.[54] Finally, Sprouse (2009, p. 330) demonstrates that acceptability judgments are more robust than some critics have suggested.

Many authors have pointed out that it is unclear what intuition judgments are actually measuring – competence (whether a sentence is permitted by the grammar) or performance (whether a sentence is likely to be used). Devitt (2010, p. 834) for instance, calls intuition judgments "voice of competence." Aarts (1991, p. 46), drawing on a similar belief by Chomsky, claims that 'grammatical sentences' are products of competence, whereas 'acceptable sentences' are products of use (performance), and the latter does not necessarily include the former. Schütze & Sprouse (2014, p. 28) argue that "since a grammar is a mental construct not accessible to conscious awareness, speakers cannot have any impressions about the status of a sentence with respect to that grammar." Similar to this, Bader & Häussler (2010, p. 277) believe that the term *grammaticality judgment* is "a misnomer because

[53] There is empirical evidence (e.g. Cassanto (2009)) showing that whether a person is left- or right-handed influences the way in which they evaluate concepts such as 'good' and 'bad'; hence a different conceptualisation of language could also be present in left- and right-handed people.

[54] Similar to this, Myers (2009) believes that all the protocols that need to be followed in the design of formal psycholinguistic experiments (such as randomization, counterbalancing, use of control and filler items etc.) are a waste of time and effort. Instead, the author promotes the use of so-called 'small-scale judgment experiments', which are as simple and quick as informal methods, but still rely on the careful design and quantitative analysis of formal experimentation.

grammaticality cannot be judged directly. Judging a sentence always involves more than just the properties that the mental competence grammar ascribes to the sentence. What can be judged in a literal sense is the *acceptability* of sentences; grammaticality is only one of the factors which determine acceptability." Hence if a sentence is judged unacceptable, that does not necessarily mean it is ungrammatical. The authors propose *perceived well-formedness* as a superordinate term. Wasow & Arnold (2004, p. 1428) distinguish between primary (whether an item is well-formed, meaningful etc. or not) and secondary intuitions (why an item is the way it is), arguing that the latter do not constitute empirical evidence.

Cowart (1997, p. 9) believes that "there is no such thing as an absolutely acceptable or unacceptable sentence, although there might be such a thing as an absolutely grammatical or ungrammatical sentence." As we have seen in the discussion in Section 2.1, usage-based theories do away with the competence/performance distinction, so most theorists use the terms grammaticality and acceptability interchangeably. In the instructions of their rating task, Arppe & Järvikivi (2007a) asked their participants to rate the 'naturalness' of the sentence as they believe, in line with Penke & Rosenbach (2004, p. 492), that "related theoretical notions such as *grammaticality*, *acceptability*, *well-formedness*, *correctness* and *interpretability* are most probably difficult to distinguish for lay informants."

Relying solely on intuition data and disregarding usage data can lead to dubious results. We know from experience that in actual speech speakers rarely utter sentences exhibiting full argument structure (i.e. containing a subject, a verb and an object), even though this is what grammars generate. Many previous works have shown that items that have been labelled as 'ungrammatical' by linguists not only occur in actual language, but are also accepted by speakers, and vice versa. For instance, Dąbrowska's (2008a) analysis of questions with long-distance dependencies (henceforward LDDs) demonstrated that the sentences that usually appear in generativist writing to exemplify the rule about LDDs never actually appear in the language. In short, the generativists' claim was that there could be any number of intervening clauses between the WH-word and the main clause without any effect on their comprehensibility or acceptability. An example Chomsky gives in support of this claim is *Who did Mary hope that Tom would tell Bill that he should visit?* Dąbrowska, on the other hand, shows that real-life sentences differ to a great extent from such

constructed ones and that more than 70% of attested sentences conform to one of the following patterns: *WH do you think S-GAP?* or *WH did NP say S-GAP?* (2008a, p. 392). So, unlike Chomsky's invented example, the majority of real-life LDD questions only have one or two intervening clauses, refer to the second person and contain the verb *think* or *say*. This opens the possibility that the informal methods that have characterized data collection in syntactic theory in the past have led to unsound theories. Similar to this, both Geoffrey Pullum and Geoffrey Sampson (separately) challenged Chomsky's purely intuitive claims on verb fronting in questions where a subordinate clause precedes the main clause: "the generative linguists shared one intuition – the relevant questions are all vanishingly rare, while Pullum and I shared a different intuition – all these questions are perfectly normal; and empirical evidence showed that we were *all* quite wrong" (Sampson (2005, p. 20), italics original). In short, they were shown to be frequent in writing but rare in speech. Similar analyses showing the unreliability of Chomsky's other intuitions were undertaken by Wasow & Arnold (2004) and Gibson & Fedorenko (2013).

Sampson (2007) finds no reason to assume that patterns in a speaker's intuitive grammaticality judgments reflect realities of their language. Similar to this, Divjak (2016, p. 21) believes that judgments may reflect properties of the rater rather than properties of the grammar. The author came to this conclusion after one of the variables with the strongest effect in her analysis turned out to be *rater generosity*. In short, participants who gave filler sentences high scores were also more likely to give trigger sentences high scores too. Bermel, Knittl, & Russell (2015) classify their respondents into several types, depending on which parts of the 7-point rating scale they used. The authors distinguish between respondents who used the full scale (largest group), those who used the endpoints and several (but not all) midpoints, permissive (did not use the lowest marks), hesitant (did not use highest and lowest marks), and categorical respondents (used only the endpoints and one midpoint). These groups are later used to analyse the respondents' behaviour on a forced-choice task, but they found no evidence to disprove the null hypothesis. In addition, the authors calculate a whole variety of other measures of respondents' behaviour (e.g. STRENGTH OF PREFERENCE, OVERALL PERMISSIVENESS etc.) and examine correlations between them.

Schütze (1996, p. 48) offers several explanations for two subjects giving conflicting reports about the same stimulus: "subject A is truly experiencing a different sensation than B; subject A was experiencing the same sensation as B, but was misreporting it; subject A is simply lying." All three, in the author's opinion, show the unreliability of introspective judgments as measure of productivity.

Furthermore, a variety of authors have shown an effect of repetition on acceptability of judgments. Surprisingly, however, different authors discovered opposing effects. Whereas Luka & Barsalou (2005) find that exposure to structures in a reading task increase the acceptability of those structures in a subsequent rating task (termed *habituation*), Nagata's (1988) findings show that repeated exposure to sentences makes the judgments more stringent (i.e. the more the subjects look at a sentence the more they find wrong with it). The finding that grammaticality judgments were easily influenced by repetition and other variables, such as embedded context, led both sets of authors to conclude that linguistic intuitions as revealed in grammaticality judgments are not absolute but relative. Snyder (2000) shows that *satiation effects* (all sentences starting to look alike) appear only in certain types of sentences, thus dismissing this effect as a property of the judgment process and therefore also as a liability for the linguist to use acceptability as an explanatory tool. Various procedures can be used to ward off such unwanted effects of order (such as counterbalancing the test material). All in all, as argued by Cowart (1997, p. 5), the utility of introspective judgments in furthering research far outweighs any limitations ascribed to them.

In sum, we are not arguing for the complete abandonment of intuition in linguistic research – as pointed out by Gilquin & Gries (2009, p. 3), "abandoning judgment data altogether would mean throwing out the baby with the bathwater." Rather, we argue that introspective judgments are the most direct way of probing into speakers' implicit knowledge of their native language. Various authors (e.g. Newell & Bright (2001), N. Ellis (2002)) have pointed out that language learning (or at least the memorising part of it) is an implicit rather than an explicit process. The features of such an implicit knowledge are that it is "intuitive and procedural, variable but systematic, usually accessed by means of automatic processing and during fluent performance, and not verbalizable" (Gutierrez, 2013, p. 424).

We end this section with a lengthy quote from the great Bosnian linguist Midhat Riđanović (2005, p. 61) who, in a very illustrative way, shows the benefits of intuition, which he calls our "most precious possession" (Bosnian original: 'naš najintimniji posjed'):

> Any respected (psycho-)linguist nowadays believes that language is the product of human instinct, similar to, for instance, digesting food. And we cannot control our instincts. Try eating a sandwich and 'ordering' your food processing system not to digest it. You clearly do not stand a chance. The grammar of our native language is an instinctive part of us as well. Try speaking ungrammatically (…) and you will see you will 'break a sweat', yet you will never be able to fluently speak ungrammatically. Just like you cannot stop or change the process of digesting food, you cannot affect your language feeling either.[55]

3.2. Corpora

With the rapid development of computer technology, linguists gained access to a particularly helpful source of data in the form of large digital collections of texts, or corpora.[56] Pullum (2007, p. 37) calls corpora "the most useful tool that has been provided to the grammarian since the invention of writing." Similar to this, Fillmore (1992, p. 35) says that "every corpus I've had a chance to examine, however small, has taught me facts I couldn't imagine finding out about in any other way."

Gilguin & Gries (2009) develop a hierarchy of linguistic data in terms of naturalness of collection. In their hierarchy, corpora of written language come out at

[55] Bosnian original: "Danas svi ozbiljni (psiho)lingvisti smatraju da je jezik plod ljudskog instinkta, otprilike kao, recimo, varenje hrane. A instinktima ne možemo upravljati. Pokušajte pojesti sendvič i 'narediti' svom probavnom traktu da ga ne vari. Očito je da nemate šansi. I gramatika maternjeg jezika je instinktivna u nama. Pokušajte govoriti negramatično (recimo da kažete **Ja doći kafe popilo* umjesto *Ja došla da popijem kafu*) i vidjećete da ćete se 'preznojiti' a ipak nećete nikad tečno govoriti negramatično. Isto onako kao što ne možete zaustaviti ili izmijeniti proces varenja hrane, ne možete uticati ni na svoje jezičko osjećanje."

[56] It is important to note that the modern-day meaning of the word *corpus* differs from traditional uses of the word. Namely, in times of Leonard Bloomfield, a corpus was a collection of recordings that were collected in not-so-natural surroundings, such as by means of interviews or observation.

the top as the most natural, whereas intuition judgments occupy the opposite endpoint. Since a corpus is made up of texts that had already been created for some purpose other than linguistic research, there is no way for the researcher to bias the contents of the texts (however, see below for matters of biasing the contents of the corpora). Another major advantage of a corpus is its size – for instance, the Corpus of Contemporary American English (COCA) contains 450 million words, whereas in recent years it has been made possible to use the entire World Wide Web as a corpus (the WebCorp project), whose size is several orders of magnitude larger. However, as already mentioned in Section 2.2.2, Kilgariff (2005, p. 266) argues that size is not necessarily a good thing. "If we increase the sample size, we would ultimately reach the point where all null hypotheses would be rejected."

Performance data retrieved from a corpus is often interpreted as a "window to the mind" (Gilquin, 2010, p. 89), a way to study something unobservable (i.e. the human mind) by means of something observable. In other words, "although corpus data do not reflect the characteristics of mental grammars directly, we do consider corpus data a legitimate source of data about mental grammars (…) Characteristics observable in usage reflect characteristics of the mental processes and structures yielding usage, even though we do not know the exact form of these mental representations" (Divjak & Arppe, 2013, pp. 229-30).[57]

In recent years, the standard in cognitive linguistic research has been to analyse corpus data using a variety of statistical methods. Many proponents of this kind of analysis argue that this way of doing research brings corpus linguistics closer to psycholinguistics. From this perspective, as pointed out by Stefanowitsch (2011, p. 272), the linguistic corpus becomes "a model of linguistic usage (…) of an 'average' speaker and the (quantitative) methods applied to it become a (partial) model of the way in which this average speaker derives linguistic knowledge from usage."

On the other hand, as pointed out by Martin Hilpert (in Arppe, et al. (2010, p. 15)), even many psycholinguists remain unconvinced that corpus data allow us to draw conclusions about cognition. Biber, Conrad, & Cortes (2004, p. 376) say that frequency data should in no way be regarded as explanatory. "In fact, we would argue

[57] However, note that even some corpus linguists (e.g. Teubert (2005)) believe corpus linguistics should not be concerned with the psychological aspects of language, but only social ones.

for the opposite: frequency data identifies patterns that must be explained. The usefulness of frequency data is that it identifies patterns of use that otherwise go unnoticed by researchers." However, cognitive linguists by and large maintain that frequency distributions nevertheless have some kind of significance. Schmid (2010, p. 118), for instance, believes that "what frequency counts in a corpora reflect more or less directly are degrees of conventionalization of linguistic units or structures." We turn to the issue whether corpora are able to, in fact, represent the average speaker/listener next.

3.2.1. Issues with corpora

In the generative view, corpora only supplement the real source, which are linguists' intuitions and should not be used to draw any kind of conclusions about the grammaticality of any given structure. The generativists' rejection of corpus data is best illustrated in the debate, referred to in Sampson (2005, p. 32), between a respected generative linguist R. B. Lees and Nelson Francis, one of the authors of the first-ever corpus of English – the Brown Corpus. Lees was reported to have said to Francis that compiling the Corpus was "a complete waste of time because as a native speaker in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text."

The main critique directed at corpus linguists from the generative camp has focused on the fact whether data from corpora can in fact faithfully represent the language of the whole population? Generative linguists deny corpora any explanatory power at all. In Chomsky's view, any natural corpus will be skewed. "Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description [of language based on the corpus] would be no more than a mere list" (Chomsky, 1957, p. 159). Furthermore, Newmeyer (2003, p. 696) points out that "there is no way that one can draw conclusions about the grammar of an individual from usage facts about communities, particularly communities from which the individual receives no speech input." Newmeyer objects to methods which make inferences about child language based on corpora which consist of, e.g. texts from the *New York Times* (considering no child reads the *NY Times*).

Since it would be impossible to include every single utterance ever produced in a language in a corpus, we need to choose a representative sample to stand as a proxy for the complete language production. Unfortunately, even one of the pioneers of corpus design, Douglas Biber, admits that a corpus that is truly representative of the language we encounter in everyday life would have to consist of 80% conversation, 10% television shows, 2% radio broadcasts, 2% letters and other writings, 1% newspapers, 1% novels, and 4% other texts (Biber & Jones, 2009, p. 1288). However, since present-day corpora are more concerned with representing the full range of variation that is present within a language rather than the actual proportions of variation, they are more than appropriate for research purposes.

Another one of Chomsky's objections to corpora is that they cannot provide negative evidence. In other words, if a linguistic item does not appear in a corpus, we cannot possibly know whether it is, to use terms by McEnery & Wilson (2001, p. 11), "unseen but grammatical" (absent because of the corpus is not large or diverse enough) or "ungrammatical and unseen" (absent because it is not possible in the language). In such a situation, in Chomsky's opinion, only intuition judgments can provide reliable data on the (un)grammaticality of an item. However, Stefanowitsch (2006), (2008) has shown how it is possible to distinguish between what he calls *accidentally* and *significantly absent* items by calculating expected frequencies from observed frequencies. More importantly, Stefanowitsch also shows that the speakers are able to make the same calculations in their minds.[58] Our surveys have also shown that items that are accidentally absent will never be perceived as absolutely unacceptable for native speakers.

3.3. Methodological pluralism – comparing intuitions and corpus data

The two types of data analysed above are different in so many ways – we have already explained the differences in naturalness, interpretation etc. Dylan Glynn (in Arppe, et al. (2010, p. 7)) calls corpus data *found* data, whereas experimental data examining intuition constitutes *elicited* data. Featherston (2005b, p. 205) illustrates

[58] However, Stefanowitsch also points out that neither intuition judgments nor his method can tell us *why* an item is significantly absent – only linguistic theory can do that.

the difference between the two in the following way: "if you want to know *what* people say, choose frequencies, but if you want to know *why*, you are better off with judgments" (italics original).

However, linguistics has for a long time been criticised and impeded in its development because of its so-called methodological monism – we have already seen that intuition judgments were the method of choice in generative syntax. Corpus linguistics is in danger of repeating the mistakes of its predecessors if corpus analysis remains its sole interest. Rather, as Wasow and Arnold (2004, p. 1484) point out, "it is typical of all scientific work to test hypothesis against multiple types of evidence using multiple methods." Conclusions drawn from a single data source can leave an incomplete picture, whereas the replicability of results using different methods only increases their robustness. As pointed out by Arppe and Järvikkivi (2007b, p. 108), since language is a multimodal phenomenon, we can expect to understand it fully and comprehensively only by combining multiple methods and multiple sources of evidence.

Hence it is common practice nowadays for patterns from corpora to be cross-checked against the behaviour of real language users, on tasks as diverse as rating forms, making a choice among several forms, imitating a stimulus, naming a picture or simply reading at your own pace (or combinations of these), during which not only one's responses but also reaction times, eye movements and other performance data can be recorded. Although the two types of data each have their own advantages and disadvantages, when brought into conjunction, more often than not, they give converging results and solve problems that the researcher would face if only one method had been employed. As pointed out in the discussion in Arppe, et al. (2010, p. 21), "it is rare that a corpus study entirely nullifies the results of a good intuition-based study." However, as noted in the same discussion (2010, p. 11), even if the two measures do not consistently yield the same result, this does not automatically undermine the validity of one of them. The goal of such a multi-methodological analysis is, in Pullum's (2007, p. 40) opinion, "an optimal fit between a general linguistic theory (which is never complete), the proposed rules or constraints (which are quite as conformant with the general theory as we would like), the best grammaticality judgments obtainable (which are not guaranteed to be veridical), and facts from corpora (which may always contain errors)."

Fillmore (1992) argues for a methodology of linguistic analysis where corpora are used as a means of maintaining authenticity of a form while augmenting this data with native speaker intuitions as a way of filling out paradigms and exploring possible analyses. A multitude of works have used this 'converging evidence' approach in order to see how frequency affects mental representation (Gries (2002), Kempen & Harbusch (2005), (2008), Hoffmann (2006), Divjak (2008), Schönefeld (2011), Bermel & Knittl (2012), Caines (2012), etc.).[59] We will proceed with analysing a few of them to see whether their results offer converging or diverging evidence on the relationship between the frequency of an item and its status in the speakers' mental grammars, which can be expressed with numerous terms – *acceptability, salience, prototypicality* etc.

Bader & Häussler (2010), for example, "found no instance where a syntactic structure S1 was judged as more acceptable or grammatical than a syntactic structure S2 but S2 occurred with greater corpus frequency than S1." A similar pattern has emerged in all the studies quoted above and could be said to represent a basic effect of frequency on the status of a linguistic unit in mental grammars.

Arppe & Järvikivi (2007a) conducted a large-scale study of Finnish verbal synonymy in which they compared data obtained from three different sources: corpora, forced-choice experiments and acceptability rating experiments. Besides confirming conclusions derived from a pure corpus analysis, the two questionnaire studies also allowed the authors to draw some reliable conclusions about two rarer cases in Finnish which could not have been deduced from the corpus data itself. We will present their conclusions in more detail as they have shown that the relationship between frequency and acceptability does exist but is not so simple to describe. They put forward seven formulas which illustrate this relationship:

i) frequent → acceptable
ii) unacceptable → infrequent
iii) (acceptable → frequent) ∨ (acceptable → infrequent)
iv) ¬ (infrequent ↔ unacceptable)
v) ¬ (acceptable ↔ frequent)
vi) frequent ↔ preferred

[59] Gilquin & Gries (2009) searched through two large bibliographical databases and found 85 such articles to date.

vii) infrequent ↔ dispreferred (Arppe & Järvikivi, 2007a, p. 151).

In other words, a highly frequent item will also be highly acceptable (i), but at the same time this does not mean that an infrequent item will be unacceptable (iv). Rather, even infrequent items show moderate levels of acceptability. Conversely, acceptable items come from both high and low frequency ranges (iii). The last two statements, derived from their forced-choice task, say that frequency correlates with preference as does infrequency with dispreference. As we can see, the relationship between the two concepts is very complex and it needs to be explored further.

Bresnan (2007) does not use raw frequency data to predict speakers' behaviour, but rather probabilities from a statistical model (which was fed the frequency data). The author's hypothesis was that the subjects' ratings of alternative dative constructions in English would correspond to the model probabilities, meaning that a sentence in which the model predicted a prepositional dative with a high probability would also be rated by speakers more favourably with a prepositional dative. This hypothesis was later confirmed. Similar analyses have later been performed by Arppe & Abdulrahim (2013) and Divjak, Dąbrowska, & Arppe (2016). This kind of reasoning will also be used in one of the questionnaire studies described in Chapter 6, where the behaviour of native speakers will be compared to the performance of two exemplar-based models of language.

After discovering contradictory patterns when using different measures of frequency (token vs. log frequency) on their own in predicting the use of subject pronouns in Spanish, Erker & Guy (2012, p. 546) make an argument that frequency does not make its unique contribution to the probability that a form will occur or not, but rather that its effect comes from the interaction with other predictors (Arppe (2009) also suggests a multivariate analysis of frequency). More specifically, the authors believe that frequency affects linguistic variation indirectly, "as a limitation on the sufficiency of evidence." Erker & Guy define the *frequency threshold* as the level at which a speaker has enough evidence to formulate reliable estimates of distinctive lexical properties. "Below some frequency threshold, items are too rare to formulate rich representations that include collocational information. Above the threshold, language users have sufficient information about each lexical item to individuate them" (2012, p. 549).

However, defining that threshold has proven itself to be a difficult task for authors. In Erker & Guy's original set-up, the distinction between frequent and infrequent forms was arbitrarily set at 1% of the corpus, although their subsequent analysis has shown this boundary to lie somewhere around the values of log 1.5-1.7. Gordon & Alegre (1999) discover no frequency effects for regular inflections with a frequency below six words per million. Bybee (2010, p. 18), on the other hand, claims that "there is no way for frequency to matter unless even the first occurrence of an item is noted in memory. Otherwise, how would frequency accumulate?" Similar to this, Luka and Barsalou (2005) also provide evidence that *mere exposure* to moderately ungrammatical structures has a positive effect on ratings and that the effect shows already after one or a small number of exposures to such structures. However, as a point of caution, Featherston (2005b) finds that different thresholds seem to operate in different syntactic domains. This point of minimum exposure will be revisited in one of the subsequent questionnaire studies described in Chapter 6.

Kempen & Harbusch (2005) found a systematic discrepancy between the frequency counts and grammaticality ratings. First of all, the authors observed that "the grammaticality judgments tend to be more lenient than the corpus data" (2005, p. 337). Contrary to their expectation that constructions in the middle range of the grammaticality spectrum would appear with moderate frequencies in the corpus, they were conspicuously absent. In their view, grammaticality ratings are sensitive to the number and seriousness of violations of the production-based rule. Hence constructions that represent mild violations of the rule will receive medium-range grammaticality scores, even though the grammatical production mechanism cannot produce them at all. The explanation for this discrepancy is that "somewhere on the grammaticality continuum ranging from 'perfectly well-formed' to 'seriously ill-formed' there is a critical value called the *'production threshold'*. Structures with grammaticality values above this threshold will occur in corpora with moderate-to-high frequencies. Structures whose grammaticality scores lie slightly above or slightly below the threshold will have zero or, at best, very low frequencies – they are 'marked'" (2005, p. 343).[60]

[60] However, once again, the definition of this threshold eludes the authors. They only say that a sufficient number of exposures is needed for a variant to attain the status of a more or less stable

Gries (2002) wonders what results a linguist should take as more relevant when the two methods just described lead to diverging conclusions. In other words, should the linguist in cases like these describe the performance-based or judgment-based grammar?

> If results from corpus data contradict results from acceptability judgements and both could be explained equally well but differently, I would always tend to accept the hypothesis supported by the corpus data: the production of linguistic utterances/texts that happen to end up in a corpus occurred under completely natural circumstances and is, thus, less likely to be subject to experimental bias than questionnaire data. Moreover, I would in general consider corpus data to be more precise in the sense that factors such as register, prescriptive attitudes and medium can be filtered out, whereas we can never be sure to what extent they influence subjects' reactions in experimental settings (Gries, Evidence in linguistics: Three approaches to genitives in English, 2002, p. 28).

As opposed to this, Rosenbach (2013, p. 293) does not think that we can assign *a priori* any privileged status to any type of evidence in the case of diverging results. Just like experiments can be conducted in a careful or in a sloppy way, corpora can be analysed in a sound or in a superficial way. Schütze (1996), on the other hand, argues that grammaticality judgments offer a much more direct insight into the internal grammar of speakers as performance data are necessarily prone to various kinds of performance errors, caused by processing load, memory lapses etc. On top of this, Bauer (1993, p. 8) also believes that 'editorial interference' introduces bias into the corpus data. Namely, writers are always encouraged by their editors to use the prescribed forms and their original wording is sometimes corrected, hence these forms will predominate in the corpus. In elicitation studies there is no such direct interference (note, however, the discussion on the *observer's paradox* in Section 6.1).

Boersma (2004) argues that mismatches between acceptability judgments and corpus frequencies can be explained by a difference between the linguistic task of

grammar item, without attaching any kind of numerical value to the label 'sufficient' (Kempen & Harbusch, 2008, p. 189).

production and the task of providing acceptability judgments. In production, the speaker chooses a pronunciation that best harmonizes speaker- and listener-based requirements. On the other hand, acceptability judgments involve choosing, for a given meaning, an overt form that comes closest to realizing it. Mismatches between acceptability judgments and production occur because of the role speaker-based preferences play in the production process.

Featherston (2005b, p. 196) believes that the two data types do *not* measure the same factor (italics original) – "we can therefore exclude categorically the possibility that relative judgments merely reflect frequency or probability of occurrence." Of the sixteen different constructions the author tested, two-thirds of them did not appear in the corpus at all; yet his subjects gave different ratings to all of them.

Finally, Divjak (2016, p. 19) believes that the reason why usage frequency has problems predicting acceptability judgments, especially at the low end of the frequency spectrum, is because language researchers have been looking at the wrong kind of frequency data, focusing on raw or co-occurrence frequencies rather than on higher-dimensional conditional probabilities.

In sum, we feel that a pluralistic approach, in which the results of several methods of data collection are juxtaposed to each other, should be the preferred way of doing linguistic analysis. Whereas statistically analysing data from corpora could reveal interesting patterns, it is only by testing real speakers that we can be certain that these patterns have a psychological basis. Such an approach is adopted in the present dissertation.

Chapter 4. Morphological doublets in Croatian - a corpus approach

Like any Slavonic language, Croatian has a rich morphological system. Nouns and adjectives are inflected for seven cases, three genders, and two numbers;[61] adjectives, on top of this, also distinguish between definite and indefinite forms, while personal pronouns also have stressed and unstressed forms. All Croatian verbs have one present tense, four past tenses and two future tenses, two conditionals, two verbal adjectives (participles) and two verbal adverbs (gerunds). With this multitude of forms and endings, it is no wonder that we can find one ending appearing in different cells (syncretism) as well as cells which are populated by more than one ending (allomorphy and doubletism). This chapter gives an overview of the inflectional system of Croatian, focusing especially on cells where doublets arise. For each of the cells, we present data from both grammar books (and other reference manuals) and corpora of Croatian.

The majority of Croatian grammars in circulation are to a large extent prescriptively oriented. However, as argued by Kačić (2001, p. 48), "in order to get to a good prescription of language, one first needs an accurate description of it. And a good description needs to contain: a) a comprehensive list (corpus) of written texts from various genres and b) spoken language of a variety of social groups" (translation mine).[62] The descriptions of Croatian are inadequate in this respect, to say the least. The majority of Croatian grammar books and other reference manuals are still written in the old tradition, drawing their data either from a) older, 'good' writers or b) the authority of the grammarian and their 'language feeling'. Rarely are their examples

[61] Unlike its close relative Slovene, Croatian has no category of dual number, although some vestiges of it can be seen in case endings of some nouns. What Croatian does have, like many other Slavic languages, is syntactic agreement with numerals '2', '3', and '4', which is sometimes labelled as 'paucal'. Corbett (2010, p. 16) objects to the use of this label as paucal is a number form parallel to singular, plural and dual and is used for a small number, which is not the case in Croatian (the form *čovjeka* cannot be used independently to refer to a small number of men), hence he prefers the label '234 form'.

[62] Croatian original: "Da bi došli do dobrog propisa moramo imati dobar opis. A dobar opis mora sadržavati: a) što obuhvatniji popis (korpus) pisanih tekstova svih vrsta, b) govorni jezik svih razina i svih sociooloških grupa."

drawn from present-day usage. For this reason, Anić (2009, p. 474) calls the Croatian language norm "impressionistic".

When it comes to the question of using literature as a linguistic source, Brozović (1976, p. 113) argues that morphological forms that can be found in literary texts written in the period between the 16th and the first half of the 19th century should not be considered as relevant to the standard language. These texts belong to the pre-standard period of Croatian, hence their inclusion in any corpus of Standard Croatian is not justified. Ivić (1965) also argues that writers are language innovators by vocation, hence we should not rely solely on their language in the process of language standardization.[63] Marković (2012, p. 220) believes that "if we were to stop looking up to old role models, the morphological description of Croatian would be far more informative" (translation mine).[64] With regards to the latter source of evidence, the authors of *Hrvatski jezični savjetnik* [*Croatian Language Adviser*] strongly argue that 'the language feeling' cannot be used as a criterion for anything, let alone for a language standard. "We cannot use something that is by its nature inconsistent, changeable, subjective, individual, and personal to create something that is supposed to be supra-personal, general, obligatory and relatively stable, such as a standard language" (HJS (1999, p. 49), translation mine).[65] However, at another place in the same text, when trying to determine a reliable criterion for the distribution of the second palatalization (more on which in Section 4.3.2.1), the authors go on to say that "the state of affairs in written sources and the living language also cannot be relevant" (HJS (1999, p. 34), translation mine).[66] We wonder, then, what the authors would consider relevant data for a proper description of language.

[63] However, an opposing opinion is expressed by Jonke (1965b, p. 11), who calls writers "the guarantors of the organic development of our language" who pay special attention to their linguistic expression [Croatian original: "Književnici su garancija za organičan razvoj našeg književnog jezika"].

[64] Croatian original: "Možda bi bilo grubo, ali vjerojatno ne bi bilo daleko od istine kazati da bi s manje ugledanja u stare uzore morfološki opis suvremenoga hrvatskog bio znatno obavjesniji."

[65] Croatian original: "Ne može se na temelju nečega što je po svojoj naravi nestalno, promjenjivo, subjektivno, pojedinačno i osobno, kakav je jezični osjećaj, izgrađivati nešto što bi trebalo biti nadosobno, opće, općeobvezatno i razmjerno stabilno, kakav bi trebao biti standardni jezik."

[66] Croatian original: "Stanje u pisanim izvorima i živom jeziku također ne može biti mjerodavno."

Exactly half a century ago, Jan Svartvik wrote that "corpus studies will help to promote descriptively more adequate grammars" (Svartvik, 1966, p. vii). Even though such views were consistently rejected by the ideology of the time, nowadays hardly anyone (apart from a few hard-core conservatives) would dare to argue against the importance of corpus studies as a source of information about the grammar of a language. Biber (2012, p. 11) believes that "when specific quantitative findings are not reported [in the grammar], the reader must simply take it on faith that the identified lexical-grammatical patterns are in fact 'common' or 'frequent'". Moreover, "most linguistic phenomena are not distributed in a simple binary opposition of 'frequent' versus 'rare'. Rather, there is normally a continuous range of variation, and quantitative findings are required to adequately describe those patterns." Hence one of the goals of this chapter is to give a synchronic empirical description of doubletism in Croatian using state-of-the-art corpora.

The reader might get the impression at one point or another that our purpose is simply to criticise the work of Croatian grammarians. While we have offered some criticisms, it is with a different goal in mind. In Pennington's (2013, p. 248) view, "it is the objective linguist's duty to question the validity of prescriptive grammars and when possible to turn to a comprehensive analysis that includes diachronic, sociolinguistic and corpus linguistic details rather than settling for *ad hoc* explanations that do more to confuse than to enlighten." This is the real purpose of the present chapter. We hope the data presented in this, as well as other chapters, will contribute to the ongoing debate over how grammars of Croatian (and other languages) should be written and how to use the tools at our disposal to describe the language more accurately and usefully.

The chapter is organised as follows. In Section 4.1 we give a short historical overview of the development of Standard Croatian with the intention of identifying key periods and processes that led to the emergence of doublets in the language. Section 4.2 gives information on the corpora we used in our research. Section 4.3 is the longest section, which presents the empirical data for all four nominal declension patterns. Section 4.4.summarises the findings and patterns emerging from the data.

4.1. Development of Standard Croatian[67]

The whole South-West Slavonic area has been the subject of debate as to its linguistic make-up. The best way to describe this geographical area would be to use the term *dialect continuum* as per Peti-Stantić (2008). Several dialects are spoken throughout this area and there is "a more or less unbroken possibility of communication from one dialect to another" (Peti-Stantić, 2008, p. 437). The Kajkavian dialect[68] is spoken in north-west Croatia and all of Slovenia, Čakavian is spoken along the coast of Dalmatia, Istria and the islands, Torlak is spoken in south-east Serbia. The fourth and largest dialect is called Štokavian and it is spoken on the rest of the territory. This dialect will be the focus of attention throughout this overview because it serves as the basis for both Standard Croatian and Standard Serbian. The Štokavian dialect itself is further divided into three sub-dialects, which differ in the reflexes of the Old Slavonic phoneme *jat* (ě): the *ekavian* or eastern (ě > e), *ikavian* or western (ě > i) and *ijekavian* or southern (ě > (i)je).

Literature in all three dialects spoken on Croatian territory has been thriving ever since the 16th century (especially in Štokavian-speaking Dubrovnik and Dalmatia). Soon after that, first grammars and dictionaries started to appear (*Institutionem linguae Illyricae*, written by Bartol Kašić in 1604, is often labelled as the first Croatian grammar). Since there was no unified Croatian state at that time (Croatian territory was split between the Turkish empire, Austro-Hungary, Venice and other rulers), there was no unified Croatian language either, hence those manuals actually described specific dialects (notice that even Kašić does not use the term 'Croatian language' but rather 'Illyric' to refer to the Štokavian dialect).

The 19th century turned out to be a crucial period for the development of both Croatian and Serbian language standards. From that period onwards, it is impossible

[67] The complexity of the linguistic situation in the South-West Slavonic area would call for another dissertation-sized piece of work, hence in this section we only attempt to trace back the origin of doubletism in Croatian, without going into issues not directly relevant for our arguments, e.g. whether Serbo-Croatian was a real language or matters of language policy. The major part of this section is a summary of the history of Croatian presented in Katičić (1974) and Moguš (1993). For additional intakes on the topic, see e.g. Peti-Stantić (2008), Greenberg (2008) etc.

[68] The names of the different dialects (apart from Torlak) are derived from the forms of the interrogative pronoun 'what' in the respective dialects, namely *kaj, ča, što*.

to trace the development of Croatian without taking into account the parallel development of Serbian. Hence a couple of words about Serbian need to be said as well.

The language situation in 17th-century Serbia was no less complex. Just like Croatia, Serbian territory was not unified, but rather split between several rulers (Austro-Hungary, Turkey and the independent Kingdom of Serbia). The language of medieval literature was Church Slavonic of the Serbian recension, written in Cyrillic script. All literary output was connected to the Orthodox church since monks were the only literate people at that time. In the majority of schools children were taught Church Slavonic of the Russian recension. The spoken language on the territory of Serbia up to the end of the 18th century is often referred to as '*slavjanoserbski'* – it was a mixture of Church Slavonic with some elements of Russian and local Serbian (Štokavian) speech. At the beginning of the 19th century there stepped to the fore a self-taught linguist named Vuk Stefanović Karadžić, who became one of the crucial figures in the whole south-west Slavonic territory. Karadžić's seminal work, *Srpski rječnik* (*'The Serbian Dictionary'*), published in 1818, also contained an overview of grammar in the appendix. This book marks the official beginning of Serbian as we know it today.[69] Karadžić's most radical step was to break away from the Church Slavonic tradition and the increasing Russian influence. Instead, he proclaimed that the 'new' language of Serbians should be based on the speech of the 'common Serbian folk'. More specifically, as the basis of the standard language Karadžić took the dialect of his home region, Eastern Herzegovina, which is a Štokavian ijekavian dialect. Having spent years performing the Sisyphean task of trying to persuade the authorities and fellow Serbians that this was, in fact, a step in the right direction, his standard was introduced to Serbian schools in the 1860s.[70]

[69] Peti-Stantić (2008) claims that the reason why first works describing the Serbian language did not appear before the 19th century, much later than for other South-West Slavonic languages, was because of the Orthodox mindset, which had a different relationship towards language. According to Peti-Stantić (2008, p. 433), "writing grammar and lexicographical works was a sin against the Orthodox doctrine."

[70] It would be interesting to compare Karadžić's reform to similar trends taking place in other languages across 19th-century Europe. For instance, Thornton summarises the main ideas of Alessandro Manzoni's reform of Italian. In many respects, Karadžić and Manzoni seem to have completely divergent views on standardization. Manzoni's goal, among other things was "elimination of forms that were typical of his local dialect (Lombard-Milanese) in favour of more common forms" as well as "elimination of all

This digression about Serbian was necessary in order to understand the next period in the development of Croatian. At the same time as Karadžić was reforming the Serbian language, a movement called 'the Croatian revival period' (or the *Illyric movement*), headed by Ljudevit Gaj, was underway in Croatia.[71] Croatian territories were beginning to unite under a common ruler at that time, so the need for a common language arose. In spite of German or Latin, the official languages of the Austro-Hungarian Empire, being spoken (or at least understood) by the majority of people, they were considered languages of the oppressor and neither of them was native to the area, so the natural choice for this common language was one of the three dialects of Croatian. Even though the Illyric movement thrived in Zagreb, the centre of political power, which is located in Kajkavian-speaking territory, the Revivalists decided to take the Štokavian dialect instead as the basis of the 'new', standard Croatian language. This was justified by the rich literary tradition of writers from Dubrovnik and Dalmatia, who wrote in Štokavian. All that was left to do was to codify that language by means of grammars and dictionaries. At that time, news of Karadžić's reform reached Zagreb and many Revivalists were inspired by his works. "His writings, with the way they presented data and his attitudes and the great international recognition they received, seemed to fit current needs more than the eighteenth-century manuscripts which basically determined the Croatian Štokavian standard. Many Croatian writers looked with awe at the old man who needed to fight for recognition even in his native Serbia" (Katičić (1974, p. 237), translation mine).[72] Since the Illyric movement was by its nature unitarian, with the aim of unifying all South Slavonic nations, the adoption of Karadžić's language served that purpose. Whereas in the Serbian case, the choice of Štokavian represented a radical switch from

kinds of doublets, both morphological and lexical" (Thornton, Reduction and maintenance of overabundance: A case study on Italian verb paradigms, 2012, pp. 198-9).

[71] It is important to note that the Illyric movement was primarily a political movement rather than a literary-linguistic one. However, Gaj and his followers often used linguistic arguments as a means of awakening the nation's consciousness (i.e. if our language disappears, so will we as a nation).

[72] Croatian original: "Njegovi su spisi načinom svojega izlaganja i svojim stavovima, pa čak i velikim međunarodnim priznanjem koje su dobili, više odgovarali potrebama vremena nego stariji spisi osamnaestoga stoljeća koji su zapravo utemeljili hrvatski štokavski standard. Mnogi hrvatski pisci gledali su zadivljeno u velikoga starinu koji se u svojoj Srbiji još morao boriti za opće i službeno priznanje."

the established tradition, its choice for Croatian seems to represent a more natural sequence of events.

In 1850 major Croatian and Serbian literary figures signed an agreement in Vienna, which laid the groundwork for a complete integration of the language and determined future language policies. From that time on we can talk about Standard Neo-Štokavian,[73] which, through the course of time, established itself in two variants: Standard Croatian (or 'western') and Standard Serbian (or 'eastern') (and later Standard Bosnian and Standard Montenegrin, as these languages also have the same dialect as their basis).[74] However, soon after the 'Croatian revival period' ended in the late 1860s, some mixed feelings about this 'new language' came to be expressed.

At that time, there were two major 'schools of philology' in existence in Croatia – the Zagreb School of Philology (*zagrebačka filološka škola*, represented by Adolfo Veber Tkalčević and Bogoslav Šulek) and the Croatian Vukovites (*hrvatski vukovci*, i.e. the followers of the work of Vuk Karadžić). Both of these schools produced their own grammars of Croatian, based on the corpus of older literary works. Some of the areas where their attitudes differed were the following: members of the Zagreb School used the language of medieval and 18th-century Croatian writers as their corpus, whereas the Vukovites referred only to the works of Karadžić and his contemporaries. Furthermore, members of the Zagreb School were in favour of the principle of tri-dialectalism of Standard Croatian (i.e. introducing the elements from all three dialects into the standard), whereas the Vukovites argued for a 'pure' Štokavian standard (so the speakers of Kajkavian and Čakavian would have to adapt and 'learn' Štokavian). Finally, the Vukovites promoted the phonetic principle of

[73] This is a term many Croatian linguists (e.g. Brozović (1976)) prefer to the awkward 'Serbo-Croatian' or 'Croato-Serbian', for which Katičić (2001) claims was never a functioning matter-of-fact reality but rather existed merely as a project. The term is derived by analogy with Standard Tuscan (for Italian) or Standard Castilian (for Spanish).

[74] Present-day Standard Serbian has, in the meantime, drifted away from the Vukovite tradition by adopting the urban dialect of the capital, Belgrade, as its basis. Standard Serbian nowadays is based on the Šumadija-Vojvodina ekavian dialect, whereas Standard Croatian has remained faithful to its ijekavian tradition.

orthography (popularised by Karadžić in his maxim "Write as you speak"), whereas the Zagreb School promoted the morphophonological (etymological) principle.[75]

The most influential work by Croatian Vukovites was *Gramatika i stilistika hrvatskoga ili srpskoga književnog jezika,* written by Tomo Maretić in 1899. This grammar was accepted as the 'official' one, even though it contained doublets, inconsistencies, and was based on a very limited corpus (only the writings of Karadžić). Since the native speech of Karadžić (Štokavian of eastern Herzegovina) is limited to a small geographical area, naturally some of the words and grammatical patterns used there would differ from those used in the rest of the Štokavian-speaking territory. However, since Maretić considered Vuk's language the purest form of Štokavian,[76] he introduced all of its grammatical features into the Croatian standard to co-exist with the already established forms. This caused a lot of criticism in Croatia. However, even the most prominent critics of the Vukovites' interventions could not help but admit that, despite all of its pitfalls, Maretić's grammar was the work of an expert philologist and a very well elaborated and detailed piece of work (Moguš, Povijest hrvatskoga književnog jezika, 1993, p. 175). Furthermore, people were tired of debates on orthography and they needed stability; this grammar brought that with it (Katičić, 1974, p. 239).

Croatian writers of the time, on the other hand, could not easily make peace with the fact that the rich Croatian literary tradition had been so easily cast aside and they continued writing as the older writers did. They claimed to be writing in the Croatian language as well, even though their language differed to some extent from the prescribed (i.e. Maretić's) norm (this primarily meant using elements from the other two Croatian dialects). Most of their works have been edited to fit the actual norm, but some of them have managed to survive in their original form to date.

Later 20th-century Croatian grammarians could not ignore Maretić's norm because it was too strongly established in language by then, but they also could not

[75] There was another school as well that was active even in the Revivalist period. That was the Zadar School of Philology, which did not approve of the ijekavian variant at all, but was of the opinion that the ikavian dialect should be used instead, due to the long tradition of ikavian writers in Dalmatia.

[76] In the Preface to his *Gramatika*, Maretić equates the role Karadžić played in the development of the language to the role Cicero played for the development of Latin.

ignore the strong literary tradition in Croatia, so in the majority of cases they made a compromise by adopting both. Hence in situations where these two traditions differ, we are likely to find doublets and competing solutions in different manuals. As we will see below, such situations are numerous in Croatian.

The aim of this section was to give a short overview of the development of contemporary Croatian and the tradition of grammar writing. As much as Croatian language purists today try to diminish the role of Vuk Karadžić, one cannot but conclude that his work is incorporated into the Croatian language as we know it today, a fact also pointed out by one of the greatest Croatian linguists of all times, Vladimir Anić (2009, p. 162).[77]

Before proceeding with a quantitative analysis of doubletism in Croatian, in the following section we present the corpora we used as the source of our quantitative data.

4.2. Croatian corpora

There are currently three large general-language corpora available for Croatian, alongside several smaller specialised ones (learner corpora, parallel corpora etc.). Below we give a short description of all three of them, without going into technical details on how they are annotated or how searches are conducted. Although useful as sources of information, it is important to note that all three corpora have their flaws in terms of user-friendliness and ease of retrieval of information, all of which had to be tackled differently during our research process.

In general, we feel that Croatian lags behind other Slavonic languages by a substantial margin in terms of the development level of the corpora, which makes linguistic analyses of the language harder. Czech and Russian have for a long time been the pioneers of corpus development in the Slavonic area. For instance, the Czech corpus was designed based on a survey of Czech people's reading habits, in order to ensure appropriate levels of representativeness and balance (see Králík and Šulc

[77] Croatian original: "Moramo zaključiti da je djelo Vuka Karadžića ne samo ugrađeno u naš jezik kakav on danas jest nego i da je mjera našega odnosa prema standardnom jeziku mjera naše veće ili manje dosljednosti prema Vuku i ukupnoj prijevukovskoj i poslijevukovskoj tradiciji."

(2005)). There have been no attempts of such a survey for Croatian, hence none of the corpora below fare particularly well on either of these matters. There are also numerous issues with annotation and retrieval of information, which we will not go into.[78]

4.2.1. Hrvatski nacionalni korpus (Croatian national corpus, HNK)

The Institute of Linguistics at the Faculty of Arts and Humanities in Zagreb has the longest tradition of producing corpora of Croatian. The first major attempt was *The one million token corpus* compiled by Milan Moguš over the course of twenty years (1976-1996). Based on that corpus, a frequency dictionary of Croatian was created (Moguš, Bratanić, & Tadić, 1999), the only Croatian frequency dictionary to date. This project was succeeded by the HNK project, headed by Marko Tadić. Several versions of this corpus have been produced throughout the years. Version 2.5 contains 101 million tokens, whereas the latest available version (3.0) contains 216 million tokens. The texts in version 2.5 of the corpus are represented as follows: 74% are informative texts (newspapers, magazines, journalistic writing), 23% are literary texts and 3% are texts of heterogeneous character. One important thing to note is that all the versions of HNK contain only post-1990 texts.[79] As Tadić (1998, p. 338) explains, "we all intuitively feel that from then on we could use Croatian more freely, more spontaneously, or put poetically, we could finally breathe it to the fullest" (translation mine).[80]

[78] For a more detailed description of problems a corpus linguist might come across when using Croatian corpora, see Stojanov & Vučić (2012), who compared the performance of HNK and HJR on several levels: document and content, presentation of results, and forms and characters.

[79] This is not to say that the texts from older authors are not present in the corpus as well; they are, but only in their post-1990 edited versions, i.e. editions adapted to the current standard.

[80] Croatian original: "Svi, dakako intuitivno, osjećamo da smo od tada hrvatski mogli rabiti 'slobodnije', 'spontanije' ili, gotovo pjesnički rečeno, mogli smo ga konačno 'disati punim plućima.'"

4.2.2. Hrvatska jezična riznica (Croatian Language Repository, HJR)

HJR was first developed by Damir Ćavar and Dunja Brozović Rončević under the auspices of the Institute for Croatian Language and Linguistics (IHJJ) in 2005. Since the primary goal of the project was, according to the authors, to serve lexicographical projects at the IHJJ, their primary concerns when creating the corpus were not matters of representativeness and balance. "Instead of one fixed balanced corpus, our goal was to create an annotated text corpus, as large as possible, that could be dynamically mapped on individual sub-corpora for specific research and development interests" (Ćavar & Brozović Rončević, 2012, p. 52). Although the authors' original goal was to expand the corpus indefinitely, due to lack of funding the project was stopped in 2011. However, the latest version of the corpus is still available online. This version contains more than 100 million tokens and the ratio of fiction to specialised texts (non-fiction prose, scientific literature, newspapers) is 28: 72.

4.2.3. Croatian web corpus (hrWaC14)

HrWaC is the latest addition to the list of Croatian corpora and is the largest available corpus of Croatian to date. It is part of the worldwide Web as Corpus project. Project leaders are Nikola Ljubešić and Željko Agić and it has been developed under the auspices of the Department for Information and Communication Sciences at the Faculty of Arts and Humanities in Zagreb. The current version, 2.0, which is the version we used, contains 1.9 billion tokens (see Ljubešić & Klubička (2014) for more information). The main difference between this corpus and the previous ones is that it consists, in its entirety, of texts published online (texts with a *.hr* web domain), such as online newspaper articles, blogs, forum discussions etc. and barely any literary texts.

After introducing the reader to the various Croatian corpora, in the following section we present what the respective corpora revealed about the phenomenon of doubletism in Croatian. In our research we have consulted all three corpora, depending on which one of them was available at the time of search or more appropriate for the type of query that was needed.

4.3. Declension patterns of Croatian nouns

Croatian nouns are traditionally classified as belonging to one of three declensional patterns, based on the genitive singular ending.[81] In the past, other classification criteria have been used as well (e.g. based on gender, nominative ending etc., see Tafra (1981), Marković (2012, pp. 268-9) for an overview); however, we present our data using the former classification as it is the one that is most widely used and contains the least subdivisions and exceptions. We start each section by presenting the endings of the whole declension class, followed by an individual analysis of the cells which exhibit doubletism. Since giving a complete frequency list for all nouns in the respective cases would make the table and the whole dissertation unreasonably long, the frequency tables below contain only a selection of instances that showcase variation. Information as to which corpus the frequency distributions about a particular case come from is given next to the individual tables. If the distributions differ substantially in two different corpora, this is indicated in the text. Unless stated otherwise, in all the tables below, lexemes are sorted by *cumulative frequency*, i.e. a sum of individual frequencies of the two (or more) doublet forms in a particular case.

In general, doubletism in Croatian can be of two types: *doubletism of endings* (when two endings are used interchangeably) and *stem doubletism* (when the stem of the word either undergoes a certain phonological change or not). The majority of the cells described below exhibit either one or the other; there is only one cell which combines the two types (masculine plurals, Sections 4.3.1.3 and 4.3.1.4), resulting in tripletism.

Before proceeding with the analysis of individual cells, we would like to explain one alternation that will keep reappearing in several sections below. It is the vowel-zero alternation in stems, widely referred to as 'fleeting *a*' (Croatian: "nepostojano *a"*). This refers to the vowel /a/, which appears in the stem of words ending in *-ac*, *-ak*, *-alj*, *-am*, *-ar* and *-at* only in the nominative singular (Nsg) and genitive plural (Gpl) and disappears in all other cases (e.g. Nsg *ručak* 'lunch' > Gsg

[81] Some grammarians (e.g. VHG (2007), Marković (2012)) distinguish indeclinable nouns, such as foreign names (*Ines, Dagmar*) as a separate, fourth declension class.

ručka, Dsg *ručku*, …, Gpl *ručaka*).[82] This /a/ replaced the Old Slavonic semivowel *yer* (ь), which was dropped when in the weak position in later phases of development of Common Slavic. Throughout this text, we will use round brackets around the vowel (a) to indicate that it is fleeting in a particular word.

4.3.1. A-declension

Only masculine and neuter nouns belong to this pattern. The majority of masculine nouns end in a consonant in Nsg; however, nouns that end in -o (*auto* 'car', *dečko* 'boy', neuters), -e (personal names like *Hrvoje,* neuters) and foreign nouns in -i and -u (*kivi* 'kiwi', *hobi* 'hobby', *intervju* 'interview', *Peru* 'country name' etc.) also belong to this pattern. The whole paradigm is given in Table 1.

Table 1. A-declension endings

Case	Singular	Plural
Nominative	-ø, -e, -i, -o, -u	-i
Genitive	-a	-ā
Dative	-u	-ima
Accusative	=N (animate), =G (inanimate)	-e
Vocative	-e, -u	-i
Locative	-u	-ima
Instrumental	-om, -em	-ima

4.3.1.1. Doubletism in the vocative singular (Vsg)

In general, nouns that end in a hard consonant take -e, whereas nouns that end in a soft consonant (palatal) take -u in this case. However, grammar books identify some nouns that can take both, such as those in *-ar*, *-er* and *-ir* (*gospodar* 'master' > *gospodare / gospodaru, leptir* 'butterfly' > *leptiru / leptire*), -dak (*predak* 'ancestor' > *preče / pretku*), *-tak*, *-sak*, *-šak*, *-zak*, *-žak*. The third option in this cell is to use the Nsg form. This regularly happens with foreign names; however, Šipka (1957) argues for this principle to be extended to proper nouns of Slavic origin as well (e.g. *putnik*

[82] Brozović (1976) believes that instead of saying that /a/ disappears in the majority of cases (diachronic view), it is better to say it is inserted in Nsg and Gpl (synchronic view).

'traveller' > *putniče*, but surname *Putnik* > *Putnik*). Since the vocative is the least frequent case in Croatian and rarely appears in written language, no corpus search in any of the corpora returned enough results for a more thorough investigation of this case. Previous analyses of this case (Šipka (1957), Mlađenović (1977)) have noticed the tendency of -e to spread to all nouns and a gradual elimination of -u.

4.3.1.2. Doubletism in the instrumental singular (Isg)

Similar to Vsg above, the ending in this case is also phonologically conditioned; nouns that end in a hard consonant take -om, those that end in a soft consonant (as well as vowels -e and -i) take -em. However, there are some groups of nouns for which grammar books license both endings.

a. Nouns ending in -ar

This family consists of complex nouns derived from other nouns or verbs which express the meaning of Agent, i.e. a person doing the action expressed by the root (e.g. *kuhati* (v) 'to cook' > *kuh-ar* 'a cook, person doing V'; *politika* (n) 'politics' > *političar* 'a politician, person doing N').[83] How did this doubletism arise? From a synchronic point of view, since /r/ is a hard consonant, it would be natural for it to take -om. However, in an earlier period of the language these nouns used to end in a soft semivowel (/rь/). So for this subclass of nouns the use of -em would be primary and historically older and the use of -om natural from a present-day perspective.

Our original search in HNK retrieved 252 *-ar* nouns used in Isg; there were 2,003 tokens of *-arom* and 286 tokens of *-arem*. A later search in a much larger corpus, hrWaC14 (which was not available at the time of the original search), retrieved even

[83] This principle would therefore exclude simple nouns, such as *dar* 'present', *papar* 'pepper', *požar* 'fire' etc. as well as nouns in *-ar* borrowed as a unit, such as *žandar* 'gendarme' (< French), *bećar* 'bachelor' (< Turkish), *sekretar* 'secretary' (< Latin). This latter group cannot be decomposed into the stem (**žand-, *beć-, *sekret-)* and ending as the other *-ar* nouns. For both of these groups -om is the only licensed ending. The only exception is the noun *car* 'emperor', which is a simple noun but with the same etymology as the other *-ar* nouns (< OS *сьsarь). This noun is one of the rare lexemes that appears in the corpus more frequently with -em (21 tokens of *carom* and 44 of *carem* in HNK).

larger numbers, but the general pattern was the same: -om is in general the more frequent form, but the actual proportions of the two vary from one lexeme to another. We identified very few nouns where -em was the more frequent choice, usually in low frequency words. We noticed that nouns ending in *-čar* show the greatest amount of variation. A selection of these nouns is presented in Table 2 below.

Table 2. Frequency of Isg doublet forms of nouns ending in -čar (Source: hrWaC14)

Lexeme ('gloss')	**N (-om)**	**N (-em)**
političar ('politician')	422	372
kritičar ('critic')	129	116
matičar ('registrar')	87	158
alkoholičar ('alcoholic')	121	106
tehničar ('technician')	131	75
desničar ('right-winger')	105	87
komičar ('comedian')	93	97
mehaničar ('mechanic')	81	65
ljevičar ('left-winger')	72	62
otmičar ('kidnapper')	35	67
knjižničar ('librarian')	51	26

Is it possible that there are certain contexts which favour the use of -om and others where -em is used? The only potential factor that is occasionally mentioned in grammar books in connection with this case in other declension classes is the presence or absence of a preposition. To check whether the same principle applies here, we performed a simple analysis of the dataset retrieved from HNK. Of the 2,003 tokens of -om, 1,480 were preceded by a preposition, 523 were not. Of the 286 tokens of -em, 201 were preceded by a preposition and 85 were not. The result of a chi-square test[84] for this distribution is 1.671 ($p = .196$). This small-scale analysis shows that the presence of a preposition does not play a role in the choice of ending in this instance.

b. Nouns ending in a soft consonant

As was said above, the general rule across grammars of Croatian is that nouns that end in a soft consonant should only take the -em ending. However, what has not

[84] A chi-square test is used in statistics to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories and tries to determine whether this difference is due to sampling variation or it is a real difference.

been said is that there are additional sub-rules and exceptions to this rule, which often differ from one reference manual to another. For instance, Raguž (1997, p. 11) prescribes that "-em follows palatals **and** /c/, **except** in foreign nouns" (translation and highlight mine). The authors of *Velika hrvatska gramatika* (VHG, 2007, p. 321), on the other hand, state that apart from nouns ending in a palatal, -em should also be used with nouns ending **only** in *-(a)c* and *-ic* (there is no mention of foreign nouns, highlight mine). Therefore, for instance, according to Raguž (1997), the noun *princ* 'prince' should take only -om, but according to VHG (2007), -em should be the ending. What happens in practice is that this noun appears with both endings (55 tokens of *princom* / 37 tokens of *princem* in HJR). However, the examples of doubletism do not end here.

The above rule that -em should be attached to nouns that end in a soft consonant is not applicable to all such nouns. Rather, in situations when the final soft consonant is preceded by the vowel /e/, -em is replaced by -om, so as to avoid the occurrence of two /e/'s in succession, This change, which takes place in other languages as well, is known as *dissimilation.*[85] However, as in the previous case, the descriptions of this principle in grammars are vague and contradictory. For instance, Raguž (1997, p. 12) says, "if /e/ precedes the final palatal, the ending is **usually** -om: *bodežom, lupežom, padežom, crtežom, keljom, Senjom, Bečom, koledžom*. But -em is also **not unusual**: *keljem, crtežem* etc." (translation and highlight mine). A further complication lies in the fact that dissimilation is not applicable across the board either; rather, it only applies to mono- and disyllabic nouns. Nouns with three or more syllables follow the general rule for soft consonants and take -em.[86] Given this multitude of rules and sub-rules, it is no wonder our corpus search identified numerous instances of both endings appearing with these nouns. Mlađenović (1977, p. 52) points out that the use of -om with this family of nouns "demonstrates a tendency for the modern language

[85] Nevins (2012) gives other names for this phenomenon, such as *haplology, repetition avoidance*, or *anti-homophony*. The underlying principle behind this phonological change is the morphological Obligatory Contour Principle, proposed by Joseph Greenberg in the 1950s, which prevents the combinations of two homorganic phonemes (phonemes with the same place of articulation).

[86] Different grammars interpret this differently as well; whereas VHG (2007, p. 322) says dissimilation does not apply to nouns with *three or more* syllables, HG (2005) says it does not occur in nouns with *more than three* syllables.

to develop in one direction, i.e. an overgeneralisation of -om (it is spreading out to classes to which it does not originally belong)" (translation mine).[87]

We give a selection of the more interesting examples in Table 3 below. Nouns containing /e/ in the final syllable, which should undergo dissimilation, are underlined, foreign nouns are marked with an asterisk. We can see that most variation appears in nouns ending in *-ej*, *-š* and foreign nouns.

Table 3. Frequency of Isg doublet forms of nouns ending in a soft consonant (Source: hrWaC14)

Final palatal	**Lexeme ('gloss')**	**N (-om)**	**N (-em)**
-c	udar(a)c ('kick')	32	9,875
	maslac ('butter')	148	2,664
	mjesec ('moon, month')	1,905	37
	pal(a)c ('thumb')	124	1,490
	princ* ('prince')	462	274
	šverc* ('smuggling')	602	2
	stric ('uncle')	44	383
	blic* ('flash on a camera')	252	162
	vic* ('joke')	75	164
	plac* ('market')	91	46
-č	mač ('sword')	26	3,688
	Kovač (surname, Smith)	109	672
	grč ('cramp')	61	509
	Beč* ('Vienna')	434	12
	kič* ('kitsch')	25	278
	kauč* ('couch')	90	189
	meč* ('match')	149	71
	linč* ('lynching')	92	99
-dž	imidž* ('public image')	433	509
	bedž* ('badge')	47	25
-j	odgoj ('upbringing')	73	3,245
	tramvaj* ('tram')	83	2,838
	muzej* ('museum')	1,429	345
	sprej* ('spray')	385	1,280
	volej* ('volley')	205	407
	trofej* ('trophy')	189	279
	Sergej* (personal name)	156	240

[87] Croatian original: "Ovo pokazuje tendenciju razvitka današnjeg srpskohrvatskog književnog jezika u određenom pravcu, tj. težnju da se nastavak -om što više uopšti (odnosno širi se i na imenice kojima izvorno ne pripada)."

	Jur(a)j (personal name)	72	314
	displej* (‘display’)	117	68
-lj	ravnatelj[88] (‘principal’)	79	3,571
	detalj* (‘detail’)	106	1,072
	portfelj* (‘portfolio’)	155	476
	roštilj* (‘grill’)	82	530
	kelj* (‘kale’)	83	18
	štambilj* (‘stamp’)	45	44
	hmelj (‘common hop’)	35	17
-nj	konj (‘horse’)	28	1,339
	glež(a)nj (‘ankle’)	21	173
	Rovinj (town name)	97	70
	Senj (town name)	140	7
-š	miš (‘mouse’)	283	3,528
	Bush* (surname)	840	269
	sportaš (‘athlete’)	16	624
	gulaš* (‘goulash’)	153	113
	finiš* (‘race finish’)	75	161
	marš* (‘march’)	58	91
	leš* (‘corpse’)	126	13
-ž	križ (‘cross’)	223	4,664
	staž* (‘internship’)	1,301	1,410
	crtež (‘drawing’)	880	115
	pejsaž* (‘landscape’)	151	97
	kolaž* (‘collage’)	86	72
virtual palatals[89]	plašt (‘cloak’)	1,023	864
	dužd* (‘Venetian doge’)	25	18
-t’ [90]	kut (‘angle, corner’)	2,765	1,991
	šut* (‘shot’)	1,008	416
	autoput (‘motorway’)	773	294

[88] Even though its final vowel is /e/, this lexeme should not undergo dissimilation because it has three syllables.

[89] “Virtual palatals” is a label given to clusters /št/ and /žd/ by VHG (2007, p. 322) – even though the final phoneme is hard, the preceding soft consonant has a significant influence on the perception of the whole cluster as palatal-sounding, hence the authors license *-em*. Historically, these nouns did actually end in a soft phoneme (e.g. *dažd* < OCS дъждь, cf. Russian *dožd’* ‘rain’), so once again, as in the example of *-ar* nouns, we see language keeping an archaic feature alive, although from a synchronic perspective, -om should be licensed.

[90] Similarly as with *-ar* nouns, the final consonant of the nouns in this cell is hard in present-day language, but diachronically it used to be soft (t’ < OS *tь).

We omitted one very frequent noun, *put* 'path, road' (< OS pǫtь) from the last row of Table 3. The reason for this is of a technical nature. Namely, one of the Isg forms of this noun, *putem*, has over time grammaticalized into a preposition with the meaning 'by means of, by way of'. However, all the tokens of *putem* (of which there are more than 100,000) in all three corpora have been annotated as a preposition, even though just by looking at the first couple of occurrences we could notice numerous examples of this form being used in its original, nominal, meaning. We concluded that a manual separation of these two uses would be too time-consuming, so we had to consult other sources to determine the distribution of the two forms. We bring some observations from Ham's (2002) article. Maretić (1899) introduced a functional distinction so that *putom* be used when preceded by a preposition and *putem* on its own.[91] Ham dismisses this principle as she found a roughly equal distribution of both forms in the governed position in older Croatian writers. The author also conducted a survey among 106 speakers of Croatian, asking them to use this noun in both positions. Unfortunately, the author does not provide us with much additional information about the design of the survey or respondent structure. The results were the following: in the governed position -om was selected 94 times, while -em was selected 140 times; in the ungoverned position, -om was selected 35 times, -em 199 times. The author does note that the respondents seemed visibly confused while filling out the survey and that this confusion and insecurity was often exhibited by the respondents correcting their previously marked answers (Ham, 2002, p. 135).

Some more recent grammars (e.g. VHG (2007)) introduce a different distinction – *putom* should be used in a literal meaning, 'a road for travelling, journey', whereas *putem* should be used metaphorically to refer to the way of performing an action (e.g. *slati elektroničkim putem* 'to send something electronically'). Literary sources once again do not give credit to such a principle as both forms are used in both meanings in literature with *putem* dominating. Ham's survey has shown similar patterns – 35 of her respondents used *putom* and 199 *putem* in the literal meaning; in the metaphorical meaning, 45 used *putom* and 189 *putem*. Hence the author believes there is no semantic distinction between the two forms either. A similar semantic distinction could also be assumed for the noun *mjesec*, which can have two meanings.

[91] Weisser (2006) calls the former 'governed' and the latter 'ungoverned' instrumental.

However, a manual analysis of the 37 tokens of the form *mjesecem* has shown 23 uses of this form in the meaning of 'moon' and 14 uses in the meaning of 'month'. So we consider the possibility of a semantic distinction unlikely for this noun as well.

c. Nouns ending in -io

The final group of nouns which show variation in Isg are nouns ending in *-io*. These are mostly nouns of foreign (usually Latin) origin, such as *radio*, *studio*, *trio*, personal names of Romance origin such as *Dario, Mario, Antonio*, various Italian surnames (*d'Annunzio*) etc. The majority of grammars do not deal with these nouns. Some of them only discuss the genitive singular forms, assuming that other cases should be predictable from the genitive stem (cf. Babić (1979, p. 88)). For instance, if the genitive of *Mario* is *Marij-a* (which all normative manuals agree upon), then its stem is *Marij-*. The instrumental should therefore be *Marij-em* (since /j/ is soft), and it is this ending that appears more or less consistently in older Croatian writers, as well as grammars by Raguž (1997) and VHG (2007). However, some other manuals (HG (2005), HJS (1999)) introduce opposing solutions by noting the genitive as *Marij-a*, but the instrumental as *Mari-om*. The most recent orthographic manual of Croatian (HP, 2013) is the only one that licences both forms. Table 4 provides the frequency distributions of the two variants of the most frequent *-io* nouns in the corpus, sorted by cumulative frequency in Isg.

Table 4. Frequency of Isg doublet forms of nouns ending in -io (Source: HNK, version 3.0)

Lexeme	**N (-iom)**	**N (-ijem)**
Mario	151	49
Antonio	75	27
radio	38	52
Silvio	27	52
Dario	26	29
studio	14	25
DiCaprio	18	12

4.3.1.3. Doubletism in the plural paradigm

In the plural we do not encounter doublet endings; what we encounter is a different issue. The plural of nouns in the a-declension is formed by either (a) attaching

the appropriate set of endings from Table 1 directly to the stem or (b) extending the stem with the morpheme -ov- or -ev- and then attaching the same set of endings. For convenience's sake, in the remainder of the text we will refer to (a) as the 'short' plural and (b) as the 'long' plural. The long plural is, therefore, characterized by *extended exponence*, marked both by the stem extension and the plural endings. The choice between the short and the long plural is, in general, established on the basis of the number of syllables a noun has, in the following way:

1) monosyllabic nouns – all Croatian grammars agree that all monosyllabic nouns form the long plural (with a few exceptions, such as *konj* 'horse' > *konj-i*, nationalities and currencies). However, they allow for some nouns to appear with the short plural as well. The list of such words differs from grammar to grammar. The short plural is often labelled by grammarians as 'poetic' or 'expressive' and is mostly used in literature (more on that below). However, Anić (1973, p. 15) claims that the short plural is never ungrammatical with any of these nouns.
2) disyllabic nouns with 'fleeting *a*' – we have already explained the term 'fleeting *a*' at the beginning of Section 4.3 above. Since the vowel /a/ appears only in Nsg and Gpl, the stem of such nouns is disyllabic only in those two cases, whereas in all other cases it becomes monosyllabic. When it comes to plural formation, this group is quite heterogeneous. Grammars have not been able to devise a precise rule for plural formation but rather give lists of words which form the short plural, those which form the long plural and those that can form both (the most extensive list appearing in Samardžija (1987)). Into this group we also include nouns ending in *-ao*, where /a/ fleets and /o/ devocalises to /l/ in the oblique cases, resulting in a monosyllabic stem (e.g. *pos(a)o* 'work' > *posl-*).
3) 'proper' disyllabic nouns – for all nouns of this group, the short stem is the licensed one. However, nouns of a certain prosodic structure (short-descending accent on the first syllable) allow the long stem as well. But, contrary to what we saw in monosyllabic nouns, such alternative forms are not labelled as 'poetic'. Anić (1973, p. 15) argues against invoking "poetic purposes" as a distinguishing feature of the two forms. Just like a shorter plural of a monosyllabic noun might be used for the purposes of rhyme or rhythm, a

longer plural form of a disyllabic noun might be necessary to retain dactyl rhythm. However, the latter is never considered as an option in grammars. Nouns with more than two syllables form only the short plural.

Short plurals are diachronically older than long plurals. The short form was inherited from the large, open class of Old Slavonic o-stems, whereas the long forms were typical for u-stems, which was a small, closed class of high-frequency nouns. The long plural first started to be used in Gpl to distinguish it from Gsg (even though the -ā in Gpl is phonetically longer, the macron above it is rarely used in writing). The stem extension later started to be applied to other plural cases by analogy.

There are a few nouns where the short/long plural carries a semantic distinction, e.g. *sat* > *sati* 'hours' / *satovi* 'clocks, watches', *akt* > *akti* 'office papers' / *aktovi* 'nude portraits', *čin* > *čini* 'magic spells' / *činovi* 'military ranks', *trak* > *traci* 'rays of sun' / *trakovi* 'road lanes' etc. The distinction is more or less accurately reflected in language use with all abovementioned lexemes, with the exception of *trak*. A manual analysis of 119 tokens of *trakovi* retrieved from HNK identified 39 uses of this lexeme in the meaning 'rays of light'. Similarly, we found 12 tokens of *traci* (out of a total 139 tokens) used in the meaning 'road lanes'. There are no indications of such a semantic distinction for other lexemes in this class.

The frequency distributions of the two plurals of selected nouns are given in Table 5. We give separate frequencies for the literature and specialised sub-corpora of HJR, to check the grammarians' claim that short plurals are 'literary'. The numbers in the table reflect the cumulative frequency of a lexeme in all seven plural cases. However, as we have said, in some cases (especially in groups 1 and 3), the short form of Gpl is identical to Gsg,[92] so it was often impossible to determine the actual frequency of short Gpl forms. In the end, for this case we retrieved only the forms preceded by an attribute in Gpl (unambiguously signalled by the ending -ih).

[92] However, see Vukušić (1990), who claims that the Gpl of nouns such as *vuk, bog, rog* 'horn' is not *vukā, rogā, bogā*, but rather *vucī, rozī, bozī*, i.e. identical to Npl. On the other hand, Raguž (1997, p. 17) notes that the majority of nouns that have short plural as a stylistic marker cannot have short genitive plurals; if they do, they are very rare.

Table 5. Frequency of doublet forms in the plural paradigm (Source: HJR)

Number of syllables	Lexeme ('gloss')	Fiction		Non-fiction	
		N (short)	N (long)	N (short)	N (long)
1σ	znak ('sign')	216	1,115	288	1,807
	val ('wave')	228	957	8	716
	dar ('gift')	108	652	1	996
	zvuk ('sound')	281	470	304	401
	bog ('god')	127	837	1	452
	ključ ('key')	112	202	6	761
	grijeh ('sin')	431	16	489	37
	vuk ('wolf')	166	179	11	576
	vijek ('life span')	85	338	13	73
	vrag ('devil')	119	125	2	218
fleeting *a*	pos(a)o ('work')	164	3,166	89	24,572
	ot(a)c ('father')	748	180	409	590
	vjet(a)r ('wind')	112	400	1	379
	lak(a)t ('elbow')	204	205	7	41
	puc(a)nj ('bang')	15	8	98	147
	vrš(a)k ('tip, top end')	198	16	30	3
	pijet(a)o ('rooster')	78	84	3	91
	bljes(a)k ('flash')	12	33	2	92
	glež(a)nj ('ankle')	36	47	0	43
2σ	slučaj ('case')	20	1,292	111	9,512
	korijen ('root')	132	13	934	1
	vitez ('knight')	79	345	2	211
	golub ('pigeon')	47	297	2	267
	prsten ('ring')	21	51	69	406
	pauk ('spider')	80	3	93	14
	ležaj ('berth')	31	6	37	111
	jablan ('poplar')	116	35	11	3
	plamen ('flame')	47	75	5	9
	pojas ('belt')	29	21	70	9
	kolut ('hoop')	63	29	3	13
	klaun ('clown')	2	3	25	45
compounds	stereotip ('stereotype')	21	4	191	39
	pradjed ('great-grandfather')	11	91	0	18
	sudrug ('partner')	16	29	4	10
	velegrad ('metropolis')	2	18	1	18
	polubog ('demigod')	4	8	0	10

When it comes to the question whether the short plural is more typical of literary register, the table shows that this might be a lexical matter rather than a general principle. Whereas it can be claimed that the forms *bozi* and *dari* are indeed more typical of literature, the same claim cannot be made for e.g. *zvuci* and *oci*.

Our corpus search revealed an interesting pattern with compound nouns, whose plural is usually not dealt with in grammar books (with the exception of Raguž (1997)). Being polysyllabic nouns, compounds would be expected to appear in the short plural, e.g. *skok* 'leap' > *skokovi,* but *vodoskok* 'fountain' (literally 'water-leap') > *vodoskoci*. However, we have noticed that when the second element of the compound is a monosyllabic noun with doublet plural paradigms, more often than not this doubletism is transferred to the compound as well (see last row of Table 5).

4.3.1.4. Doubletism in the long plural

Additionally, just like in Vsg and Isg, the choice of the morpheme used for forming the long plural is phonologically conditioned. Nouns that end in a hard consonant extend the stem with -ov-, whereas those that end in a soft consonant extend it with -ev-. However, once again, we come across instances of nouns that can take either of these two morphemes, especially with nouns ending in a soft consonant (and 'historic' soft consonants /r/ and /t/). In some instances, this leads to the co-existence of three plural forms, one short and two long (e.g. *pojas* 'belt' > *pojas-i / pojas-ov-i / pojas-ev-i*). The distribution of the two long variants is given in Table 6.

Table 6. Frequency of doublet forms in the long plural (Source: hrWaC14)

Final phoneme	**Lexeme ('gloss')**	**N (-ov.*)**	**N (-ev.*)**
-r	čir ('ulcer')	23	1,358
	kotar ('district')	49	877
	žir ('acorn')	124	197
-s	pojas ('belt')	139	4,427
	as ('ace')	1,389	943
	boks ('boxing')	1,546	602
	nos ('nose')	453	1,315
	faks ('fax, college')	782	862
	bas ('bass')	692	374
-t	put ('path, road')	22,188	22,530

	kut (‘angle, corner’) šut (‘shot’)	4,615 252	2,835 2,512
-z	knez (‘duke’) mraz (‘frost’) mlaz (‘gush’)	2,646 598 638	744 471 175

Ham (2002) analyses the two long plural forms of the highly frequent noun *put* (disregarding the short plural *puti*). In older normative manuals and literature, *putovi* was exclusively prescribed and used. However, a small-scale survey which the author conducted among Croatian speakers has shown a much greater preference for the form *putevi*, which Ham connects with the dominance of the analogous Isg form, *putem*. In other words, since speakers use *putem* more (see Section 4.3.1.2.b), they are also more likely to use *putevi* by analogy. The author dismisses any kind of semantic difference (described in Section 4.3.1.2.b) and concludes that *putovi* is typical for all functional styles of Standard Croatian except the conversational style.

4.3.1.5. Doubletism in the genitive plural (Gpl)

Putting aside the discussion from Section 4.3.1.3, where the whole plural paradigm was involved, there are also several highly frequent nouns in this paradigm which can have competing endings only in Gpl (-ā / -ī / -ijū). The last of these, -ijū, is a remnant of the Old Slavonic i-stem dual, which in its original function has fallen out of use a long time ago and is no longer productive as a Gpl marker.[93]

Table 7. Frequency of Gpl triplet forms of some masculine nouns (Source: HJR)

Lexeme (‘gloss’)	**N (-ā)**	**N (-ī)**	**N (-ijū)**
prst (‘finger’)	36	13	734
zub (‘tooth’)	325	404	23
mrav (‘ant’)	132	39	*
crv (‘worm’)	42	30	*
gost (‘guest’)	*	34	4,907
dečko (‘lad’)	*	38	16
nokat (‘nail’)	83	*	79

[93] This ending also appears as the sole Gpl ending of highly frequent neuter nouns *oko* ‘eye’ and *uho* ‘ear’, as well as one of the possible endings in a number of feminine nouns in the other two declension patterns, such as *kokoš* ‘hen’, *kost* ‘bone’, *ūš* ‘louse’, *grudi* ‘bosom’, *prsa* ‘chest’.

There are also masculine nouns that end in a consonant cluster, which have the option of reinserting the fleeting *a* or not. This also happens with some neuter nouns, the majority of which end in -Cce. A selection of the most frequent instances is given in Table 8.

Table 8. Frequency of fleeting a doubletism in Gpl of masculine nouns (Source: HJR)

Family	Lexeme ('gloss')	N (-ā)	N ((a)-ā)
-Ck	kiosk ('newsstand')	86	4
-Cl	ansambl ('ensemble')	24	194
	artikl ('item of goods')	4	181
	bicikl ('bicycle')	73	75
	spektakl ('spectacle')	17	33
-Ct	objekt ('object')	12	4,419
	projekt (project')	43	3,743
	incident ('incident')	8	937
	projektant ('project designer')	7	53
neuter	koplje ('spear')	7	57
	drvce ('small tree')	12	38
	jajašce ('small egg')	7	30
	deblo ('tree trunk')	3	25

4.3.1.6. Isolated examples of stem doubletism in A-declension

In some cases, doubletism occurs not because there are two endings at the speaker's disposal, but rather because a certain phonological change may or may not take place in the stem of the noun. We have already explained the phenomenon of fleeting *a* at the beginning of Section 4.3.1. The name *Jur(a)j* can, for instance, be declined with the /a/ (*Juraja, Juraju* etc.) or without it (*Jurja, Jurju* etc.). The former possibility is analogous to all other nouns that end in -aj (*kraj, utjecaj*) and the latter to other names with fleeting *a* (*Pet(a)r* > *Petra, Petru*). The forms of *Juraj* containing /a/ appear in HJR 126 times (all singular cases combined), the forms without /a/ 940.

In Section 4.3.1.3, we have also seen the process of stem extension at work in the plural paradigm. This process also takes place, to a lesser extent, in the singular oblique cases of some nouns in this declension class (proper names such as *Mile*, *Rade*, neuter nouns ending in *-ce* etc.). The issue is again whether to add the stem extension

(-et-) or not (e.g. *Mile* > *Mile*, *Mili, ...* / *Mileta*, *Miletu, …* etc.).[94] A semantic difference between the two forms appears only for one lexeme. When the noun *drvo* refers to 'tree', it is declined without the extension (*drva*, *drvu* etc.); when it is used to mean 'wood' as material, it is declined with the extended stem (*drveta*, *drvetu* etc.). No other lexemes show such a distinction.

Table 9. Examples of stem doubletism in the singular paradigm (Source: HNK, version 3.0)

Lexeme ('gloss')	**N (no stem extension)**	**N (stem extension)**
Mile (first name)	476	30
mjestašce ('small town')	110	2
zvonce ('small bell')	54	3
seoce ('small village')	17	8

4.3.2. E-declension

This declension pattern includes all nouns that end in -a in Nsg; the majority of them are feminine, but there is a significant number of masculine ones as well (*kolega* 'colleague', male personal names *Ivica, Jurica* etc.).

Table 10. E-declension endings

Case	**Singular**	**Plural**
N	-a, -o, -e	-ē
G	-e	-ā, -ī
D	-i	-ama
A	-u	-e
V	-o, -a, -e	-e
L	-i	-ama
I	-ōm	-ama

4.3.2.1. Doubletism in the dative/locative singular (Dsg/Lsg)

In this case we have another example of stem doubletism, as we are not dealing with doublet endings but rather with the (non-)execution of a phonological change.

[94] Using similar arguments as for the fleeting *a*, Brozović (1976) claims that it is better to say that these nouns shorten their stem in Nsg rather than extend it in the majority of cases.

The change in question is known as second palatalization, which involves the sybillants /k/, /g/, and /h/ changing to /c/, /z/, and /s/ respectively when followed by /i/. Since the Dsg/Lsg ending is -i, this would mean that all feminine nouns ending in -ka, -ga, -ha should undergo this change in both of these cases. However, normative manuals of Croatian are not completely clear on this phenomenon. For instance, the authors of *Hrvatski jezični savjetnik* (HJS, 1999, p. 34), admit that no coherent or valid rules of palatalization could be determined, except for proper nouns.

In older periods of the language the second palatalization was consistently carried through. However, in the modern language, it seems to be more of an exception rather than a rule. For instance, Silić & Pranjković (2005, p. 109) say that "in some geographical terms it has become habitual to change /k/, /g/, /h/, but with others it is not (*Rijeka > Rijeci, Amerika > Americi, Korzika > Korzici, but Volga > Volgi, Krka > Krki, Malaga > Malagi*)" (translation mine).[95] Monosyllabic nouns, personal names, terms of endearment and foreign nouns tend to not undergo it, whereas in other instances doubletism abounds, as visible from Table 11 below.

Table 11. Frequency of second palatalization in Dsg/Lsg (Source: HJR)

Phonol. family	**Lexeme ('gloss')**	**N (non-palatalized stem)**	**N (palatalized stem)**
-Vka	(r|R)ijeka ('river', town name)	3	6,016
	Amerika ('America')	12	3,580
	utrka ('race')	0	2,985
	šaka ('fist')	15	183
	buka ('noise')	5	140
	Krka (Croatian river)	112	1
-čka	točka ('dot')	1,169	72
-jka	majka ('mother')	37	1,569
	djevojka ('girl')	13	837
	bajka ('fairy tale')	11	99
-lka	alka (ancient Croatian game)	117	95
	jelka ('fir tree')	26	1
-ljka	križaljka ('crossword')	3	71

[95] Croatian original: "Pri uporabi nekih zemljopisnih imena stekla se navika da im se *k, g, h* mijenjaju a pri uporabi nekih da im se ne mijenjaju."

	školjka (‘shellfish’)	12	44
-mka	zamka (‘trap’)	4	41
-nka	banka (‘bank’)	24	3,366
	stanka (‘pause’)	14	261
	stranka (‘political party’)	29	131
-pka	crpka (‘pump’)	133	0
	klopka (‘trap’)	6	25
	zipka (‘cradle’)	2	25
-rka	zbirka (‘collection’)	11	521
	petorka (‘five people’)	52	232
	nuklearka (‘nuclear power station’)	178	82
	kćerka (‘daughter’)	53	40
	barka (‘small boat’)	15	37
-ska	vojska (‘army’)	26	1,973
	daska (‘board’)	39	84
	Aljaska (‘Alaska’)	11	78
	ljuska (‘crust’)	11	21
	maska (‘mask’)	18	4
	guska (‘goose’)	18	2
-ška	podrška (‘support’)	3	480
	Gradiška (town name)	215	51
	greška (‘mistake’)	48	191
	puška (‘rifle’)	38	46
-tka	bitka (‘battle’)	131	370
	čestitka (‘greeting card’)	195	92
	tvrtka (‘company’)	254	16
	pripovijetka (‘short story’)	38	26
	krletka (‘bird cage’)	26	31
-vka	Trešnjevka (area in Zagreb)	15	428
	Podravka[96] (proper name)	115	101
	pretpostavka (‘hypothesis’)	2	142
-ga	knjiga (‘book’)	17	4,777
	snaga (‘strength’)	7	2,417
	tuga (‘sadness’)	28	198
	vaga (‘scales’)	206	9
	sloga (‘concord’)	26	136
	žega (‘sweltering heat’)	33	14
	prisega (‘oath’)	19	26

[96] Težak (1986, p. 401) claims that in this case second palatalization is used to indicate a difference between two possible senses of the word. *Podravki* would refer to a woman from the north-east Croatian region of Podravina whereas *Podravci* would refer to a food company based in the town of Koprivnica. Our corpus analysis did not confirm such a pattern as both forms are used predominantly in the latter sense.

-ha	svrha (‘purpose’)	11	354
	ovrha (‘audit’)	13	106
	snaha (‘daughter-in-law’)	26	19
	zaliha (‘stock’)	21	2

Usage data indicate that the language is developing in the direction of the non-execution of second palatalization, especially in lower-frequency nouns, a tendency that was noted already by Težak (1986) in his small-scale survey of native speakers. As was said at the beginning of Section 4.3, all the tables contain only instances of doubletism; we did not include lexemes that appear only with one form. If we had, there would be, for instance, more nouns ending in *-ha* that do not undergo palatalization than those that do. It is possible that the tendency for words not to undergo this change spread from the Kajkavian and Čakavian dialects of Croatian as in those two dialects this change does not take place. Furthermore, as argued by VHG (2007, p. 385), de-palatalization probably arises from the tendency for the word not to become too phonologically distant from its original stem” (translation mine).[97]

4.3.2.2. Doubletism in the vocative singular (Vsg)

As in the previous declension pattern, the vocative singular of some nouns in this pattern can have two different forms. For instance, personal names, both male and female (*Ana, Marija, Ivica*), nouns in *-ica* (referring to female occupations), kinship terms, hypocoristics etc. can either have Vsg identical to Nsg or they can take a distinct vocative ending (-e or -o). Most normative manuals will claim that there is a stylistic difference between the two: the Nsg form is more neutral, whereas -e/-o has a more “emotionally loaded value” (VHG, 2007, p. 390) (translation mine).[98] Jonke (1965a, p. 277) mentions only the name *Marija* as having two possible forms (*Marija* and *Marijo*), without providing any additional qualifications.

Težak (1982) conducted a small-scale study of this case. The author distributed a forced-choice questionnaire among almost 200 students asking them to select the

[97] Croatian original: “Desibilarizaciju zacijelo potiče i težnja da se oblik riječi glasovno ne udaljuje od prvotne osnove.”

[98] Croatian original: “nastavak -o poprima stilski izrazitije, osjećajnije obilježje.”

most appropriate form. His results showed a great deal of variation (e.g. *Eva* was chosen by 97 and *Evo* by 85 respondents). The author also found some regional differences between students from Zagreb and Osijek, the former using more of the -o ending. Also, according to the author, nouns in *-ica* develop the abovementioned stylistic difference even further: -o carries negative connotations, whereas -e shows more affection towards the referred person (e.g. *prijateljico* 'you fake friend' vs. *prijateljice* 'my dear friend').

It is interesting to note that the vocative is the case which has shown a great deal of variation in both declension classes we dealt with so far. This is a somewhat unusual situation, as per Brown, Tiberius, & Corbett (2007, p. 523), who argue that the higher a function's frequency, the greater the number of forms associated with it, the typical intuitive assumption being that "it would be more taxing on memory to learn many forms for a function which occurs infrequently."[99] With the vocative being the least frequent case in Croatian (in spoken dialogue it is often replaced with the nominative form),[100] the fact that it can be expressed with a great number of forms could, in this view, be regarded as somewhat unintuitive.

4.3.2.3. Tripletism in the genitive plural (Gpl)

This is an interesting case of free variation because it is one of the rare occasions where the speakers have three, rather than two, endings at their disposal: (1) -ā, (2) -ī and (3) re-insertion of the fleeting *a* + -ā (e.g. *naranča* 'orange' > *narančā* / *narančī* / *naranačā*).[101] Nouns whose stem ends in a single consonant (-VCa) only

[99] The authors note, however, that the same argumentation could be applied in support of referrals. "If we assumed that in order to learn a referral-based system it is only necessary to acquire the rule which say that the form of the accusative is the same as the genitive (if animate) or nominative (if inanimate), then this is possibly less taxing on memory than learning all of the inflections as directly associated with the accusative" (Brown, Tiberius, & Corbett, 2007, p. 523).

[100] Many authors question the status of the vocative as a separate case, primarily for this reason. Brozović (1976, p. 130) claims that according to syntactic and semantic criteria it is not a real case, but according to word formation and morphological criteria it is, hence it is relevant to consider it here with other example of case marking.

[101] Similar to what we saw in the long plurals of a-declension, (3) is an example of extended exponence, as the Gpl meaning is expressed both by the presence of the fleeting *a* and the case ending.

take (1) (however, as we will see below, nouns with a syllabic /r/ also appear with doublets), whereas nouns whose stems end in a consonant cluster (-V(C)CCa) can have as many as three Gpl forms. It is this latter family that we are interested in. Of the three endings, the last one (henceforward a-a) is the most unambiguous marker of Gpl. As we have seen in Table 11, -a is also the ending for Nsg, whereas -i also appears in Dsg/Lsg.[102] The fleeting *a*, on the other hand, appears only in Gpl. From a diachronic perspective, (1) is the traditional ending (developed from the old -ah), whereas (2) is the most recent addition to the language standard (introduced by Maretić (1899); it does not appear in any older Croatian writer before him).

Similar to Dsg/Lsg above, this is another instance where Croatian grammars cannot seem to agree on the most accurate description of what happens in this case. Most of them give lists of consonant clusters which have one, two, or three options respectively. It seems that the number of nouns in this family that have only one form is extremely low. Most manuals will say this is the case with nouns ending in *-st, -št, -zd, -žd, -šč,* and *-šć*. But as we will see below, doublets are not that uncommon in these classes either. Silić & Pranjković (2005, p. 110) state that when all three forms are possible, they are equally acceptable, but that there are cases in which one of the three variants is uncommon (e.g. *majka* > *majaka*). "However, this does not mean that when such a form occurs, it is grammatically incorrect" (translation mine).[103] Some manuals (VHG, 2007) tend to give preference to variant (3) due to its unambiguity. HJS (1999), on the other hand, gives an advantage to (1), except in cases when (3) is "more common", but also states that (2) is most common in conversational style.

Normative manuals mention several examples of a semantic distinction between the different forms. For instance, the form *banaka* should be used only in the meaning of a financial institution, whereas *banki* only to refer to banknotes; *maraka* only to refer to the former German currency and *marki* for industrial brands; *stranaka* in reference to political parties and *stranki* to mean 'clients'. The numbers in the corpus do confirm this pattern, although there are occasional examples of forms being used to express the 'other' meaning. In all other cases, the variation is completely

[102] Both of these endings are phonetically longer than their singular counterparts, but they are rarely marked with a macron in written language, hence the distinction is no longer prominent.

[103] Croatian original: "No to ne znači da to kad se dogodi nije u skladu s gramatičkim pravilima."

unconditioned. In some instances, one form is preferred so as to avoid ambiguities with other forms. For instance, the form *snimaka* could be Gpl of either *snimka* (feminine) or *snimak* (masculine), both meaning 'recording'; hence, *snimki* is more unambiguous as the Gpl of the former. When we were performing our corpus analysis for lexemes such as these, it was impossible to determine what the actual Nsg form was unless there were other markers in the surrounding text. Furthermore, animate nouns that end in -Cka (referring to a feminine Agent) cannot have variant (3), although it is not explicitly stated anywhere that (1) is impossible as well.

One of the earliest empirical analyses of this case (and unfortunately a rare one as well) was performed by Težak (1980). Since no corpus of Croatian was available at that time, the author had to manually collect examples of this case from both spoken (TV and radio shows) and written language (newspapers and magazines) of the time. His corpus consisted of around 3,300 tokens. The majority of tokens (48.9%) were forms with a-a, followed by -ī (33.9%) and -ā (17.2%). The author also conducted a survey among students of Slavonic languages at the University of Zagreb asking them to choose the most appropriate variant for a number of lexemes. For the majority of nouns, the -ī forms were chosen most often. Even though the respondent sample was quite small (90) and not very representative (all respondents were students), these results show: a) a lot of disagreement between speakers and b) a dominance of -ī (cf. results of our questionnaire study reported in Section 6.4). The author concludes his article by stating that the multitude of forms should in no way be considered a defect of our standard literary language as they enable the user to "adapt to the requirements of style, euphony, rhythm and sometimes unambiguity of expression" (Težak (1980, p. 15), translation mine).

Distribution of the three genitive plural forms in present-day language is presented in Table 12 below.

Table 12. Frequency of doublet forms in Gpl (Source: hrWaC14)

Phonological family	Lexeme ('gloss')	N (-ā)	N (-ī)	N ((a)-ā)
-st, -št, -zd, -žd, -šč, -šć	vrsta ('breed, class')	42,018	1,050	0
	cesta ('road')	10,164	86	72
	pasta ('paste')	118	184	0
	pošta ('post office')	570	34	0
	fešta ('festivity')	175	249	0
	zvijezda ('star')	13,942	307	*[104]
	žlijezda ('gland')	1,860	131	*
	brazda ('furrow')	54	74	*
	gošća ('female guest')	72	278	*
-Cb	odredba ('decree')	50	13,850	4,332
	skladba ('piece of music')	66	5,530	17
	narudžba ('order')	14	3,700	112
	jednadžba ('equation')	13	1,153	4
	postrojba ('army troop')	1,063	11,900	3
	molba ('request')	21	1,035	0
	bomba ('bomb')	170	6,190	1
	borba ('struggle, fight')	201	6,080	1
	glazba ('music')	31	249	0
	služba ('service')	665	25,850	6
	optužba ('accusation')	50	9,330	521
-Cc	ovca ('sheep')	20	28	8,603
	licenca ('licence')	25	2,990	0
	vrpca ('ribbon')	73	329	3
-Čč	naranča ('orange')	28	219	207
	kopča ('buckle')	14	130	0
-Cd	milijarda ('billion')	551	99,392	3
	sekunda ('second')	445	39,900	5
	Ande ('the Andes')	476	41	0
-ck	kocka ('cube')	6	460	440
-čk	točka ('dot, full stop')	44	986	20,150
	igračka ('toy')	17	175	7,125
	pljačka ('robbery')	14	948	20
-ćk	voćka ('fruit tree')	2	185	1,482
	srećka ('lottery ticket')	0	69	48
-jk	djevojka ('girl')	68	138	16,494
	majka ('mother')	48	3,633	9

[104] Nouns ending in *-zd, -žd, -šč, -šć* cannot have the form with the fleeting *a* as they never used to contain the Old Slavonic semi-vowel *yer* from which the fleeting *a* developed.

	brojka (‘numeral’)	15	3,276	254
-ljk	biljka (‘plant’)	65	550	16,121
	pošiljka (‘parcel’)	10	790	1,460
	svjetiljka (‘lamp’)	26	911	261
	udaraljka (‘percussion’)	3	655	0
-mk	iznimka (‘exception’)	21	890	2,430
	zamka (‘trap’)	20	1,200	9
-nk	stranka (‘political party’)	519	354	43,441
	banka (‘bank’)	200	800	41,313
	stotinka (‘hundredth of a second’)	5	3,700	3
	znamenka (‘digit’)	0	736	422
-pk	tipka (‘key on a keyboard’)	7	2,274	1,085
	crpka (‘pump’)	4	1,118	0
-rk	zbirka (‘collection’)	73	3,933	48
	zamjerka (‘objection’)	14	2,005	15
	novinarka (‘female journalist’)	6	286	*[105]
-sk	daska (‘board’)	14	119	1,788
	vojska (‘army’)	71	1,705	4
	maska (‘mask’)	30	1,677	35
	guska (‘goose’)	4	47	652
-šk	greška (‘error’)	40	269	12,850
	puška (‘rifle’)	3	48	2,320
	bilješka (‘note’)	2	1,220	765
	kruška (‘pear’)	5	23	1,112
	kriška (‘slice’)	8	253	22
-tk	tvrtka (‘company’)	237	64,667	654
	rešetka (‘metal bar’)	18	363	4,844
	bitka (‘battle’)	7	300	2,506
	čestitka (‘greeting card’)	11	962	370
	pripovijetka (‘short story’)	3	134	1,211
	zagonetka (‘riddle’)	3	428	78
-vk	pretpostavka (‘assumption’)	35	2,485	105
	olovka (‘pencil’)	3	130	725
-Cl	varijabla (‘variable’)	29	1,335	0
	igla (‘needle’)	12	75	1,123
	kugla (‘sphere’)	40	700	8
	cigla (‘brick’)	26	335	186
	jasle (‘crib’, pl.t.)	20	137	139
-Clj	zemlja (‘earth, land’)	711	39	140,995
	baklja (‘torch’)	18	950	5

[105] This is an animate noun referring to a female occupation; all grammars agree that such nouns cannot have the option with the fleeting *a*.

-Cm	pjesma (‘song’)	470	206	60,654
	firma (‘company, firm’)	87	6,676	2
	norma (‘norm, quota’)	395	5,016	1
	čizma (‘boot’)	12	101	1,399
	terme (‘spa’, pl.t.)	400	1,087	0
-Cn	usna (‘lip’)	24	57	8,073
	kazna (‘punishment’)	160	6,970	0
	akna (‘acne’)	15	3,500	1
	tajna (‘secret’)	140	2,943	0
-Cnj	prijetnja (‘threat’)	205	10,200	0
	radnja (‘action’)	204	8,071	0
	sumnja (‘suspicion’)	79	2,511	0
	vožnja (‘driving’)	40	1,927	0
	trešnja (‘cherry’)	5	42	1,341
	lignja (‘squid’)	4	442	541
-Cp	Alpe (‘the Alps’)	1,657	332	0
	pumpa (‘pump’)	25	995	0
-Cr	igra (‘game’)	279	470	47,411
	litra (‘litre’)	2	273	30,797
	sestra (‘sister’)	120	72	18,638
	jezgra (‘nucleus’)	90	1,766	648
	smotra (‘parade’)	22	680	0
	šifra (‘code’)	7	483	86
-Cs	nijansa (‘nuance’)	125	5,256	2
	šansa (‘chance’)	27	5,090	3
-Ct	karta (‘map, ticket’)	40	465	21,384
	komponenta (‘component’)	105	5,419	2,595
	sorta (‘species’)	130	3,773	1,770
	varijanta (‘variant’)	50	3,016	268
-Cv	žrtva (‘victim’)	125	323	36,184
	crkva (‘church’)	95	1,210	9,860
	rezerva (‘reserve’)	65	4,936	0
	smokva (‘fig’)	2	120	1,825
	gužva (‘crowd’)	65	1,658	1
	bačva (‘barrel’)	4	443	632
	ljestve (‘ladder’, pl.t.)	9	360	516
	breskva (‘peach’)	3	119	586
syllabic /r/	mrlja (‘blot, blemish’)	1,605	34	*
	zakrpa (‘patch’)	300	164	*
	kovrča (‘hair curl’)	107	52	*
	svekrva (‘mother-in-law’)	88	30	*

The corpus analysis presented in Table 12 above has shown one interesting pattern: within each phonological family the majority of nouns appear more frequently with -ī; however, the most frequent nouns in each family have one of the other two endings as the most frequent one (e.g. *ovca, točka, djevojka* etc.). It seems these highly frequent nouns somehow become autonomous from other words in their phonological families and create a pattern of their own. It is a well-known fact in morphology that the most frequent items are irregular and less prone to regularization (see Bybee & Hopper (2001)). Besides frequency, two more factors are said to influence the degree of autonomy, namely semantic simplicity and morphophonemic regularity (Bybee J. L., 2007, p. 14). This idea of autonomy of highly frequent forms will be revisited in later sections (Chapters 5 and 6).

There are a few additional nouns which also exhibit doubletism in this case, but with other endings involved. For instance, *ruka* 'arm', *noga* 'leg' and *sluga* 'servant' take both -ā and -ū, the latter being a remnant of the Old Slavonic ā-stem dual. All three nouns appear more frequently with the latter ending in HJR. *Pluralia tantum* in this declension class, *usta* 'mouth', *vrata* 'door' and *prsa* 'chest', can take both -ā and -ijū, with the former variant being more frequent in HJR.

4.3.3. I-declension

This declension pattern includes feminine nouns that end in a consonant. Babić (2006) claims that this declension class is closed in present-day Croatian, or more specifically, semi-closed – it takes only new derivations with *-ost* (which is a highly productive suffix), *-ež* and *-ad*, whereas any recent feminine neologism or loanword would go to the e-declension. The paradigm of these nouns is presented in Table 13.

Table 13. I-declension endings

Case	Singular	Plural
N	-ø	-i
G	-i	-ī
D	-i	-ima
A	-ø	-i
V	-i	-i
L	-i	-ima
I	-i, -ju	-ima

4.3.3.1. Doubletism in the instrumental singular (Isg)

This is the only case that we have come across so far where doubletism is a feature of all nouns in the declension pattern rather than being present in a certain phonological subfamily or a few isolated examples. Similar to what we have seen in Gpl of the e-declension, one of the endings is polyfunctional (-i), whereas the other one (-ju) is not. The former ending is a direct reflex of the inherited class in PIE (*-eh$_1$), whereas the latter was borrowed from Old Slavonic ā-stems. As with several previous cases, the doubletism was introduced by Maretić (1899). Most Slavonic languages use only the latter ending for nouns of this declension pattern (cf. Russian *kósťju*, Polish *kością*). Older Croatian writers also used -ju almost exclusively.

However, some normative manuals (starting again with Maretić (1899)) also try to delimit the usage of the two endings. For instance, the grammar by Silić & Pranjković (2005, p. 111) says that -ju can be used in all instances, but that -i should be reserved for instances when the noun is preceded by an adjective, pronoun or preposition. In other words, -ju should be used whenever it might not be clear from the context which case we are dealing with (as -i appears in the majority of cells in this declension pattern). However, Težak (1989, p. 39) notes that such ambiguous situations are extremely rare in Croatian. Labov (1994) explicitly argues that the drive to reduce ambiguity, as useful as it might be to language users (and therefore language use), is never much of a driving force in grammatical change. Hence it is unlikely that a language will develop a functional distinction between two of its elements solely because one of them is polyfunctional.

The usage data presented below show that the two are used almost equally, regardless of the syntactic environment. This led the authors of *Hrvatski jezični savjetnik* to finally admit that "attempts to separate the two endings so that -ju is used in isolation and -i when preceded by a preposition or adjective have not borne fruit" ((HJS, 1999, p. 85), translation mine).[106]

Težak presented 37 Croatian language teachers with a questionnaire in which he asked them which instrumental form they found acceptable (the author does not

[106] Croatian original: "Pokušaji da se raspodjela tih nastavaka razgraniči tako da se nastavak -i upotrebljava uz prijedlog ili pridjev, a nastavak -ju samostalno, ne daju ploda."

give information about the scale used, but we assume it was a binary scale). All examples in the questionnaire were instrumentals without a preposition or attribute. The relationship of the preferred morphemes was: 261 -i: 184 -ju: 68 both. "From this it can be implied (1) that teachers show disagreement with grammars, which in certain syntactic surroundings (such as the one in question) give precedence to -ju endings and (2) that it would be possible to speak of the connection between the instrumental ending and the final consonant" (Težak, 1989, p. 37) (translation mine).[107] For instance, the majority of authors will agree that -ju is more common with nouns ending in -*st* (but not -*rst*), -*št*, -*ao*, -*đ*, and -*ć*; both endings are used equally with nouns ending in -*č, -s, -š,* and -*v*, whereas -ju is almost never used with nouns ending in -*d, -n, -r, -t* (excluding -*st*), and -*ž*. However, as was already pointed out, the authors of these normative manuals do not provide any information as to how they arrived at such distributions.

The largest class of nouns in this family are abstract nouns ending in -*st*. This class has been researched quite extensively in recent years by Grčević (2006), (2007), both from a synchronic and diachronic perspective. For instance, Grčević (2007) performs a corpus analysis of these nouns in the 220-million token media sub-corpus of HNK. The author's analysis concludes that the nouns in question always take -ju if they are not preceded by an instrumental marker. Some exceptions which appear with the ending -i do not interfere with the abovementioned principle. On the other hand, when they are preceded by an instrumental marker, they tend to use both and the distributions of the two tends to vary. The nouns that use -ju the most, even when they are preceded by an instrumental marker, are those ending in -*ost*, which form the largest subclass within this family. The author concludes that "with these nouns -i serves as a marginal ending which makes the linguistic expression richer and more diverse" (Grčević, 2007, p. 21) (translation mine).[108] Some of the distributions the author found are given in Table 14 below.

[107] Croatian original: "Očito je na prvi pogled (…) da su [nastavnici] manje suglasni s gramatikama koje u danim sintaktičkim položajima daju prednost nastavcima -ju/u ili bar dopuštaju ravnopravnost s nastavkom -i, c) da bi se moglo govoriti o vezanosti instrumentalnih nastavaka uz određeni tip imenica, na što osim upotrebne čestoće utječe i završni suglasnik."

[108] Croatian original: "S imenicama na -*ost* nastavak -i ima ulogu rubnoga nastavka kojim se jezični izričaj čini raznovrsnijim i bogatijim."

Table 14. Frequency of Isg doublet forms of nouns ending in -ost (adapted from Grčević (2007))

Phonological family	Lexeme ('gloss')	N (-i)	N (-ju)
-ast	vlast ('position of power')	297	1,283
	ovlast ('jurisdiction')	140	26
	mast ('fat')	11	14
-est	bolest ('illness')	208	642
	povijest ('history')	44	413
	obavijest ('notification')	41	17
-ost	javnost ('the public')	224	5,765
	sigurnost ('security')	58	2,516
	mogućnost ('possibility')	99	1,067
	prednost ('advantage')	53	588
	vrijednost ('value')	144	551

Grčević also notes two interesting patterns concerning the use of -i: 1) it is used almost exclusively when these nouns are used as proper nouns (e.g. *Mladost* as the name of various Croatian sports clubs) and 2) it is much more frequent than -ju in non-suffixed nouns, such as *čeljust, kost, korist, ovlast* etc.

Finally, Ham (1996) believes that the authors who argue for the exclusive usage of -i (regardless of its polyfunctionality) are led by the 'nationalist' principle – as -i is virtually non-existent in Isg in Serbian, they argue it should be used in Croatian precisely for that reason. The author believes this principle is justified only when used in combination with traditional and systemic reasons, but never on its own. The author adopts Težak's (1991, p. 89) view, who concludes that "whenever there is a device in language that enables faster and safer reception of information, it should always be used. This means that -ju will never be incorrect, even when preceded by an attribute or preposition" (translation mine).[109]

4.3.3.2. Gender doubletism

There is a small class of nouns that end in -ež whose gender is unclear. They are sometimes classified as feminine nouns, hence belonging to this declension

[109] Croatian original: "Kada u jezičnom sustavu postoji razlikovno sredstvo koje omogućuje brže i sigurnije primanje obavijesti, treba ga iskorištavati, a to znači da nastavak -ju neće biti nekoristan, a ni nepravilan niti uz prijedlog ili atribut."

pattern; in other cases they are classified as masculine and belonging to the -a declension pattern. Data from grammar books and dictionaries, as with plenty of cases before, show a lot of variation and inconsistencies. A questionnaire study conducted by Bošnjak (2005) has shown that native speakers show the same level of variation. Only in a couple of nouns did all respondents (N = 34) give a unanimous answer (e.g. *crtež* (m) 'drawing', *mladež* (f) 'youth'). Interestingly enough, these were the more frequent nouns in the group and also nouns for which corpus data reveal a strong preference for one of the paradigms.

Besides this family of nouns, gender doubletism also appears in some individual nouns in this declension pattern (*bol* 'pain', *glad* 'hunger', *čar* 'charm', *večer* 'evening', *pelud* 'pollen'). Ignjatović (1960) argues that, in the case of the plural of *bol*, there is a semantic distinction between the two genders – the feminine form *boli* should be used for psychological, mental pain, whereas masculine *bolovi* is reserved for physical, tangible pain. To wish someone a "Good evening" in Croatian, the speaker has three options: *Dobar večer* (masculine), *Dobra večer* (feminine) or *Dobro veče* (neuter). The masculine form is the most frequent in written language (124 / 55 / 22 in HJR), although the other two appear more frequently in speech. Historically speaking, *večer* used to be masculine and is masculine in a majority of Slavonic languages (cf. Polish *wieczór*, Czech *večer*). Ham (2004) analyses the gender of the noun *pelud*. The author gave 100 students sentences in which they had to use this noun in various cases. 26 respondents used it exclusively in the masculine, 9 used it exclusively in the feminine, whereas the rest mixed the two genders, using a-declension endings in some cases and i-declension in the other.

4.3.3.3. Isolated examples of doubletism in I-declension

The noun *kći* 'daughter' has, in recent years, started to undergo an interesting analogical process. Namely, this noun extends its nominative stem in all the oblique cases (Gsg *kć-er-i*). However, in recent years this stem extension started to spread to the nominative as well, so numerous examples of *kćer* as the Nsg form can be found

(1,970 *kći* / 535 *kćer* in HJR).[110] The same analogy occurs with the e-declension noun *mati* 'mother' > Gsg *mater-e* > Nsg *mater*.

4.3.4. Declension of adjectives and pronouns

Croatian adjectives, besides the usual gender-number-case distinction, also mark an additional category – definiteness. However, since this category is syntactically conditioned, an example such as *dobar čovjek / dobri čovjek* is not an example of doubletism of forms. In short, possessive adjectives in *-ov*, *-ev*, *-in* as well as possessive pronouns can have only indefinite marking, adjectives in *-ski*, *-ški*, *-čki*, and all comparatives and superlatives can only have definite marking.[111] Only descriptive adjectives can appear in both forms. The main difference between the two uses is semantic, similar to the difference in meanings expressed by articles in English. Indefinite adjectives highlight only one feature of the noun (N is X, not Y, Z etc.), whereas definite forms identify the noun they modify (the N that has already been described as X). Definite and indefinite adjectives have separate declensions (although this distinction is evident only with masculine adjectives). In Table 15 below we present only the definite declension pattern as it is here that doubletism occurs.

Table 15. Pronominal-adjectival declension endings

	Singular			**Plural**		
	masc.	**fem.**	**neut.**	**masc.**	**fem.**	**neut.**
N	-ø, -i	-a	-o, -e	-i	-e	-a
G	-og(a), -eg(a)	-e	-og(a), -eg(a)	-ih	-ih	-ih
D	-om(u/e), -em(u)	-oj	-om(u/e), -em(u)	-im(a)	-im(a)	-im(a)
A	= G (animate) = N (inanimate)	-u	=N	-e	-e	-a
V	= N	= N	=N	= N	= N	= N

[110] A similar analogy can be found in some masculine nouns, namely those ending in *-in* (e.g. *državljanin* 'citizen'), which shorten their stem in the plural (*državljani*). Looking in the corpus, one can find numerous examples of the short form in the Nsg (*državljan*) as well. However, unlike in the example of *kćer*, where recent grammars started to acknowledge this analogy, most grammars still label *državljan* as ungrammatical.

[111] However, both Težak & Babić (2004) and Matasović (2008) acknowledge that from 19th century onwards, possessives started appearing in the definite declension as well.

L	= D	= D	=D	= D	= D	= D
I	-im(e)	-om(e)	-im(e)	= D	= D	= D

Pronouns follow the same declension pattern as adjectives (this pattern is sometimes referred as the pronominal-adjectival declension, see Marković (2012)). Furthermore, personal pronouns have accented and unaccented forms – the former are used when following a preposition and the latter in the second position in a sentence (known as Wackernagel's Law).

The distinction between allomorphs -og and -eg in Gsg and -om and -em in Dsg/Lsg is, similarly to nouns, phonologically conditioned: if the adjective ends in a hard consonant, the root vowel will be -o-, if it ends in a soft, it will be -e-.

4.3.4.1. Navezak

The doubletism we are interested in when it comes to this declension pattern is indicated with brackets in Table 15. Namely, adjectives and pronouns in Croatian exhibit a feature called *navezak*, which is an optional vowel attached to the end of Gsg, DLsg, Isg and DLIpl forms.[112] The vowel itself is semantically empty as it carries no grammatical meaning. This feature is quite unique for Croatian in comparison to other Slavonic languages; the only other example of a similar variation that we were able to find across the whole Slavonic world appears in the instrumental singular of Russian feminine adjectives (-ой, -ей / -ою, -ею); however, according to our Russian informants, the latter forms have been losing ground for a couple of centuries now and are by this point relegated to poetry, where they can be used to force rhymes and make up an extra syllable in meter where needed. In Croatian, on the other hand, *navezak* is still very frequent.

Recent research on old Croatian literature has shown that up to mid-19th century there was a differentiation in (grammatical) meaning between the Dative and Locative forms – in Dsg writers used only the long forms -omu/-emu, whereas in Lsg

[112] Ćorić (2007) calls this type of variation allomorphic doubletism as it occurs on a sub-lexical level. In other words, it is the morpheme -eg/og that can take either -ø or -a, not the whole word that takes -eg/og or -ega/oga.

they used the short -om/-em exclusively (see Ham (1996)).[113] However, since Vuk Karadžić used these endings more or less freely without such a distinction in meaning in his idiom, Maretić (1899) introduced free variation. Some authors (VHG, 2007) argue that this distinction should be retained in present-day language, whereas others (HJS (1999), Matasović (2008)) argue for the use of -ø and -e in Lsg and -u in Dsg.

Various normative manuals propose different principles to delimit the usage of these short and long forms. For instance, Alerić & Gazdić-Alerić (2013, p. 15) recommend the use of the long form in the following situations: a) when the adjective is used as the head of the noun phrase, b) in inversion (adjective following the noun), c) to prevent the co-occurrence of two similar phonemes at word border (e.g. *novo**g** **g**rada* > *novoga grada, staro**m** **m**jestu* > *staromu/e mjestu*), and d) only on the first adjective in situations where several adjectives are used as attributes of a noun. Whereas principles a)-c) are more or less accepted in practice, the last one is not, as Table 16 below shows.[114]

Table 16. Frequency of various combinations of navezak in the phrase hrvatski književni jezik 'Croatian literary language' (Source: HJR)

Case	Form	N
Gsg	hrvatskog**a** književnog (jezika)	132
	hrvatskog književnog (jezika)	69
	hrvatskog**a** književnog**a** (jezika)	18
	hrvatskog književnog**a** (jezika)	1
DLsg	hrvatskom književnom (jeziku)	56
	hrvatsko**m**e književnom (jeziku)	48
	hrvatsko**m**u književnom (jeziku)	2
	hrvatsko**m**e književno**m**e (jeziku)	2
	hrvatsko**m**u književno**m**u (jeziku)	1
	hrvatskom književno**m**e (jeziku)	0
	hrvatskom književno**m**u (jeziku)	0
	hrvatsko**m**e književno**m**u (jeziku)	0
	hrvatsko**m**u književno**m**e (jeziku)	0

[113] At this point there is still no attestation of -ome as the alternative long variant, however.

[114] For instance, one of the manuals that also argues for such use, VHG (2007) breaks this principle in the book's very title: *Glasovi i oblici hrvatskog**a** književnog**a** jezika* 'Sounds and Forms of Croatian Literary Language' (if this principle was closely followed, it should be *hrvatskog**a** književnog*).

We also performed a search in HJR for a phrase which uses only one attribute (*hrvatski jezik*) and we got the following distribution: in Gsg the two forms were almost equally distributed (618 *hrvatskog jezika* / 675 *hrvatskoga jezika*); in DLsg the short form was much more frequent than the two long forms (788 *hrvatskom jeziku* / 177 *hrvatskome jeziku* / 34 *hrvatskomu jeziku*). Only the short form was used in both cases, whereas the two long forms were used exclusively in their respective cases (-e in Lsg and -u in Dsg).

However, it seems that the two variants carry a stylistic difference. Težak's (1984) analysis consisted in listening to radio news, where he noticed a predominance of short forms (90%). On the other hand, in written (i.e. edited) texts the author analysed long forms appeared more, although short forms were still dominant. The author believes this is so because editors prefer the long form as it is in general considered a feature of higher style.

Pennington (2013) argues that the use of *navezak* (which he calls ALFA, Adjectival Long Form Allomorphy) in adjectives and relative pronouns mimics the usage of accented and unaccented forms of personal pronouns, as per Wackernagel's Law (i.e. long forms would be used when preceded by a preposition). This notion of morphology mirroring syntax has been termed the *Mirror Principle* by Baker (1985). Pennington's corpus analysis of HNK does show some indication of this to be the case, but only for monosyllabic pronominal and adjectival roots. For instance, the combination [preposition + *toga*] appears almost four times as often as [preposition + *tog*], whereas [preposition + *ovog*] and [preposition + *ovoga*] appear with roughly equal frequencies. However, after replicating some of the corpus queries performed by Pennington, we noticed several errors in the frequency data the author presents, hence his results should not be taken as conclusive, especially since no statistical tests were included in the analyses. The author also distributed a sociolinguistic questionnaire among speakers of all three variants of BCS and noted that ALFAs are used in Croatia at a much higher rate than in Serbia or Bosnia; moreover, Serbians who have been long-term residents of Croatia consider them 'Croatianisms' and reject their use.

Maretić (1899) declares that the form *mnom* as the instrumental singular of the 1st person pronoun is "slightly more customary" than *mnome* (Croatian original: 'malo

običniji'). However, as pointed out by Anić (2009, p. 635), all subsequent grammars copied this observation without asking the question what Maretić actually meant by the term 'slightly more customary'. The majority of grammarians argue that the doubletism in Isg of personal pronouns is not an example of free variation but is rather syntactically conditioned, where long forms such as *mnome*, *njime, time* should be used in the ungoverned position and *mnom*, *njim, tim* only when governed by a preposition. Our corpus search found numerous counterexamples of this tendency.

Table 17. Frequency of doublet forms in Isg of personal pronouns (Source: HJR)

Nsg form ('gloss')	**Isg form**	**N (+ preposition)**	**N (- preposition)**	**Total N**
ja ('I')	mnom	3,878	8	3,886
	mnome	8	33	41
on ('he')	njim	11,580	212	11,792
	njime	2,619	2,774	5,393
ona ('she')	njom	5,466	366	5,832
	njome	744	1,594	2,338

According to the same grammarians, the same principle should apply in DLIpl. However, in this case it is more closely followed. For instance, of the 170 tokens of the form *našima* 'ours' in HJR, 161 appear as either NP heads or in inversion.

4.3.4.2. Doubletism in possessive pronouns

Possessive pronouns *moj* 'my', *tvoj* 'your' and the reflexive-possessive pronoun *svoj* 'one's own' exhibit an additional doubletism in Gsg and DLsg. Namely, these three pronouns can attach the inflectional endings to the Nsg form (*moj-eg, moj-em* etc.) or to the shortened stem (*m-og, m-om* etc.). If we also consider the aforementioned phenomenon of *navezak*, we end up with four possible Gsg forms for these pronouns (five in DLsg).

The third person feminine possessive pronoun ('hers') can also have a short (*njen*) and a long form (*njezin*), but in this case they recur through the whole paradigm, not only in Gsg and DLsg. *Njen* was introduced into the grammar by Maretić (1899) and was first attested in Karadžić's writing. The majority of normative manuals acknowledge a stylistic difference between the two forms – the long form is considered

neutral and the short conversational. The frequency distributions of all abovementioned pronouns are given in Table 18 below.

Table 18. Frequency of possessive pronouns with stem doubletism (Source: HJR)

Case	Pronoun	N (short stem)	N (long stem)
Gsg	moj	4,835	763
	tvoj	1,291	173
	svoj	30,193	4,213
DLsg	moj	3,273	983
	tvoj	833	175
	svoj	25,510	3,996
all cases	ona	22,003 (*njen*)	43,090 (*njezin*)

4.3.4.3. Isolated cases of doubletism in pronominal-adjectival declension

The quantifying pronoun *sav* 'all' shows doubletism in two of its forms – the neuter form can be either *sve* or *svo*. Once again we encounter the problem of diachronic versus synchronic motivation for the existence of a form. The former variant is derived from an older form *vse*. Since /s/ is a soft consonant, -e was attached by analogy to all neuter nouns that end in soft consonants. However, throughout its history, this form underwent a metathesis of consonants and now ends in /v/, a hard consonant, hence from a synchronic perspective, -o would be the natural ending.

This pronoun also has doublet forms in Gpl, namely *svih* and *sviju*. The former is analogous to the Gpl form of all other adjectives and pronouns, whereas the latter is the remnant of the old dual,[115] which we already came across in some feminine nouns. Even though none of the normative manuals attempt to make a distinction between the two, it seems *sviju* is slowly retreating from the scene and becoming archaic.

[115] This ending also appears in Gpl of numbers 2-4 (*dviju*, …) and the quantifier *oba* 'both' (*obiju*).

Table 19. Frequency of doublet forms of the quantifier sav (Source: HJR)

Form	Gram. meaning	N
svih sviju	Gpl	38,670 985
sve svo	neuter	> 5,750[116] 394

4.3.5. Adjective comparison

Croatian speakers have several endings at their disposal to create the comparatives and superlatives of adjectives. The most straightforward of these is -ši, which is no longer productive in the language and which only appears with a few adjectives,[117] namely *lak* 'easy' > *lakši, lijep* 'beautiful' > *ljepši, mek* 'soft' > *mekši*. These three adjectives take only this ending and are therefore not so interesting for our analysis. There are also some 'irregular', or suppletive forms: *dobar* 'good' > *bolji, zao* 'evil' > *gori, malen* 'small' > *manji* etc., which will also not be the subject of our analysis.

The other two endings in use are -ji and -iji respectively. It is important to note that the former ending triggers some kind of a phonological change (palatalization etc.) in all the words it appears, so it rarely appears in its full form (hence some authors prefer to label it as -i, see Moguš (1970)). As can be seen, the two endings in question differ in the number of syllables. For that reason, in the remainder of the text, we will refer to -ji as the 'short' and to -iji as the 'long' ending.

We will first present the rules found in grammars followed by a presentation of usage data. Most grammarians assign the comparative ending based on the adjective's accent. However, sometimes there seem to be more exceptions to a rule than there are adjectives that actually abide by it.

[116] Since this form also appears in several masculine and feminine cases, it was impossible to determine the actual number of neuter forms. Hence this number includes only the tokens of the most frequent *sve* + neuter noun combinations retrieved using the collocation list function in HJR (e.g. *sve vrijeme* etc.).

[117] However, in the Kajkavian dialect it has been retained to a much greater extent (e.g. *starejši* 'older', *boljši* 'better') and is still productive.

a) Monosyllabic adjectives with a short accent take the long ending

This is a large family of adjectives and most of them strictly follow this rule, e.g. *lȍš* 'bad' > *lošiji, svjȅž* 'fresh' > *svježiji, vjȅšt* 'skillful' > *vještiji, nȍv* 'new' > *noviji*. However, most grammarians mention three to four monosyllabic adjectives from this family which nevertheless take -ji: *dȕg* 'long' > *duži, strȍg* 'strict' > *stroži, tȉh* 'quiet' > *tiši* (sometimes adding *mr̂k* 'glum' > *mrči*). But in actual fact, all of these exceptions appear in the corpus with both forms, with the short form being more frequent (see Table 20 below).

Some authors (Jonke, 1965a) argue against the longer form *strožiji*. However, other authors (Okuka (1982), Gudkov (1988)) argue in favour of its grammaticality. Although its natural comparative form, based on its prosodic structure, should be long, it became an exception due to sharing phonological characteristics with *dug*. It is questionable how *mrk* became a member of this class since diachronic data show that the 'regular' longer form started to be used earlier. It seems this adjective has never fitted very well into the class of exceptions.

All in all, we can see that -iji slowly started penetrating the class of exceptions (*tih* is the only adjective unaffected by it). This could signal a tendency for the paradigm to be levelled out, i.e. the default ending for the whole family is spreading to the smaller class of exceptions.

b) Monosyllabic adjectives with a long accent take the short ending

In the case of this rule, exceptions are even more numerous: *pûst* 'deserted' > *pustiji, slân* 'salty' > *slaniji, svêt* 'holy' > *svetiji, lijên* 'lazy' > *ljeniji*, etc. Every subsequent grammar introduces new exceptions to this rule, with some of them allowing for doublet forms.

Some authors (Vince (1977), Moguš (1977)) argue that -ji is no longer productive (as is the case with -ši), so any new adjective with the same prosodic structure would actually form the comparative using -iji. Furthermore, their claim is that every adjective in this family will eventually develop an alternative form with -iji, which will, at one point or another, either replace the original short ending or at least establish itself as a secondary form. Based on this, we could conclude that, in theory

at least, every short comparative in this family could have an alternative long form, but not vice versa.

Moguš (1977) also argues that grammarians should permit the usage of both forms if they have both been derived following valid word-formation rules and not try to artificially prescribe the use of one over the other. This leads to absurd situations such as that a minimal pair of adjectives, *gûst* 'thick' and *pûst* 'deserted', are supposed to form comparatives with two different endings (*gušći* but *pustiji*) (also see the frequencies of another minimal pair *tijésan* and *bijésan* in Table 20 below). "The path of language is such that one of the forms will become archaic and slowly fall out of use; *usually the shorter one*" (Moguš (1977, p. 148), translation and italics mine).[118] The author does not explain why he believes that the shorter form should fall out of use, but we assume it is due to its aforementioned unproductivity.

c) Disyllabic adjectives take the long ending (regardless of the accent)

The majority of adjectives in this family follow this rule unexceptionally; however, there is a supplement to this rule, which states that *disyllabic adjectives ending in -ak, -ek, and -ok delete the final syllable and take the short ending.* However, some of these adjectives do not delete the final syllable and either take the long ending exclusively (*gibak* 'elastic' > *gipkiji, pitak* 'drinkable' > *pitkiji*) or have developed doublet forms.

The adjective *blizak* 'close' is a special case in this family as it is the only one that has developed a semantic distinction between the two forms: *bliskiji* is used to express emotional, spiritual closeness, whereas *bliži* is used for physical, concrete closeness. The corpus data confirm this distinction.

Table 20 presents the frequency distributions of the two forms. The numbers represent a sum of the frequencies of all inflected forms of both the comparative and superlative of each adjective.

[118] Croatian original: "Ako je varijantnim oblicima tvorba u skladu s pravilima, mislim da treba dopustiti njihovu dvojaku uporabu. Jezični je hod takav da će jedan od oblika postati arhaičan i polako se izgubiti; *obično onaj koji je kraći*."

Table 20. Frequency of doublet forms of adjective comparatives/superlatives (Source: hrWaC14)

Family	Lexeme ('gloss')	N (short)	N (long)
1σ, short accent	strog ('strict')	12,020	829
	čist[119] ('clean')	6,177	2,840
1σ, long accent	crn ('black')	3,523	644
	bijel ('white')	385	220
	lijen ('lazy')	25	446
	plav ('blue')	55	205
	siv ('grey')	57	64
2σ, -ak#	gladak ('smooth')	540	215
	mrzak ('hateful')	281	216
	gorak ('bitter')	326	24
2σ, -an#	tijesan ('narrow')	1,073	219
	bijesan ('furious')	45	214

4.4. Intermediate summary

This overview of Croatian declension patterns aimed to show how widespread doubletism in Croatian is. We identified instances of doubletism in a number of cells in each of the four nominal declension classes dealt with in this chapter. For some of those cells we tried to determine whether there was any factor conditioning the distribution of the respective forms but we were unable to find one. For instance, in Isg of the a-declension, we did not find an effect of syntactic context (preceding preposition). We noticed stylistic factors determining the use of the short plurals in some masculine nouns of a-declension, but they were absent in many other words of the same class. The use of *navezak*, primarily with adjectives, also seems to be unconditioned.

We have also noticed that the data from grammar books and other reference manuals of Croatian are often not an accurate reflection of actual language use. Some of the rules and principles these manuals define for distinguishing between doublet forms are not followed and cells for which grammarians prescribe only one form often appear with doublet or triplet forms. For instance, none of the Croatian grammars mention doubletism in connection with nouns ending in a soft consonant, but rather

[119] It seems that Croatian is not the only Slavonic language which allows doublet endings for this adjective. In Polish, for instance, both *czystszy* and *czyściejszy* appear.

all prescribe -em. However, as we have seen, doubletism is widespread in this family of nouns. We believe Croatian is in dire need of a grammar book that would be based on synchronic usage data and would be able to minimise such discrepancies.[120]

We were also unable to find evidence of the claim proposed by Kroch (1994) and later authors that it is a natural tendency of language to eliminate doublets and that one of the competing forms will eventually disappear. Although none of the corpora we used are diachronic corpora, we were nevertheless able to determine that the number of lexemes appearing with doublet forms in Croatian seems to be increasing rather than decreasing (cf. the long comparative ending spreading to the domain of the short ending, or increasing de-palatalization in DLsg).

Finally, we are unable to make any conclusions about the potential location of doubletism in the frequency spectrum. For instance, Fehringer (2004) observes that weak preference occurs in low-frequent words, whereas highly frequent words show a strong preference for one of the forms. Thornton (2012), on the other hand, does not find a similar pattern in Italian. The verb *dovere* 'must', for instance, has the highest frequency considering both the cumulative frequencies of the cell-mates in each of the overabundant cells, and the global frequency of the lexeme, but it does not have the highest ratio between the two cellmates in its overabundant cells. Also, *sedere* 'to sit' has a much lower frequency than *dovere*, but a higher ratio between the two cell-mates in each of its overabundant cells. The author concludes that "clearly, overall lexeme or cell frequency is not a factor that can explain reduction or maintenance of overabundance on its own" (Thornton, Reduction and maintenance of overabundance: A case study on Italian verb paradigms, 2012, p. 196). Our data give credit to both authors – highly frequent words in our data generally show stronger preference for one form, but there are numerous exceptions (see for instance the distribution of *putovi/putevi* in Table 6, *njen/njezin* in Table 18 etc.).

We also noticed that highly frequent nouns often act as 'exceptions' (i.e. they favour the ending that other, less frequent members of the same phonological family

[120] One of the earliest such grammars written for English was *Comprehensive Grammar of the English Language* (Quirk, et al. 1985). Other examples are *Collins COBUILD English Grammar* (1990), *Longman Grammar of Spoken and Written English* (Biber, et al. 1999), *Cambridge Grammar of English* (Carter & McCarthy 2006), all of which heavily rely on corpus data.

disfavour), especially in Gpl of e-declension (see Table 12). This apparent "exceptionality" of highly frequent items is a fairly common phenomenon in language (e.g. note that the most frequent verb in any language, *to be*, will typically have suppletive inflection). Dye, Milin, Futrell, Ramscar (2016) account for this phenomenon by using arguments from information theory – high-frequency items tend to have inflections that are highly informative about their inflected form, whereas low-frequency items have generic inflection patterns that are less specifically informative. Since context will often fail to distinguish highly frequent items, a great deal of uncertainty is inevitable and a greater level of uncertainty reduction is called for. Eliminating highly predictable competitors would be "a boon for communicative clarity", as it would improve the predictability of other low predictable items.

The question that often arises in connection with the alternative endings is whether they are allomorphs or two separate morphs. For instance, Marković (2012, p. 120), with regards to the two endings for the instrumental singular of i-declension, -i and -ju, legitimately asks if [*čovjek*-] and [*ljud*-] (singular and plural stems of the noun 'man') are suppletive morphs, why are [-ju] and [-i] not suppletive morphs? "If we cannot use the term suppletive for inflectional morphemes, what is their relationship? Because we are definitely dealing with morphs with identical meaning, or the manifestations of one and the same morpheme {instrumental, singular, i-declension}." The author concludes that the term that best describes the relationship of the various endings is *morphome*, a term introduced by Aronoff (1994), defined as a set of morphemes with the same function.[121] For instance, the morphome {Isg} can be expressed as [i, ju] and the morphome {comparative} as [ji, iji].

[121] Martin Maiden later introduces the *morphomic level* as a whole new level of linguistic structure, "intermediate between and independent of both phonology and syntax" (Maiden, 2004, p. 138).

Chapter 5. Doublets in Croatian – computational approaches

The analysis from Chapter 4 has shown two things: i) that the rules prescribed by normative manuals of Croatian do not accurately capture the variation found in contemporary standard usage and ii) that the distribution of the competing forms in present-day Croatian is very much free and cannot be easily predicted by using intra- or extralinguistic factors. Although it might be argued that ii) is a consequence of i) (i.e. different speakers applying rules from different manuals), an alternative explanation is also possible. Perhaps the reason why rules differ from one grammar to another is because no rules can be reliably applied in this situation, which, in turn, leads to the question of whether a description involving rules actually captures what a Croatian native speaker is doing when making these choices. As pointed out by Wulf (2002, p. 121), "traditional rule-based systems can be helpful in summarising language behaviour, but sometimes offer little in the way of predictive power." Perhaps speakers use a completely different processing mechanism in this case.

The mechanism we argue for can be subsumed under the label *memory-based* or *exemplar-based*. Chandler (2002, p. 54) notes that "highly respected observers of the debate all conclude that exemplar-based models are better supported empirically by the experimental data than are the rule or schema abstraction models." More importantly, they can handle the same range of data that is handled in the dual-processing model (2M, see Section 1.2) by two mechanisms with a single processing mechanism. The basic idea behind such models, as explained by Daelemans & van den Bosch (2010, p. 156), is that "learning and processing are two sides of the same coin. Learning is the storage of examples in memory, and processing is similarity-based reasoning with these stored examples." The main features of such models are summarised by Daelemans (2002, p. 160) as follows:

> (1) there is no all-or-none distinction between regular and irregular cases because no rules are used to determine what is regular or irregular; (2) fuzzy boundaries[122] and leakage between categories

[122] The term *fuzzy boundaries* means that it may not always be clear whether an item (especially a less typical member of a category) is a member of a given category or of a different category. This leads to

occurs; (3) the combination of memory storage and similarity-based reasoning is cognitively simpler than rule discovery and rule processing; (4) they show robustness and adaptability.

Croatian, or Slavonic languages in general, can serve as an ideal test case for such models because, as Mirković, Seidenberg, & Joanisse (2011, p. 650) argue for Serbian, "among the words of the [Serbian] language it is difficult to draw a clear distinction between 'regular' (rule-governed) and 'irregular' forms. From the rule perspective, Serbian looks quite different from English – there is a huge leap in the complexity of the rules and in the range of properties they are conditioned on." Furthermore, the authors describe Serbian inflectional morphology as quasi-regular since it exhibits numerous regularities and subregularities, as well as seeming 'exceptions'.[123] The same statement can be made for Croatian.

In this chapter we will proceed with applying two exemplar-based models to Croatian data to see how successfully they can predict the seemingly unpredictable data we have seen in the previous chapter. The models we will be using are Skousen's Analogical Model (AM) and the Tilburg Memory-Based Learner algorithm (TiMBL). The cases we are modelling are the instrumental singular of the a-declension and the genitive plural of the e-declension. Proponents of an exemplar-based approach argue that grammars produced by such models resemble grammars actually acquired by native speakers more than grammars devised by linguists do.

Section 5.1 gives an overview of various computational models in existence, followed by a description of the two models we use in Section 5.2. The results of the implementation of the models on Croatian data are presented in Section 5.3.

some more typical members of a given category being misinterpreted as belonging to a different one due to sharing features with some of its members, which is termed *leakage*.

[123] Rytting (2002, p. 126) describes the Turkish system of consonant alternations in a similar way: "the exceptional classes are too regular to be listed as exceptions, yet too irregular to be fully described by rules."

5.1. Computational models in existence

All the computational models of language that have been developed so far are in their nature implementations of the classification problem that is well-known in machine learning (ML), in which we want to find an algorithm for classifying a new item into an existing category by extracting information from existing data. In the terminology of ML, classification is considered an instance of *supervised learning*, meaning that the computer is presented with example inputs and their desired outputs, given by a human, and the goal is to correctly map inputs to outputs. In *unsupervised learning*, on the other hand, no labels are given to the learning algorithm, leaving the computer to find structure in the input on its own. The models do so by using a variety of learning principles: similarity/distance metric, probability/likelihood, association rules, decision trees, artificial neural networks, clustering, Bayesian networks, reinforcement learning etc.

One of the earliest attempts of modelling the grammar of a language without the use of any rules was Rumelhart & McClelland's (1986) *connectionist* model of the English past tense (also called Parallel Distributed Processing), in which the mental lexicon is organised as an associative network of connections between the stored entities. The entities in the network are assumed to represent neurons and the connections represent synapses, like in the brain of a human being, thus giving this model a neurological basis.

It is important to note that the connectionist model does not compare individual items, but rather that each item is decomposed into features so that the features of one item are linked to the features of another (Rumelhart and McClelland coded for phonetic features, such as ±liquid, (un)voiced etc.). The more features two items share, the stronger the associative connection between them. The model is first trained by feeding in a number of input and output forms. Every time an item is encountered in the dataset, it becomes activated and this activation later spreads to all other units connected to it. This spreading activation is a unique feature of such models and it is argued to mirror the spread of neural impulses in the brain.

Learning occurs by comparing the output produced by the model with the intended output and adjusting connection weights on the basis of statistical contingencies in the training set. This feeding, comparing and adjusting process is

repeated until the model is able to produce the correct forms of all samples in the training set. It is important to note that the networks are self-organizing, i.e. the researcher does not have to (and cannot) influence the process. This overall learning pattern displayed by the model in the training stages is argued to mirror the sequence children go through in acquiring the English past tense. All in all, the clear link between neural activity and cognition have made connectionism a very appealing model for many authors.

However, Pinker & Prince (1988) criticise Rumelhart & McClelland's model for its overreliance on feedback. As has been mentioned above, connectionist models constantly receive feedback on their actual output so as to be able to adjust their connection weights and proximate the intended output. This, however, is far from being a realistic reflection of the actual acquisitional behaviours of children since a child hardly ever receives feedback concerning their own output. The connectionist model has also been criticised for its rather poor performance in generalising to verbs that it has not been trained on, especially regular verbs. A number of revised connectionist models have been developed in the last two decades trying to remedy this weakness, by adding either further layers of connections or learning algorithms (e.g. MacWhinney & Leinbach (1991), also MacWhinney's (2000) lexicalist connectionist model, in which the lexical item rather than the phonetic feature plays the central controlling role in language learning and processing). All of these models have been able to predict the correct past tense form of both regular and irregular, seen and unseen verbs with equal success, thus showing that it is not necessary to propose separate mechanisms for regulars and irregulars.

Mirković, Seidenberg, & Joanisse (2011, p. 641) find connectionist models difficult to analyse – "they may produce correct output without providing much insight about the underlying mechanisms." Rather, what we, as researchers, want to be able to do is, as explained by Divjak & Arppe (2013, p. 237), "extract linguistically meaningful weights on the contextual explanatory variables the model is operating on", or put more illustratively, "to take a look under the hood". Fortunately, there are numerous other computational models which allow us to do that, some of which we describe below.

The main difference between the various computational models is "in how they compare a new input form to the remembered exemplars, how they select certain candidate examples from the data base to provide the possible basis for an analogical response, and how they choose from among those possible candidate forms one or more alternatives to provide the basis for an actual response" (Chandler, 2010, p. 376). One problem of all these models, as often pointed out by their critics, is that they are unable to produce the actual phonological output of the items in question; rather, what they produce is the relative probability that the alternative outcomes presented to it for a given lexeme are likely to be chosen as the desired form of that lexeme.[124] However, even those probabilities provide us with enough information as they can be used to compare the behaviour of native speakers, which will be the goal of one of the studies in Chapter 6.

The Generalized Context Model (GCM) was developed by Robert Nosofsky in the 1990s. In short, this similarity-based model "categorizes a target item by comparing it feature by feature with the sum of the features of all the training exemplars given to the model. The model then sums the similarity of the target item – the number of shared features – over all the members of a given category of interest and divides that value by the sum of the item's feature-by-feature comparison to all members of all the categories being compared" (Chandler (2010, p. 382). The Minimal Generalization Learner (MGL), developed by Adam Albright and Bruce Hayes, is "an algorithm that iterates over pairs of words in the lexicon, hypothesizing generalizations conservatively on the basis of any phonological features that are shared across the words. A rule is scored according to how many items it applies to in the lexicon, weighted against cases in which the inferred phonological context is present but the rule fails to apply" (Rácz, et al. (2014, p. 1)). Notice that both of these models contain the word 'generalization' in their names in one form or another, which means that they presuppose some kind of abstraction. Rácz, et al. (2014) compare these two models and show that GCM does a better job of predicting variation across

[124] Scholars working with the Analogical Model are finding ways to correct this drawback. Chandler (2009), for instance, created an algorithm to be used for English past tense prediction that would actually produce the phonological form based on the segments of the target form that were deleted when determining supracontexts.

items than MGL. Albright & Hayes (2003), on the other hand, argue that the MGL outperforms the GCM in predicting participant behaviour.

The most recent widely-used computational model is the Naïve Discriminative Learner model (NDL), developed by Harald Baayen and collaborators. NDL is based on the Rescorla-Wagner equations. The basic idea behind this model is that animals (including humans) learn in a cue-outcome fashion. However, this learning in the model takes place not on the basis of isolated words but on the basis of word n-grams. "The association strength of a cue to an outcome is strengthened when cue and outcome co-occur. The association strength is decreased whenever the cue occurs without the outcome being present" (Baayen, et al. (2011). This work has showed that a simple naïve discrimination network can account for a wide range of empirical findings in the literature on morphological processing. Furthermore, a discrimination network was used to model the dative alternation in English (Baayen (2011)) and rival forms in Russian (Baayen, et al. (2013)), performing with accuracy on a par with that of other well-established classifiers, such as generalized linear mixed model, logistic regression, or tree & forest model.

5.2. Computational models used in present research

Milin, Keuleers, & Filipović Đurđević (2011, p. 69) argue that "it is perfectly possible to successfully model the same phenomenon using different machine learning approaches. What is important is the contribution that different approaches give to our understanding of the phenomenon." Even though we assume, based on previous comparisons of the various models described above, that they could all model the cases we are analysing with more or less equal accuracy, in the present research we decided to use only two exemplar-based models – Skousen's Analogical Model (AM) and the Tilburg Memory-Based Learner algorithm (TiMBL). Because they do not involve any kind of abstraction and have raw data at disposition at any time, these memory-based models are often also called 'lazy' learning models, as opposed to 'eager' learning models, such as rule induction, decision trees, statistical models or artificial neural networks, which, once they have extracted a generalisation, forget about the data (Daelemans & van den Bosch, 2010).

The reason why we chose those two particular models has to do with the ability to 'take a look under the hood' – both of them provide much more information about the decision-making process in addition to the mere output. Furthermore, only original data are used to make predictions rather than some intermediary system of relationships, or higher-level generalizations. As argued by Theijssen, et al. (2013, p. 231), "when storing all experience with [an] alternation, there is no reason to abstract away from the original input that we hear by defining higher-level features. This makes the role of the higher-level features used in existing research unnecessary and, using Occam's razor, implausible." Furthermore, various authors (see e.g. Kidd, Lieven, & Tomasello (2010)) have shown that children heavily engage in exemplar-based learning during acquisition. The two models are described in the following two sections.

5.2.1. Analogical Modeling

The notion of analogy has played a central role in traditional linguistics. Great 19th-century German linguist Wilhelm von Humboldt claimed that analogy is the central aspect of language (referenced in Itkonen (2005)). In more recent years, Hofstadter (2001, p. 537) claimed that "analogy is the lifeblood of human thinking." Analogy was traditionally defined as structural similarity and was considered to represent the basis for forming new words.[125] Generativism, on the other hand, has been known for its hostility towards analogy. Chomsky (1957) claimed that there was no discovery procedure for grammars, i.e. no method of deriving grammars from linguistic data. Since any such method has to be analogical (or inductive) in character, it followed that there was no use for analogy in linguistics. However, as pointed out by Chapman & Skousen (2005, p. 334), even in traditional descriptions analogy appeared too unsystematic. "Traditional accounts lack constraints on the operation of analogy, so that there is no principled means of telling when or where analogy will operate."

[125] For instance, some examples of formation by analogy are provided by Derwing & Skousen (1989, p. 48): "if *four* aeroplanes can be said to fly in *formation*, then *two* must fly in *twomation*; if it can be *too* hot in one place, why not *three* hot in another?"

Hence in 1980s Royal Skousen developed the Analogical Model, which tried to precisely define when, where, and how analogy would operate. As the name suggests, this model finds analogies between various elements of the dataset. Each occurrence in the dataset is specified in terms of a set of *variables* and an assigned *outcome*. The model goes through the database of given exemplars and groups them into supra- and subcontexts by virtue of their shared features. A *supracontext* is a more general representation of an exemplar – a supracontext always has one undefined variable more than all of its corresponding *subcontexts*. For instance, in a set of three variables {A}, {B}, and {C}, the different subcontexts would be {A, B, C}, {?, B, C}, {A, ?, C}, {A, B, ?}, {?, ?, C}, {?, B, ?}, {A, ?, ?}, and {?, ?, ?}, where '?' stands for an undefined variable. The supracontext {A, ?, ?} would include the subcontexts {A, B, C}, {A, B, ?}, {A, ?, C} and {A, ?, ?}. The model then tries to predict the outcome for a new item based on previous exemplars by comparing these contexts. Each item that appears in the database has an equal probability of being included into a subcontext, no item is pre-excluded; however, when it comes to the actual analysis, not all items can be included. This is the main difference between AM and traditional analogy, in which virtually any occurrence can serve as an analogical source. The probability that an exemplar will be used in the comparing process depends on three properties:

> (i) proximity: Database items that share more features with the given form will appear in more supracontexts and will therefore have a greater chance of being used as an analogical model;
>
> (ii) gang effect: When a group of similar examples behave alike, then the probability of selecting one of these examples as an analogical model is substantially increased;
>
> (iii) heterogeneity: An example cannot be selected as the analogical model if there are other examples more similar to the given form which have different behavior (Skousen, 1989, p. 4).

Heterogeneity is determined by a system of what Skousen defined as *pointers* that go from each occurrence in the dataset to every other occurrence, including itself. For each pointer, the algorithm determines whether there is a change in the outcome or not (*agreement* or *disagreement*). As soon as there is disagreement between a supracontext

and one of its subcontexts, that supracontext is labelled as *heterogeneous*. Only *homogeneous* supracontexts, i.e. ones where all the subcontexts behave identically, are included in the analysis.[126] So distance (unlike in MBL, see below) is not the crucial parameter here; the most similar examples are not the only ones able to have an analogical effect – exemplars that are quite dissimilar to the test item can also have a substantial effect, provided that they all behave in a uniform manner. Once all the homogeneous supracontexts have been identified, the exemplars that they contain form the analogical set based on which the outcome of the test item is determined. This can be done in one of two ways: either the outcome that appears most frequently in the analogical set is adopted for the test item (*selection by plurality*) or an outcome is determined by *random selection*. As pointed out by Chandler (2002, p. 76), "these alternative decision rules permit a model to replicate the probabilistic behaviour characteristic of human subjects" since real people do not always respond to the same stimulus in the same way.[127]

One thing that is important to note is how AM views frequency effects. As Chandler (2010, p. 386) says, "the model does not need to posit some sort of usage counter for keeping track of usage frequency. Instead, it simply assumes that tokens of usage are recorded in memory and that frequency effects fall out as a natural consequence of the number of tokens recorded in memory." The only time that AM users talk about frequency effects is when they refer to an item appearing in more than one homogeneous supracontext, but the frequency of the item itself in the dataset would rarely be more than one. However, we believe this is not a valid assumption, especially for situations that we are dealing with, i.e. morphological doublets. One of our hypothesis in the present research is that the versions of the models which replicate

[126] Skousen defines two types of homogeneous supracontexts: deterministic, in which a single outcome appears in all the subcontexts and non-deterministic, where there are several possible outcomes, but occurring within the same subcontext – a combination of a deterministic and a non-deterministic subcontext automatically makes the supracontext heterogeneous.

[127] Chandler (2002, p. 79) also proposes an introduction of an additional decision rule. Based on previous research showing that sometimes subjects respond on the basis of item differences rather than item similarities (especially in larger datasets where individual items are less distinguishable in memory), Chandler proposes the criterion of judging an item as **not** belonging to the category represented most commonly in the analogical set.

the relative proportional distribution of the two forms in competition from natural language will be more accurate than models where each item appears only once.

5.2.2. Tilburg Memory-Based Learner (TiMBL)

The basic principle behind TiMBL is the so-called nearest-neighbour principle. In other words, the model goes through the list of exemplars and variables attached to each and tries to find an exemplar that is least different from the test item. The original nearest-neighbour algorithm was developed by Aha, Kibler, & Albert (1991) for use in machine learning. The implementation of this algorithm for linguistic data was developed by Walter Daelemans and associates at the Universities of Tilburg and Antwerp in the early 1990s, hence its name Tilburg Memory-Based Learner (for a description of a recent version of the model, see Daelemans, et al. (2007)).

TiMBL inherited some of the terms it uses from information theory, where concepts such as *informativity* and *entropy* are widely used.[128] In the learning stage of the model, the algorithm first determines the behaviour of all the features (variables) in the dataset. The term *Information Gain* (IG) refers to how much information a certain feature contributes to our knowledge of the correct outcome in isolation. "The information gain of a feature is measured by computing the difference in uncertainty (i.e. entropy) between the situations without and with knowledge of the value of that feature" (Daelemans, van den Bosch, & Zavrel, 1999, p. 25). An alternative option is *Gain Ratio*, which "normalizes information gain for features with different number of values" (Daelemans, et al. (2007, p. 22)). This is one of the major differences between TiMBL and AM, as the latter does not try to determine *a priori* which variables are 'significant' and which are not. Feature weighting is used in TiMBL so as to avoid the 'exponential explosion' of all possibilities, which has for a long time been a problem in AM. As Skousen (2002, p. 45) explains, "since the model tests every combination of variables, adding each new variable increases the memory requirements and running time of the program by double." The possibility of automatically determining the relevance of features in TiMBL implies, according to Daelemans, van den Bosch,

[128] In short, information is defined by Shannon (1948) as a measure of one's freedom of choice when one selects a message and is calculated as the inverse of probability of a certain message, whereas entropy is the expected value of the information encoded in each message.

& Zavrel (1999, p. 14), that many different and possibly irrelevant features can be added to the feature set. Since the researcher cannot be sure at this stage what feature is likely to be more or less relevant, this possibility insures them against *a priori* exclusion of features.

The TiMBL model determines the similarity between the test item and each of the items in the dataset using a *distance metric*. The simplest measure of distance is the *Overlap Metric*, which is the sum of differences between the features. The Overlap Metric is an all-or-nothing measure, meaning that all the non-matching values receive the same value – zero. However, the values of each feature can also be classified in a gradient manner so that some value pairs are regarded as more similar to each other than other pairs (*Modified Value Distance Metric – MVDM*), which reduces the possibility of finding equally distant exemplars**.** *Levenshtein distance* counts the number of insertions, deletions and substitutions necessary to transform one item into the other (for other variations of the algorithm, such as exemplar weighting, tree-based indexing etc. see Daelemans, et al. (2007)).

The final element of this memory-based model is the *decision function*. The default setting in TiMBL for the nearest neighbour distance (*k*) is 1, meaning that the model makes a decision by looking at the items with only one value different (and considers all the other ones as equally irrelevant). However, this is not necessarily the optimal setting. In fact, some authors have found that the model achieves the highest accuracy rate with higher *k* values (e.g. in Keuleers & Daelemans (2007) accuracy was best at k = 3 when the MVDM metric was used, whereas Milin, Keuleers, & Filipović Đurđević (2011) achieve the best performance using a neighbourhood of seven exemplars). When k > 1, the researcher can also define a *decay function*, which determines the individual influences of neighbours at different levels. For instance, *Zero Decay* means that all neighbours have the same influence on the classification; with *Inverse Distance Decay*, neighbours are weighted by the inverse of their distance, whereas with *Exponential Decay* the more distant a neighbour is from the target exemplar, its influence on the classification of that exemplar decreases exponentially. It is up to the researcher to decide which of a variety of algorithms described above they want to use. It is important to note that AM does not offer so much flexibility in terms of parameter setting – we have seen that practically the only parameter that an

AM researcher can set is whether the outcome will be selected randomly or by plurality.

5.2.3. AM and TiMBL in practice

Both of the models just described allow for multiple possible outcomes. Each outcome has a probability associated with it, derived from the number of different exemplars in the dataset which use that outcome. In Skousen's (1989, p. 84) view, this is equivalent to real speakers changing their mind – the speaker may try one outcome, decide against it, and then try another. Most rule-based approaches do not allow for such multiple guesses.

Now that we have seen how both models work and the similarities and differences between then, can we conclude which of them is the best? Or better at least? It is hard to evaluate them in such terms. There have been numerous pieces of research comparing the outputs of the two. A study by Daelemans, Gillis, & Durieux (1997) found that the results obtained by AM were statistically superior to the ones obtained by TiMBL's default model. However, when varying degrees of noise were added, both models performed equally well (or poorly). When feature weightings were manipulated, TiMBL's performance became equal to AM. The authors suggest that AM is better at learning regularities, whereas TiMBL is better at putting to use the predictive power of (small) subregularities. Eddington's (2002) modelling of Spanish stress assignment has shown TiMBL to be more consistent in predicting antepenultimate stress and AM more adept at predicting penultimate stress. Daelemans & van den Bosch (2010, p. 160) conclude that "although algorithmically very different from and more costly than MBL (which is linear in the number of features whereas AM is exponential), empirical comparisons have never shown important accuracy or output differences between AM and MBL."

Two questions that have been recurrent in the debates about these two models are: (1) how large the datasets need to be in order for the model to make the correct predictions and (2) how many and what kinds of variables should be defined. When it comes to the first question, the consensus is usually that "larger datasets appear to result in better outcomes" (Eddington, 2004, p. 850). Some of the works discussed here used lexical databases with more than 20,000 lexemes (e.g. Nakisa & Hahn

(1996), Krott, Baayen, & Schreuder (2001)). More importantly, as Daelemans, van den Bosch, & Zavrel (1999) and Keuleers & Daelemans (2007) demonstrated, we should not try to edit the data by eliminating noise, low-frequent, or exceptional items from the dataset as that can be more harmful than beneficial in the long run because in reality it is very difficult to discriminate between noise on the one hand and productive exceptions and sub-regularities on the other.

When it comes to the second question, in the early days of both models the number of variables that could be handled by the model was limited by the memory capacity of the computer. However, with modern computers, this problem has been resolved and, in theory at least, there should be no limit as to the number of features and levels of each feature. But more important than this is the question of what kind of features should be defined. Chapman & Skousen (2005, p. 345) emphasise that "what must be avoided is to select only variables that we think are crucial to predicting the outcome." The researcher needs to be careful not to give the model too much language-specific information, thus enabling it to learn more easily. Most researchers use variables at the syllable level – onset, nucleus and coda of either the whole word or the last few syllables. They are generally represented phonologically; going down to a phonetic level (e.g. representing the distinctive features) would generally be considered too much language-specific information and tends to be avoided.[129] Eddington (2004) tried several methods of coding and alignment, but the end results did not significantly differ. The best alignment was obtained when all syllabic constituents were represented and when syllable boundaries were respected, which, in Eddington's opinion, "may indicate that words are encoded syllabically in the mental lexicon" (2004, p. 867). Keuleers & Daelemans' (2007) model had the poorest performance level when features of only the final syllable were coded.

Also, the question arises as to whether variables of different types should appear side by side. And if they do, should they be weighted equally? Keuleers, et al. (2007, p. 286) argue that "similarity is not determined by phonology alone" and find

[129] However, note that Wulf (2002, p. 116) includes information on whether the vowel is front or back, "in order to be faithful to the theory of AM."

it necessary to include orthographic information[130] into their TiMBL simulations. In accordance with their hypothesis, such a model performed better on the inflection of borrowings. Wulf (2002) uses information on whether the vowel is front or back, whereas Rytting (2002) uses the etymology of the word. Elzinga's (2006) analysis used 15 phonological and 2 morphological variables and the morphological ones did not have a significant effect. Elzinga wonders whether the reason for this was because the phonological variables were overwhelming the morphological ones. However, repeating the morphological variables several times in order to weight them more heavily did not improve Elzinga's models. On the other hand, Eddington (2002) also included certain variables several times, which significantly reduced the number of errors. Skousen (2002, p. 40) thinks that this method "may be helpful from a heuristic point of view, although it cannot be correct in principle, at least for variables of the same type." However, the author also admits that he is unclear on how weights of variables from different classes should be compared. We have seen that this 'artificial' weighting of variables is unnecessary in TiMBL as the algorithm weighs variables by itself by using Information Gain measures.

One of the things Eddington (2004) is concerned with is the relationship between token and type frequency. Bybee (1985), (2001) already established that type frequency, not token frequency was a determining factor for generalization. Eddington creates two datasets – one called the type dataset, in which every lexeme appears only once and another called the token dataset, in which the number of repetitions of each lexeme matches its frequency in a frequency dictionary. In the end, the analysis of Spanish stress assignment using types produced more accurate outcomes than the analysis using tokens. Eddington's previous simulations of Italian conjugation classes and Spanish gender assignment have demonstrated similar effects. Similar to this, Albright (2009) used GCM and MGL, and both of them performed worse when using tokens. Our present experiment also makes use of type and token models, but defined in a slightly different manner than Eddington.

The most common method used in MBL testing is *leave-one-out* in which each item in the exemplar set is temporarily excluded from it and used as the test item. This

[130] Orthographic information refers to whether a borrowed word is spelled as it would be in its language of origin or adapted to the Dutch spelling system.

method is often used to determine "analogical consistency of the dataset, that is, the degree to which each item falls within a gang of similar items with similar behaviors" (Eddington, 2004, p. 852). Other authors (e.g. Daelemans, van den Bosch, & Zavrel (1999)) use the *cross-validation* method, in which the dataset is split into N blocks and the model is run N times. In each run a different block is used as the test set and the rest of the data are used as the exemplar set. Both of these methods prevent the same item appearing in both the exemplar and test dataset. Similar to this is the *block testing* method (Wulf, 2002), in which the dataset is also split into N blocks of equal size – the exemplar set is gradually increased by one block in each run of the model, whereas the unused blocks are being used as the test set.

5.3. Applying exemplar-based models to Croatian

All the memory-based models mentioned above have, in the great majority of cases, been applied to situations in various languages where producing a form is not a straightforward matter, i.e. situations which abound in exceptions and idiosyncratic items, where making explicit rules is sometimes impossible.[131] For instance, Wulf (2002, p. 112) classifies traditional descriptions of German plurals as little more than mere taxonomies, whereby nouns are assigned to classes that take a particular ending. "Such analyses do not reliably explain why a particular noun might fall into one class or another. Therefore, there is no ability to predict the plural form given a novel word without first being told the class of that word (which is tantamount to being told the plural anyway)." By using MBL models, then, we could actually predict an inflected form of a new word without being given any kind of paradigmatic or other information (i.e. class, word origin etc.).

However, all the previous implementations of these models had another thing in common as well: there was only one "expected" outcome for any individual lexeme. What they failed to account for were situations where there is more than one possible outcome, such as in situations of morphological doubletism. In those cases we are not

[131] For example, Skousen's original (1989) work on Finnish plurals, Nakisa & Hahn (1996) and Wulf (2002) on German plurals, Eddington (2002) on Spanish stress assignment, Krott, Baayen, & Schreuder (2001) on the linking element in Dutch compounds, Rytting (2002) on the /k/ ~ ø alternation in Turkish etc.

interested in the actual outcome (as both are possible), but rather in the relationship of the two variants, which is where the probabilities that the models produce (see Section 5.1) play a role. None of the work we have come across so far spent much time discussing doublets. For instance, Keuleers, et al. (2007, p. 294) discarded all Dutch words that had two attested plural items from their analysis as their inclusion "would have needlessly complicated analyses and skewed results." Eddington (2004), when analysing the past tense in English, also came across a couple of examples of verbs which can have two past tense forms (such as *dived/dove*). To account for this, the author included both alternatives into the type dataset, but each of them only once (Rytting (2002) uses similar procedures on his data). One of our main hypotheses throughout this work (revisited in detail in Chapter 6) is that doublets that appear with a 50: 50 (or similar) distribution in natural language will have a different status in the speakers' minds than items with a distribution of e.g. 70: 30 or 95: 5. Therefore we believe that in cases of doubletism, it is necessary to include the information about the relative proportions of the two variants into the model. Hence in our analysis we create three versions of each model, differing in the dataset they use:

a) TYPE model – In this model, each lexeme appears only once in the dataset, paired with its dominant (i.e. more frequent) ending. This ending will also be the target outcome of the models' prediction. We realise it would be pointless trying to predict a form in situations where all forms are possible, hence rather than 'the attested ending', the model is trying to predict 'the dominant ending.'

b) DOUBLE ENTRIES model – similar to Eddington (2004), all the lexemes that appear in the corpus with doublet endings are entered twice into the dataset, once with one outcome and the second time with the other.

c) TOKEN model – In this model, each lexeme appears 100 times in the database; out of those 100 occurrences, the respective outcomes are assigned with the same proportions as they appear in the corpus.

Altogether, six different models are used to predict our data – AM TYPE, AM DOUBLE ENTRIES, AM TOKEN, TiMBL TYPE, TiMBL DOUBLE ENTRIES and TiMBL TOKEN. Hence we will do comparisons on two levels: comparing AM to TiMBL as well as comparing the three respective versions of the models to each other.

As a testing method, we selected a slight modification of Wulf's (2002) block testing. First we established cumulative frequencies (*f*) in the respective cases (Isg and Gpl) of all the lexemes belonging to the respective families (a-declension nouns ending in a soft consonant and e-declension nouns ending in -VCCa). The corpus used to establish those frequencies was hrWaC14. The lexemes were then sorted in descending order of frequency and split into five blocks, based on exponents of 10 (so block A consists of the most frequent items, whose frequency is >1,000, block B is between the frequency values of 100 and 1,000 etc.). In other commonly used methods, datasets tend to be split into equal blocks; in our case, this would leave items with the same frequency in different blocks, so we decided against such arbitrary divisions. In our case, the test dataset is always the full dataset (excluding the word being tested) and the training dataset increases with each run of the model: in the first run, only block A is used as the training set, in the second run, both A and B are used etc.

Prior to performing the final analysis described below, we ran several trial models to determine the optimal combination of parameters. It is important to note that none of these trial models produced considerably different results, which goes to show the strength of the models.[132] As pointed out by Skousen (2002, p. 42), "the only real way to affect the result [in AM] is in the dataset itself, by what occurrences we put in the dataset and how we specify the variables for those occurrences rather than by manipulating the [model] parameters." Since the items in the (full) exemplar set and the test set are the same, in order to prevent the predictions being based solely on the existing items, the item being tested was excluded from every individual analysis. After analysing the trial models, we decided to use selection by plurality as the decision criterion in AM. For the TiMBL analysis, *k* number of neighbours was set at 3 and the distance metric used was MVDM.

[132] Keuleers & Daelemans (2007, p. 171), however, argue against this standard practice as "there is no way of knowing what other outcomes have been predicted by simulations with different parameter settings, nor if the results of the best performing simulation are exceptional considering the results of the unreported simulations." The authors propose an alternative practice in the form of a summary of the results of all simulations instead of reporting only the best performing one.

5.3.1. Applying exemplar-based models to Croatian masculine instrumental singulars

The intricacies of the masculine instrumental singular case were described in Section 4.3.1.2. The reason why we chose to model this particular case was because there had already been several attempts of modelling allomorphy in Isg for a related language, Serbian. For instance, Mirković, Seidenberg, & Joanisse (2011) used a connectionist network to model the production of the whole Serbian case inflection system. Based on a training set of more than 3,000 inflected forms of nouns and using phonological as well as semantic constraints (animacy, abstractness etc.), the error rate for this particular case was 4%. Furthermore, one consistent finding in their analyses was that "the items produced with a morphological stem error all came from groups of nouns with relatively few members, that is, nouns where there are relatively few other nouns behaving in the same way across inflectional forms," which the authors (2011, p. 663) term the *inflectional neighbourhood size effect*. However, the model excluded the possibility of having both suffixes applied to the same stem with different probabilities, and instead implemented a simple rule that attached either -om or -em to a given stem.

Milin, Keuleers, & Filipović Đurđević (2011) model the Serbian instrumental case using TiMBL. In their analysis, the model reached its maximum very rapidly. In other words, "a very small number of exemplars was sufficient for memory-based learning to make a correct analogy and to produce human-like output of suffix allomorphy." There was no significant difference in their output once various combinations of distance (Overlap, Levenshtein, Jeffrey) and decay (Zero, Inverse, Exponential) parameters as well as neighbourhood size (1 to 16) were implemented.

The present analysis is only interested in Croatian masculine nouns ending in a soft consonant as this is the sub-family that was shown to be the hardest to describe in a rule-based approach (see Section 4.3.1.2.b). Our dataset was annotated for 9 variables:

1: number of syllables[133] (3 levels: 1, 2, 3+)

[133] We have seen that the dissimilation principle in Isg depends on the number of syllables – it does not, in theory, apply to nouns with 3 syllables or more. We included this variable to see whether the model will be able to pick up on this generalisation.

2: stress (4 levels: short-descending, short-ascending, long-descending, long-ascending)

3, 4, 5: penultimate onset, vowel, coda

6, 7, 8: ultimate onset, vowel, coda

9: final consonant

Hence the noun *princ* 'prince' is annotated in the dataset as {1, short-descending, 0, 0, 0, pr, i, nc, c} paired with the label for ending (in the TYPE dataset that was -om, in the TOKEN datasets it was -om 63 times and -em 37 times).

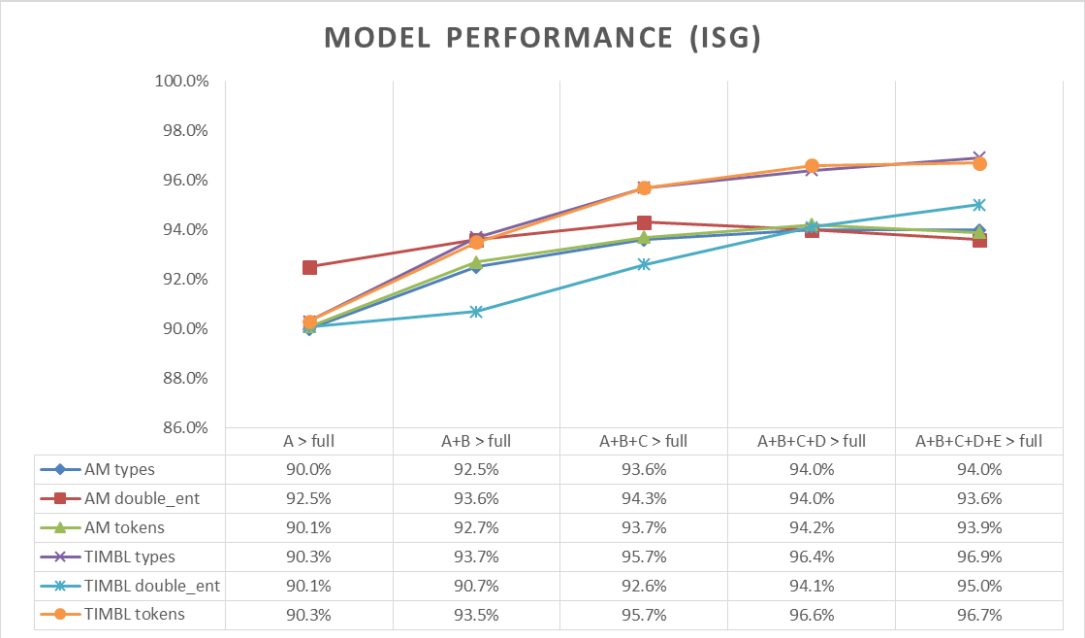
The complete (TYPE) dataset consisted of 2,582 lexemes – 2,326 of those had -em as the dominant ending and 256 had -om.[134] It is important to note that this number includes all masculine nouns ending in a soft consonant that appear in the corpus at least once in Isg, regardless of whether they appear with doublet forms or exclusively use one ending. The division of lexemes into frequency blocks produced the following distribution:

- Block A (f > 1,000): 108 lexemes
- Block B (100 < f < 1,000): 427 lexemes
- Block C (10 < f < 100): 917 lexemes
- Block D (2 < f < 10): 645 lexemes
- Block E (f = 1): 485 lexemes

Figure 1 below presents the results of the modelling. In order to be able to evaluate the success rate of our models, we need to establish the *baseline model*, i.e. a model in which all the items are either assigned the same label or the label is assigned randomly. For instance, if the model predicted all lexemes with the more frequent outcome in the database, -em, that model would still be 90.1% correct (2,326/2,582). The goal of any modelling, then, is to improve on the baseline model.

[134] The DOUBLE ENTRIES dataset consisted of 1,032 extra lines of data, representing lexemes which were attested with both endings.

Figure 1. Correctness rate (%) of the models for Instrumental singular



	A > full	A+B > full	A+B+C > full	A+B+C+D > full	A+B+C+D+E > full
AM types	90.0%	92.5%	93.6%	94.0%	94.0%
AM double_ent	92.5%	93.6%	94.3%	94.0%	93.6%
AM tokens	90.1%	92.7%	93.7%	94.2%	93.9%
TIMBL types	90.3%	93.7%	95.7%	96.4%	96.9%
TIMBL double_ent	90.1%	90.7%	92.6%	94.1%	95.0%
TIMBL tokens	90.3%	93.5%	95.7%	96.6%	96.7%

We can see that Model A performs at baseline, whereas the performance of all other models gradually improves as we increase the dataset. For instance, the improvement rate of the second model (A+B) ranges from 0.67% (for TiMBL DOUBLE ENTRIES) to 3.77% (TiMBL TYPES). Furthermore, the expansion of datasets has a much larger effect on the performance of the TiMBL models, as evidenced by steeper lines (cf. Daelemans (2002)). However, this improvement happens only up to block D. The final expansion of the dataset with *hapax legomena* does not contribute much to the accuracy of the model; in fact, sometimes the model even deteriorates in this block (AM DOUBLE ENTRIES and AM TOKENS).[135] Skousen (2002, p. 37) explains that in AM, "given sufficiently large amounts of data, stability sets in, with the result that adding more examples in the dataset will have little effect on predicting behavior." When compared to the baseline model, the most successful model of all that use the full dataset, TiMBL TYPES, improves performance by 7.55%, whereas even the least successful model in this run, AM DOUBLE ENTRIES, still shows an improvement of 3.89%.

[135] However, note Bod (1995), who asserts that removing all hapaxes degrades performance, which, in the author's opinion, "seems to contradict the fact that probabilities based on sparse data are not reliable" (1995, p. 68).

When it comes to the two comparisons we are most interested in, a first glance at Figure 1 reveals that all three TiMBL models, by the end of the analysis, fare better than the three AM models. With regards to the difference between the three versions of the models, there is no substantial difference between them. We can see that the lines for the TYPE and TOKEN models almost overlap, both for AM and TiMBL. As a reminder, both Eddington's (2004) and Albright's (2009) analysis showed better results when using type frequencies than token frequencies. The lines for the two DOUBLE ENTRIES models appear slightly different than the lines for the other model versions, especially the AM one, which deteriorates after block C. It seems that a modification of the AM model in which the majority of the individual items appears with opposing outcomes paralyses the model. It is also not superfluous to say that block C is the largest block in which the greatest number of doublets appears.

As pointed out by Eddington (2002, p. 150), in any modelling "it is important to ascertain, not only how many errors are made on irregular items, but also the direction of the errors." Reading through the output files to trace back the origin of the errors, we noticed that lexemes with -om as the dominant ending were predicted incorrectly as -em to a much greater extent than the inverse situation. For instance, in the first run of the AM TYPE model (block A as training set) all but one of such lexemes were mispredicted (255/256). Since there are only three items in block A with -om as the dominant ending, this error rate is not surprising. The model cannot be expected to predict an outcome that barely appears in the dataset. As the dataset increases (and the proportion of -om's in it), the misprediction of -om decreases. The misprediction rate of -em as -om, on the other hand, never exceeds 3%.

As a side note, it is interesting to examine which of the variables contributed the most to TiMBL's process of assigning the correct outcome. Feature 7 (ultimate vowel) had the greatest Information Gain value, followed by feature 8 (ultimate coda). This means that the model, similar to what grammarians do, identified that part of the word as being the most significant predictor of ending. The third most informative variable was feature 1 (number of syllables), which is another feature on which grammar rules are based.

5.3.2. Applying exemplar-based models to Croatian feminine genitive plurals

Since the aim of this type of analysis was to give a reliable model of the whole inflectional system of Croatian rather than producing separate models for different cases, our intention was to use the same variables as above in this model as well. This is possible in theory, but some modifications were necessary. For instance, as there are no monosyllabic feminine nouns, different values for the first variable had to be used. Furthermore, whereas the nominative singular of masculine nouns is at the same time also the stem throughout the whole declension, for feminine nouns the stem is an abstraction in the sense that it is not used on its own in any grammatical case. In Nsg, the structure of feminine nouns is of the form -V(C)CCa, where the two final consonants are parts of different syllables. The stem is formed by deleting the final vowel /a/ (see Table 10 for endings added to the stem in all other cases). Our annotation was done on the stem, i.e. up to the final consonant; however, we still considered that consonant the onset of the final syllable, even though there was no vowel to attach it to. Finally, to ensure we still had two full syllables annotated, we also included the antepenultimate syllable in the analysis.

1: number of syllables (values: 2, 3, 4+)
2: stress (same values as in previous section)
3, 4, 5: antepenultimate onset, vowel, coda
6, 7, 8: penultimate onset, vowel, coda
9: final onset

The noun *izložba* 'exhibition', for instance, is annotated as {3, short-ascending, 0, i, z, l, o, ž, b} paired with the label for ending (-ī in the TYPE datasets; in the TOKEN datasets ī was entered 78 times, a-a 20 times and -ā 2 times; in the DOUBLE ENTRIES datasets this lexeme appeared three times).

The final dataset was slightly smaller than in the previous section, containing 1,911 lexemes. However, there is more variation in this case as there are three endings at disposal.[136] 1,709 lexemes appear with -ī as the dominant ending, 101 with -ā added

[136] In the DOUBLE ENTRIES dataset, some lexemes appeared two times, others three times, resulting in 3,053 rows of data.

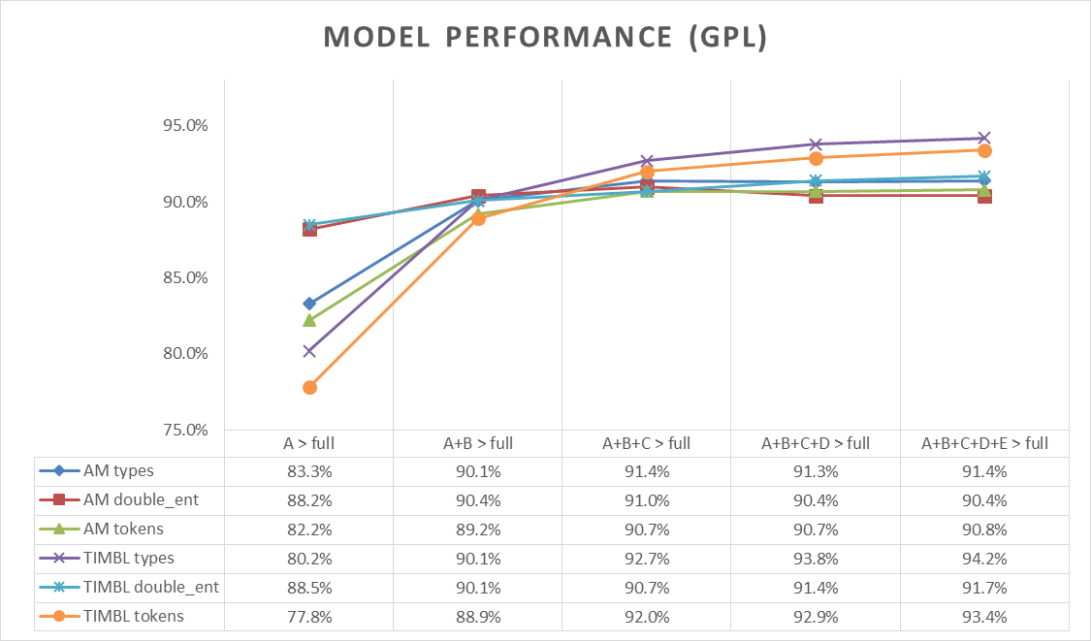
directly to the stem and 101 with the insertion of the fleeting a and ending -ā (labelled a-a).

The distribution of lexemes within blocks was the following:

- Block A (f > 1,000): 164 items
- Block B (100 < f < 1,000): 298 items
- Block C (10 < f < 100): 600 items
- Block D (2 < f < 10): 577 items
- Block E (f = 1): 272 items

The model parameters were the same as in the previous analysis (AM: selection by plurality, TiMBL: k = 3, MVDM). The baseline for the model is set at 89.4% (1,709/1,911). Results of the modelling are presented in Figure 2 below.

Figure 2. Correctness rate (%) of the models for Genitive plural



	A > full	A+B > full	A+B+C > full	A+B+C+D > full	A+B+C+D+E > full
AM types	83.3%	90.1%	91.4%	91.3%	91.4%
AM double_ent	88.2%	90.4%	91.0%	90.4%	90.4%
AM tokens	82.2%	89.2%	90.7%	90.7%	90.8%
TIMBL types	80.2%	90.1%	92.7%	93.8%	94.2%
TIMBL double_ent	88.5%	90.1%	90.7%	91.4%	91.7%
TIMBL tokens	77.8%	88.9%	92.0%	92.9%	93.4%

As in the previous analysis, we are faced with a situation where one ending is favoured by the majority of lexemes in the family (-ī). However, whereas in the previous section the misprediction rate of such an ending (-em) increased as more exemplars of the 'other' ending were added to the dataset, in this case we have the opposite situation. The largest misprediction rate of this ending occurs in the very first run of the model, when only block A is used as the dataset. When we presented the corpus analysis of this case (Section 4.3.2.3), it was noted that the 'exceptions', i.e.

lexemes whose dominant ending is not the same as the dominant ending of other members of the same phonological family tend to be the most frequent items in that family (see Table 12). This is in line with various authors who claim that high-frequency items become autonomous from other words (see contributions in Bybee & Hopper (2001)). In block A there is a roughly even distribution of all three endings, so it is not unexpected that the model predicts all three of them with roughly equal proportions as well. When other, 'normally behaving' items (i.e. those with -ī) are added to the dataset, the misprediction rate of -ī decreases and that of the other two endings increases. When it comes to the two 'other' endings, the misprediction rate of a-a is much greater than that of -ā (e.g. in the final run of the models, a-a is mispredicted in 90% of the lexemes on average). Similar to the previous analysis, we notice that the performance of all the models somehow stops increasing after block D. This is in line with Eddington's (2004) finding that mid-frequency items form the best analogical set as well as with the above claim by Skousen (2002) that stability sets in once the dataset reaches a certain size.[137]

We can see that the overall shape of lines in Figure 2 is similar to those in Figure 1 – each new run of the model produces slightly better results. However, the performance of the initial run of the model is significantly poorer here than in the previous section – only the two DOUBLE ENTRIES models approach the 89.4% rate of the baseline model, whereas the others perform much more poorly. This is probably due to the fact that there is one outcome more than there was in the previous analysis, which introduces more fuzziness. The performance of the models improves on the baseline only in the second run of the model (A+B). It is here also that the difference in performance between this model and its previous run is the most substantial. For instance, the performance of TIMBL TOKENS model A+B represents a 14,26% improvement compared to Model A. The most successful final model TiMBL TYPES) performs 5.37% better than the baseline, whereas the performance of the other full models represents an improvement on the baseline ranging from 1,12% (AM DOUBLE ENTRIES) to 4,47% (TiMBL TOKENS). In addition, TiMBL once again outperforms AM and there is no substantial difference between the TYPE and TOKEN models. We can

[137] For instance, in Derwing & Skousen (1994), this stability had already set in with only 25% of the dataset used.

also see that the two DOUBLE ENTRIES model, even though they have the highest performance at the start, are least affected by the subsequent extensions of the dataset.

The feature that had the largest Information Gain value in the TiMBL analysis was feature 8 (penultimate coda), followed by feature 9 (final onset), which shows that the model identifies the final consonant clusters as the main source of information.

5.4. Conclusion

This chapter analysed two Croatian cases using an alternative approach to the traditional rule-based descriptions. We did so by using two exemplar-based models, which make predictions on the spot, without resorting to any kind of generalisations. All models were able to predict the dominant ending with great accuracy (> 90%). Krott, Baayen, & Schreuder (2001, p. 78) note several ways in which a computational model can fail. For instance, it could have predicted the output at chance level (represented by the baseline value in our analyses). The majority of our models, especially the later ones, perform above chance level. This goes to show that the model is not simply mirroring corpus frequencies, but is rather taking other factors (namely similarity) into account as well. As pointed out by Milin, Keuleers, & Filipović Đurđević (2011, p. 16), "making decisions independent of similarity is a bad strategy." Secondly, it could have predicted the correct choice, but for the wrong reasons (e.g. basing its decision on the onset of the penultimate syllable or stress). The fact that the models give most weight to the final segments of the word show their cognitive plausibility as this is also the part of the word on which rule-based approaches condition the choice of the ending. As we will show later in the questionnaire studies, speakers' grammars are also organised in terms of phonological families, based on final consonants and consonant clusters.

All the errors that the model made were in the direction of *regularization,* i.e. hyperproduction of the most frequent ending, which is a positive thing for a model. Previous research in the area of L1 acquisition has shown that children also tend to overregularise highly frequent irregular items rather than vice versa (e.g. *goed* is

present in the early stages of grammar of many children instead of *went*).[138] Furthermore, we should not consider the models' inability to predict exceptional behaviour as a failure of the models. Rather, as Skousen (2002, p. 37) explains, "the reason it misses the exceptions is because the standard language itself fails to use the regular [form]. The real power of the model is predicting fuzziness." And the current models do this well.

Our analysis extended the existing AM and TiMBL models by including the proportional distribution of the two competing endings (TOKEN models). This did not substantially improve the model. Baayen (2010, p. 436) notes that "computational models that account for frequency of occurrence by some mechanism equivalent to a counter in the head run the risk of overestimating the role of frequency as repetition, of overestimating the importance of words' form properties, and of underestimating the importance of contextual learning during past experience in proficient reading." If the claim given at the beginning of this chapter that exemplar-based models provide realistic descriptions of mental grammars is true, based on the present finding one could argue that proportions are hence not included into our mental grammars either. However, a further argument for the opposite claim will be provided in the next chapter.

An issue that is always present when modelling alternations in language is, as pointed out by Guzmán Naranjo (2016, p. 28), that "it is not possible to know beforehand how much variability we should be able to account for with our models, and how much variability it should not be possible to model as it is likely that a degree of variation is just probability matching." In other words, we do not know *a priori* how much freedom speakers actually have when they choose one form or the other, and how much is determined by context. "This means that it is impossible in principle to ever know if the statistical model we chose reached ceiling or if there are other still unknown predictors that, if included, would increase model performance." As Kilgariff (2005, p. 264) and many others have observed: "language is never ever

[138] More precisely, children's acquisition of irregular forms shows a U-shaped pattern (as per Berko (1958)) – they first produce the irregular form (e.g. *went*), then they go through a period when they overgeneralise and produce *goed*, followed by a period when they use the two interchangeably, before finally settling for the original irregular form.

random;" however, Divjak, Dąbrowska, & Arppe (2016, p. 3) extend this claim of Kilgariff's by saying that "it is also rarely, if ever, fully predictable."

All in all, we hope to have shown that models such as these indeed offer a realistic and accurate description of the processes that take place in the speakers' minds when (processing and) producing morphological forms. In one of the sections in the next chapter (Section 6.4), we will further analyse this claim by comparing the output of the models to the results of a questionnaire study conducted among native speakers of Croatian.

Chapter 6. Doublets in Croatian - a pseudo-experimental approach

We mentioned in Section 1.2 that one of the main goals of this work is to show in what way usage data are reflected in speakers' mental grammars – whether speakers keep track of absolute frequencies (tokens), types, relative frequencies or some other frequency measure. In the same section we also examined a number of arguments for each of these measures or their combinations. On the other hand, Fehringer (2011, p. 103) argues that every time a speaker produces a form, he/she is influenced by the *preference ratio* of that form to the other. In the present chapter we test to see whether this claim by Fehringer has a psychological basis. Furthermore, Thornton (2012) uses ratios to determine the strength of overabundance in a cell – the lower the ratio, the stronger overabundance in that cell; the higher the ratio, the weaker overabundance.[139] The questionnaire studies described in this chapter will also examine, among other things, whether speakers similarly demarcate different strengths of overabundance.

In Chapter 2 we have seen that grammaticality has traditionally been defined as a categorical distinction. However, more often than not, native speakers' grammaticality (acceptability) judgments were shown to be of a continuous nature. This was usually explained by invoking performance factors (plausibility, working memory limitations, ambiguities etc.). Sprouse (2008), for instance, shows that temporary syntactic ambiguity (such as garden-path sentences) can indeed decrease acceptability, whereas temporary semantic implausibility cannot. Fanselow & Frisch (2006, p. 294) show that some processing difficulties may also make a sentence with low grammaticality *more* acceptable, which usually happens in cases when the factor making the structure ungrammatical is difficult to detect (e.g. violations that come later in a sentence). Featherston (2005b, p. 202) believes that such relative judgments reflect computational cost. Longer sentences are systematically judged worse than shorter sentences, even though there is nothing actually 'wrong' with longer sentences – it is just that more words mean more computational load.

[139] Thornton also uses ratios to determine the likelihood for the maintenance/elimination of overabundance in the language. The author's working hypothesis is that a ratio around ≥70: 1 indicates extinction of overabundance and ratios in the range of tens indicate that it is on the verge of extinction.

What the following studies intend to show, among other things, is that the differing levels of (un)acceptability are, more than anything else, due to differences in frequency of the forms in language itself which are reflected in the speakers' mental grammars. This is not to deny the existence of any of the processing effects dealt with above, but we simply add an additional dimension to the explanation.

Throughout this chapter we will re-use Fehringer's concept of preference ratios, albeit calculated in a slightly different manner – the term we will use, especially in the final two studies, is *relative proportions,* which are expressed in the form of percentages. We have already come across this concept in the previous chapter, when we developed our TOKEN models. Models using this information performed slightly better in places than models using just the majority ending (TYPE models), but this improvement was not substantial. In this chapter we try to determine whether the differences in relative proportions will also be reflected in the reaction of native speakers to doublet pairs. Our assumption is that doublet forms that appear in the language with a 50: 50 distribution will not have the same status in speakers' mental grammars as those that appear, for example, with a 70: 30 distribution nor as those with a 99: 1 distribution. On the other hand, native speakers should react similarly to all items that appear with, e.g. a 95: 5 distribution, regardless of whether this proportion reflects a cumulative frequency of 20 or 20,000. The relative proportions, defined in this way, also help us overcome the problem of low frequencies that we have encountered in some instances of doubletism in Chapter 4.

Relative proportions are represented by the concept of a BAND, which is defined as a grouping of items of similar proportions. A similar concept has previously been used by a number of authors, but only as a method of classifying various lexemes, without serving a 'higher purpose', so to say. Bresnan (2007) divided the predictions of her regression model into five probability bins, ranging from very low probability of being a prepositional dative to very high. The bins were delineated quite arbitrarily at 20% intervals. Šimandl (2008, p. 37), in his doctoral dissertation on variation in the Czech nominal system, defines five types of relation between the two variants based on the proportions of the two forms, expressed in percentages: if variant A appears more than 95% of the time, it is labelled as the *monopolistic* variant; if it appears between 60% and 95% of the time, it is labelled as *majority*; if both variants appear

between 40% and 60%, they are *equipollent*; if variant A appears between 5% and 40% it is labelled as *minority*; and if it appears less than 5% of the time it is *marginal*.

In a more recent study, Bermel & Knittl (2012) define seven frequency bands, which are then used as a predictor variable in an acceptability survey for Czech. The bands are defined as follows: when a form appears in less than 1% of examples, it is *isolated*; when its relative proportions are between 1% and 9%, it is *marked*; between 10% and 29% it is the *minority* form; both forms appearing between 30% and 69% are *equipollent*. On the other end of the scale, a form with 70-90% presence is *majority*, the one with 91-99% is *unmarked*, and the one that appears over 99% is *dominant*. This work already showed a strong correlation between the relative proportions of forms in a corpus and their acceptability to native speakers, but it also refuted the possibility of percentages from the corpus being mapped proportionally onto degrees of acceptability. Rather, they concluded that for forms with a proportion above 50%, high levels of acceptability can be predicted with great confidence; however for those below 50% no prediction can be made. Our goal in this study is to explore this relation further.

6.1. Study methodology

Two of the three questionnaire studies described below are acceptability studies, whereas the final one is a forced-choice study. We have already seen in the discussion in Chapter 3 that judging acceptability is not unanimously accepted as a natural or reliable empirical method. As pointed out by Schütze (2005, p. 470), rating (either existing or nonce) inflected forms is a task "for which we still have no theory of what people might be doing." However, that does not make it a pointless task. The same author later argues that acceptability judgments are "themselves data about human behavior and cognition that need to be accounted for; they are not intrinsically less informative than, say, reaction time measures – many linguists would argue that they are more informative" (Schütze, 2011, p. 216). Before proceeding with the description of the individual studies, we extend the discussion from Section 3.1 with a few additional points.

The majority of researchers uses Likert-type rating scales to measure speakers' attitudes, which provide a range of responses to a given statement. Virtually every

author we have come across in the literature has designed their own rating scale, in terms of the number of points the scale contains and the labels assigned to each point. For instance, Culbertson & Gross (2009) use a four-point scale with the points labelled simply 'perfect', 'okay', 'awkward', and 'terrible'. Wasow & Arnold (2004) also use four points – 4 for 'fully acceptable', 3 for 'probably acceptable, but awkward', 2 for 'marginal, at best', and 1 for 'completely unacceptable'. Divjak (2008) asked her participants to rate '*how Polish a sentence sounds*' using five points, including -2 ('unnatural Polish') for sentences "that sound strange and may even be difficult to understand", 0 ('ok Polish') for a sentence "a native speaker could produce although it isn't perfect", and +2 for a "natural Polish sentence." Endresen's (2014) five-point scale combines both acceptability and perceived frequency of a form. Points on her scale are labelled as '*This is an absolutely normal Russian word*' (5), '*This word is normal, but it is rarely used*' (4), '*This word sounds strange, but someone might use it*' (3), '*This word sounds strange and it is unlikely that anyone uses it*' (2), '*This word does not exist in the Russian language*' (1). Andrews' (1990) scale combines acceptability and naturalness in the following labels: '✓' refers to "completely acceptable and natural," '?' is "acceptable, but perhaps somewhat unnatural," '??' is "doubtful, but perhaps acceptable," '?*' is "worse, but not totally unacceptable," '*' is "thoroughly unacceptable," and '**' is "horrible". Divjak, Dąbrowska, & Arppe (2016) use a ten-point scale, where 1 is reserved for an item that sounds 'very strange' and 10 for one that sounds 'completely natural'. The scale used by Bresnan, et al. (2007) in experiments on the English dative alternation was designed in a zero-sum way – the respondents were asked to distribute 100 points between the two alternatives (75-25, 90-10 or any combination thereof).

Whichever scale he/she uses, every researcher is bound to encounter the same problem when performing data analysis – which statistical test to use for testing their hypothesis. Parametric tests are in general more powerful than non-parametric tests. However, in order for parametric tests to be used, several assumptions need to be met (data need to be on an interval or ratio scale,[140] they need to be normally distributed, the number of respondents needs to be sufficiently large etc.). Statisticians have

[140] Kuzon Jr, et al. (quoted in Jamieson (2004, p. 1218)) contend that using parametric analysis for ordinal data is the first of the 'seven deadly sins' of statistical analysis. Authors like Rasinger (2008), on the other hand, consider Likert scales interval and find no fault in applying parametric tests to them.

pointed out that many of the tests using scales such as above do not meet some of these assumptions (even though the researcher often assumes that they do), hence the researcher increases the chances of coming to the wrong conclusions if using inappropriate statistical techniques (Jamieson, 2004).

To overcome this problem, Bard, Robertson, & Sorace (1996) adapt a method widely used in psychophysics – *Magnitude Estimation* (ME) – for use in acceptability rating tasks. This method uses an open-ended, possibly infinite, scale, which the respondents are free to differentiate to as fine-grained level as desired. However, the respondents do not rate the stimuli independently, but rather compare them to the reference stimulus (*the modulus*), which is rated at the start of the experiment. The judgments are proportional – if an item is judged as twice as good than the reference example, it should also receive twice as high a score; if it is half as good, then it should receive half the score. The authors argue that this method captures the gradient nature of grammaticality (as demonstrated by Sorace & Keller (2005))[141] much better than Likert scales. Also, ME allows participants to distinguish as many levels of acceptability as they can perceive, unlike tasks in which they are limited to five or seven choices, as is common practice. According to Sprouse (2011, p. 274), ME has, in many respects, become the "gold standard" in the acceptability-judgment literature today.[142]

Bader & Häussler (2010) compare the results of an ME task and a binary judgment task. Their experiments showed no significant difference between the two methods. However, the ME results had a higher correlation with the corpus results than the binary judgment results did. The authors conclude that binary grammaticality judgments are derived from continuous ones rather than vice versa, hence the latter

[141] In short, Sorace & Keller distinguish between *hard* and *soft constraints* (an Optimality Theory term). Violations of hard constraints cause strong unacceptability and violations of soft constraints are reflected in mild unacceptability. Furthermore, soft constraints are context-dependent whereas hard constraints are immune to context effects. Featherston (2005b) also notices that violations of different constraints affect the judgments in a different way.

[142] Featherston (2008) further modifies the ME method by introducing a second reference point, which is supposed to define the scale in the same way that the 0°C and 100°C points (the freezing and boiling points of water) fix the Celsius scale. The name Featherston gives to such a scale, *Thermometer Judgements*, reflects this analogy.

type are primary (Bader & Häussler, 2010, p. 321). Weskott & Fanselow (2009), (2011) perform similar comparisons of ME to a standard seven-point scale task, using identical experimental material. The authors found that it is not the case that binary or N-scale measures are less informative than ME. In fact, the answers on the seven-point scale led to significantly less unexplained variance, suggesting that it may actually be more precise than ME. Sprouse's (2007) results suggest that subjects are not actually performing the magnitude estimation task, but rather performing a relative rating task in which the reference sentence serves as an upper bound for ungrammatical items, which brings into question the intuitiveness of this method for participants. Weskott & Fanselow (2009, p. 234) argue that even though the question the respondents pose to their grammatical knowledge *'How acceptable is this given n ranks of acceptability?'* is probably much less natural than the plain question *'Is this grammatical or not?'*, the question '*Given the assignment of value i to the modulus α, and the assignment of value j/a to the stimulus y, and the acceptability relation between x and y, what is the value k/a to be assigned to x?'* is even less so. The oddity of the task is also reflected in the fact that a training stage is necessary prior to the actual experiment, which allows the participant to familiarise themselves with the method. No such training stage is necessary for the other commonly used methods. Finally, Sprouse (2011, p. 279) argues that, even though ME might be useful in examining the magnitude of physical stimuli, such as intensity of sound or brightness of light, it is fundamentally incompatible with sentence acceptability. All of these arguments made us doubt its usefulness for our studies.

In our questionnaires we used a seven-point Likert scale as per Bermel & Knittl (2012), ranging from 1 (*totally unacceptable*) to 7 (*totally acceptable*). With interval data, the most important thing to ensure is that the intervals are truly uniform, i.e., that subjects treat the difference between 1 and 2 the same as the difference between 4 and 5. To overcome this, the other mid-points on our scale were left unlabelled. Cowart (1997, p. 71) argues that such a scale description "is meant to invite uniform treatment of the intervals without explicitly discussing the relation between different scale categories." This, in turn, allows us to view this scale as interval instead of ordinal and use parametric tests accordingly.

Another issue which has been especially problematic and widely debated is whether linguists themselves should be included as subjects in any kind of intuition

studies. As we have seen in the discussion in Section 2.1, generativists traditionally used only their own judgments, or judgments from fellow linguists, in their works as it was believed that those who know more about a topic issue more reliable judgments.[143] But, as Labov (1978, p. 199) asserted, "linguists cannot continue to produce theory and data at the same time." A theory of language derived in such a way necessarily describes only the linguist's idiolect and not the whole language. Gries (2002) and Dąbrowska (2010) have both shown that linguists' judgments diverge to a great extent from the judgments of non-linguists and therefore their judgments cannot be considered as representative of the population as a whole. In Gries' (2002) analysis of the English genitive alternation, the linguists failed to predict the importance of several variables that were judged as important based on data from corpora and from naïve speakers. Dąbrowska's (2010) experiments have shown that linguists tend to provide more categorical judgments, while non-linguists tend to use the entire scale of possible ratings. Furthermore, it seems linguists are more lenient in judging linguistic prompts (especially ungrammatical ones) as they come across them more often in their work. On top of this, Gibson & Fedorenko (2013) argue that the theoretical biases of linguists could influence their judgments.

However, Culbertson & Gross (2009, p. 725) point out that linguists are also more reliable, i.e. "among linguists there is a greater tendency for sentences they judge *acceptable* to be *grammatical*." This reliability is not to be confused with consistency, which means that the responses stay constant across different elicitations, regardless of accuracy. Snow & Meijer's (1977) experiments also demonstrated that linguists showed greater agreement with one another than non-linguists. They explain it as the following: either linguists have learned to ignore minor irrelevant differences among sentences (such as their semantic plausibility) or they have learned to apply their theory to unclear cases. Contrary to all the evidence above, Sprouse & Almeida's (2012) survey found only a 2% divergence between the two sets of judgments.

However, none of the works above that showed a difference between judgments of linguists and non-linguists have still not managed to establish the cause

[143] Valian (quoted in Schütze (2011, p. 212)) makes a case in favour of using such 'expert' judgments in psycholinguistic experiments, based on an analogy to wine tasting, which relies on the acquired ability to detect subtle distinctions that inexperienced wine drinkers simply cannot make.

of the differences. Culbertson and Gross's (2009) experiment shows that the main source of respondents' reliability is not expertise in syntax but rather having task-specific knowledge. Subjects who have previously come across similar types of experiments (e.g. in other areas of cognitive science) were much more successful than subjects who have not. Knowledge of linguistics made no difference.

Since our participants were recruited from, among other places, the humanities departments of various Croatian universities, it was expected that a number of students of either Croatian or foreign languages would also respond. However, in neither of their studies was their overall proportion so great as to skew the results in any direction. On the other hand, since all Croatian children go through ten to twelve years of formal training in their native language in school and language debates are very much present in the popular culture, the majority of Croatian speakers might be less linguistically naïve than speakers of English, for instance.

In any type of experiment using human participants, it is virtually impossible to know whether the participants are actually doing what the researcher has asked of them. Plenty of studies in the history of science have been affected by the *observer's paradox* – the tendency of participants to change their normal behaviour knowing which aspect of this behaviour is being studied. In linguistic studies, the observer's paradox would mean that the answers speakers give are not necessarily the ones they would give in unobserved speech. What is even more problematic, as pointed out by Schütze (2005, p. 464), "we do not know if variation within and across experiments is due to variation in the underlying grammatical systems we wish to study or in the task the speakers are carrying out." For that reason, the instructions the researcher gives to the participants can be crucial for the success of the experiment. Sometimes the researcher will divert the respondents' attention to a fact irrelevant to the actual experiment, thus leaving their targeted intuitions intact.[144] More importantly, as pointed out by Meyer (2002, p. 57), "subjects need to be told what is stressed over and over again in any introductory linguistics class: that no linguistic form is more 'correct' than any other linguistic form, and that when linguists study language, they

[144] However, note Klavan (2012, p. 120), whose goal was not to hide the actual purpose of the experiment; "indeed, it was the goal to encourage participants to work out a conscious response strategy in differentiating between the two constructions."

are not interested in what individuals may have been taught about correct or incorrect usage in school but in how they naturally feel about a given linguistic construction." Many authors (e.g. Arppe & Järvikivi (2007a), Divjak (2008), Caines (2012)) encourage their respondents to make judgments on the *naturalness* rather than *correctness* of the stimuli to avoid them resorting to the prescribed rules of grammar when rating. For instance, Ullman's (1999, p. 54) instructions were worded as: "*Is the verb in a form that 'sounds' right to you and that you would naturally use in your own speech?*" Cowart (1997, p. 91) asked his respondents to imagine themselves in the role of a teacher of English and judge whether their students would sound like a native speaker if they used the sentences in question. However, the author also conducted an experiment where two groups of respondents were given the same sentences but different instructions (which the author labelled 'intuitive' and 'prescriptive') – the two groups nevertheless gave more or less similar responses, showing that "informants have very little ability to deliberately adjust the criteria they apply in giving judgments" (1997, p. 58). It is unlikely that subjects have multiple judgment routines, so regardless of the instructions, they will invoke the only kind of judging they know. As long as subjects are given some explicit set of instructions, the exact content of these instruction should not matter a great deal. In all our studies, we repeatedly emphasised to the speakers that they should disregard what they had been taught in school. But ultimately, we can never know whether subjects were giving genuine responses to the experimental material.

6.2. Questionnaire study 1

The first questionnaire study focuses on the instrumental singular of nouns ending in *-ar*. The corpus analysis of this case presented in Section 4.3.1.2 has already established that -om is more frequent with the majority of such nouns. Table 23 gives more detail about the corpus data. Our search retrieved around 250 *-ar* nouns used in Isg; however, only 56 of them appeared with both endings (this is labelled *attested doubletism*). It is important to note that the lexemes that appear exclusively with either of the endings tend to be the less frequent lexemes in the family – rather, doubletism appears in the higher frequency ranges. Since there were very few lexemes used exclusively with -em and their token frequency was very low, we decided to ignore

this group, hence in the questionnaire study we focus only on the remaining two groups.

Table 21. Type and token frequency of -ar nouns (Source: HNK, version 2.5)

	-om only	**-em only**	**both (attested doubletism)**
Lexemes	181	15	56
Tokens	750	25	1514 (1253 -om + 261 -em)

6.2.1. Questionnaire design

The corpus data were divided into three frequency bands on the basis of frequency ratios of the two forms (dividing frequency of the -om form with the frequency of the -em form):

- Band 1 (*absolute dominance,* B1), which shows no variation (i.e. all the lexemes in this band appear exclusively with -om, N = 181);[145]
- Band 2 (*strong dominance,* B2), which includes lexemes where -om is more than five times more frequent than -em (N = 19);
- Band 3 (*weak dominance,* B3), including lexemes which show a fairly equal distribution of -om and -em (the ratio is between 1:1 and 5:1; N = 37).

The reader might ask why the ratio 5:1 was chosen as the border point between Bands 2 and 3. If we look at Table 23, we can see that the ratio of the token frequency of the two forms in the case of attested doubletism is around 5:1, so this seemed a natural boundary in this instance.[146]

Three words were taken from each of the frequency bands, producing a total of nine test items. Questionnaire sentences containing each of these nouns were then created, trying to replicate original sentences from the corpus. Several authors object to using artificial sentences[147] for testing hypotheses, hence we tried to use sentences

[145] This is not to say that such forms are impossible; in Stefanowitsch's (2008) terms, they are *accidentally absent*.

[146] More accurately, 1253/261 = 4.8.

[147] Roland & Jurafsky (2002) call such sentences 'test-tube' sentences.

already attested in the language as much as possible. Since the presence/absence of a preposition did not seem to make a difference in the corpus data (as demonstrated in Section 4.3.1.2), we did not design the questionnaire to incorporate preposition use as a potential factor. However, we made sure there was an equal number of sentences where the target form was preceded by a preposition and where it was not. The respondents were presented with sentence pairs, where the pair differed only in the ending of the -ar noun. This was done so as to make it clear for the respondents what exactly they were evaluating. Had the pairs been split, we feared it would make the respondents focus on irrelevant parts of the sentence, which might influence their response.

After filler material was added, the questionnaire consisted of a total of 42 sentence pairs.[148] The filler sentences contained examples of inflectional doublets in the adjectival and pronominal paradigms. The questionnaire was designed and distributed via the SurveyMonkey online platform. The SurveyMonkey software has the option of individual randomization of questions, so each respondent saw the questions in different order. Respondents were recruited individually by the author via social networks and personal correspondence. The responses were collected in July 2013.[149]

Our null hypothesis (H_0) was defined in the following way: *The acceptability of two forms in free variation is in no way related to their frequency ratios nor to any of the known sociolinguistic factors.* In other words, speakers do not invoke the actual distributions of -om or -em in Croatian (with the corpus standing as the proxy for the users' prior lexical experience[150]) when rating the two forms. To test the H_0, we would need to compare the ratings of the paired sentences, i.e. sentences containing forms with -om and sentences containing forms with -em across different bands. If there is no visible (or statistically significant) difference in ratings, we cannot reject the null hypothesis. If H_0 is rejected, the alternative hypothesis (H_1) is defined as: *The*

[148] It is standard practice in psycholinguistics for the filler and test sentences to be represented in a 2:1 ratio (see e.g. Sprouse (2009, p. 335)).

[149] Ethical approval for this, as well as other questionnaire studies described in this chapter, was obtained in accordance with the University of Sheffield's Research Ethics Policy.

[150] Or, to use a term by Baayen, Milin, & Ramscar (2016), "a mirror of collective experience."

acceptability of a certain instrumental ending mirrors its frequency in the corpus data. This would mean that words in B1 and B2 would show a clear preference for -om, due to its greater frequency. For nouns in B3, the differences in acceptability ratings would be smaller, but again, there should be a noticeable preference for the ending -om. However, there is also an alternative explanation for the variation in acceptability ratings of the two endings other than frequency factors, namely is that it is a sociolinguistic phenomenon (it depends on age, or sex, or social class, or dialect etc.). For instance, some of Labov's older works show that women and older speakers are more self-conscious in their use of language and therefore tend to use the forms they consider correct more than would men. These variables would also need to be accounted for when performing our statistical analysis.

6.2.2. Questionnaire results

The total number of respondents was 95 (61 F and 34 M); their age distribution was from 18-42, with the majority of respondents being in their twenties (mean = 27.38, SD = 5.42, median = 26). 78 of the respondents either finished university or were currently studying; hence we have an overrepresentation of females as well as of highly educated respondents. Furthermore, almost half of the respondents came from the author's region of origin (Sisak-Moslavina County), hence this region is also overrepresented.[151]

We will first compare the average ratings of particular forms and their absolute frequency, presented in Table 24. As we already said in Section 3.3, previous research has shown that this relationship is not so straightforward. It would be unrealistic to expect that there would be a proportional relationship, i.e. as absolute frequency of a form increases, so does its rating. In Divjak's (2008) analysis, in 65% of the cases raw frequency and acceptability did not match up. Erker & Guy (2012) have also noticed a non-uniform behaviour among the highest-frequency items.

[151] Over-/Underrepresentation is established by comparing the proportions of a certain demographic category from the sample with the proportions from the population. According to the last census from 2011, females constituted 52% of Croatia's population, whereas only 5% of Croatia's population lived in Sisak-Moslavina County.

Table 22. Comparison of respondents' ratings and absolute frequencies of -ar nouns

Band	Forms rated	Form frequency	Mean rating	Median rating
B1	konobar-om	30	6.55	7
	konobar-em	0	2.71	2
	kipar-om	23	6.26	7
	kipar-em	0	3.27	3
	kuhar-om	13	6.64	7
	kuhar-em	0	2.28	2
B2	novinar-om	209	6.42	7
	novinar-em	8	3.04	3
	gospodar-om	158	6.02	7
	gospodar-em	10	4.33	4.5
	vladar-om	56	5.80	6
	vladar-em	4	4.66	5
B3	vratar-om	197	6.24	7
	vratar-em	47	3.42	3
	političar-om	85	5.44	7
	političar-em	39	4.24	5
	čuvar-om	35	5.55	6
	čuvar-em	9	4.48	5

We performed a bivariate correlation analysis between absolute frequency and the mean ratings as per the table above and the result of the Spearman's rank correlation[152] was .590 (p = .01, 2-tailed), which shows a moderately positive correlation (i.e. 59% of the data go in the desired direction, but the other 41% do not). For instance, the highest rated item is the one with an absolute frequency of a mere 13 (*kuharom*), while an item with a frequency of more than 200 (*novinarom*) has a slightly lower rating (but is still highly acceptable). These data confirm one of the conclusions presented in Arppe & Järvikivi (2007a, p. 151) that it is not necessarily true that high acceptability entails high frequency ($\neg$ *(acceptable* $\leftrightarrow$ *frequent)*). Another pattern also arising from the table is that highly frequent items are consistently judged as highly acceptable.

In order to test H_0, we did a 3 x 2 repeated-measures ANOVA design with BAND (3 levels) and ENDING (2 levels) as the within-subjects dependent variables and

[152] Spearman's rho is a nonparametric measure of rank correlation, which assesses how well the relationship between two variables can be described using a monotonic function. A perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other.

demographical information (GENDER, REGION, EDUCATION) as between-subjects factors. In light of the debate on the nature of Likert scales touched upon in Section 6.1, we rely on the conclusions of e.g. Cowart (1997) and Rasinger (2008), who argue that Likert scales are interval, thus justifying the use of parametric tests (such as ANOVAs). The results of the ANOVA are presented in Table 25.

Table 23. Repeated-measures ANOVA (Dependent variable: RATING)

Independent variable	**ANOVA result**	**Significance**	**Effect size**
ENDING	F (1, 83) = 39.166	p < .001	partial η^2 = .321
BAND	F (2, 166) = 4.237	p = .016	partial η^2 = .049
ENDING*BAND	F (2, 166) = 17.296	p < .001	partial η^2 = .172

Contrary to the null hypothesis, we can see that there are several factors influencing acceptability. The partial eta-squared (η^2) value reveals how much of the variation in question can be explained by each of the individual variables, their interactions, and error. Cohen (1969, p. 23) describes an effect size of up to 0.2 as "small", whereas an effect size of 0.5 is described as "medium" and is "large enough to be visible to the naked eye." Cohen describes an effect size of 0.8 as "grossly perceptible and therefore large."

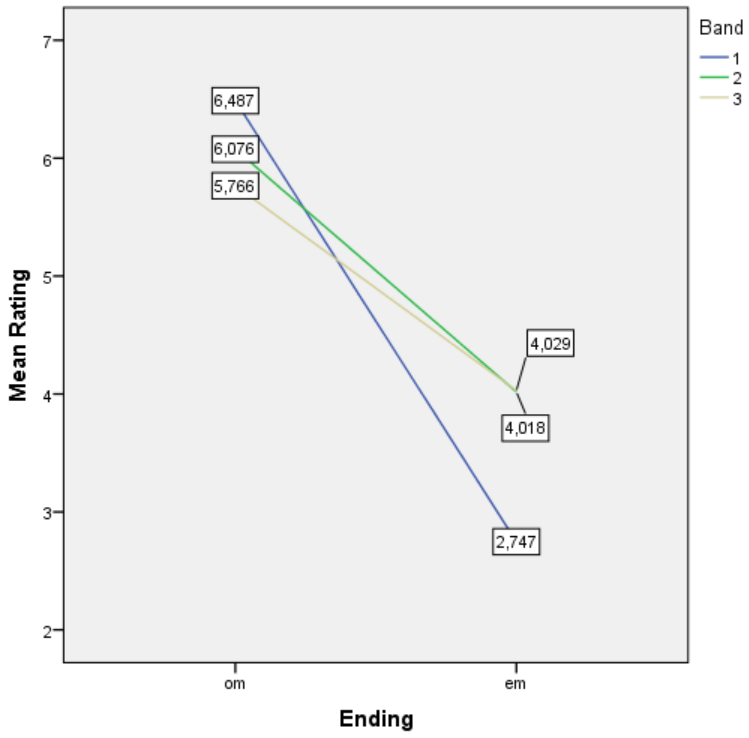
We can see that ENDING has the biggest effect on the rating, which is something we would expect. The effect size of .32 means that almost a third of the variance in ratings can be attributed to the fact that one ending is regarded more acceptable than the other regardless of any other factor. However, we can also see a significant effect of BAND itself and the combined effect of ENDING*BAND. It is this third factor we are most interested in. This tells us that the same ending is rated differently when in different bands. However, these three variables are unable to explain the full extent of the variance in ratings – the rest of the variance must be attributed to some other factor. As pointed out by Weskott & Fanselow (2009, p. 239), "if we have no hint at what might be responsible for the additional variance, we have to consider this variance as spurious." Since neither of the between-subjects factors were shown to have a significant effect, we cannot but conclude that a great portion of variance here is unexplained.

However, the ANOVA results as presented here do not tell us anything about the direction of the influence, i.e. are the ratings of a certain ending in B1 higher or

lower than its ratings in B2 etc.? As a reminder, our alternative hypothesis (H_1) was that *the acceptability of a certain instrumental ending would mirror its frequency in the corpus data.* This is a very general and non-directional statement. However, our actual assumptions were a bit more specific. We were expecting -om to be rated as more acceptable than -em throughout the questionnaire, in line with the works mentioned in Section 3.3. Moreover, we were expecting its ratings to be consistently high across the bands, i.e. B1om ≈ B2om ≈ B3om.

When it comes to -em, we also had some specific predictions. Whereas we have seen that high frequency → high acceptability, the works mentioned in Section 3.3 have shown that it is not necessarily true that low frequency → low acceptability, but rather that low acceptability → low frequency. We refrain from making any kinds of assumptions in such absolute terms. Rather, our aim was to determine the behaviour of -em between bands. We were expecting this ending to behave in a similar manner in both B1 and B2 since in both of these bands it has frequency values of zero or close to zero. Considering its frequency values are greater in Band 3, we were expecting the ratings to reflect this fact as well. Hence, our actual assumption was (B1em ≈ B2em) < B3em. However, as we can see below in Figure 3, neither of these predictions was borne out.

Figure 3. Mean ratings of -om and -em in different bands



6.2.3. Discussion

Several things can be noticed from Figure 3. First of all, we can see that the lines for B2 and B3 are more similar to each other than the line for B1 is to either of them. As a reminder, our original prediction was that B1 and B2 would behave in a similar manner. Furthermore, we can see that differences in ratings between the two endings in B1 (*absolute dominance*) are the most dramatic, i.e. the more frequent ending -om has an extremely high rating, while the less frequent ending -em (whose frequency is zero) has a low rating (but we can hardly call it absolutely unacceptable). Even though Boyd & Goldberg (2011, p. 56) say that "it is tempting to believe that the simple non-occurrence of a given form is sufficient to render it unacceptable", it seems this is not what happens with doublets. Although the -em form is not directly present in the input for a particular lexeme, it is still theoretically available in speaker's grammars for the whole *-ar* family of nouns.

When we move on to cases where both endings are present in the input, the results are not as expected. First of all, the ratings of -em drastically increase as soon as this ending appears in the corpus. This would lead us to the conclusion that even minimum exposure to a form in the input can lead to a significant increase in its acceptability. How else could we explain the fact that a form with a frequency of 0 has a mean acceptability of 2.28 whereas the form with a frequency of a mere 4 has an average acceptability of 4.66? This is in line with Bybee's (2010, p. 18) claim that "there is no way for frequency to matter unless even the first occurrence of an item is noted in memory. Otherwise, how would frequency accumulate?" This view is reinforced later in the text (2010, p. 60) when she says that native speakers can register a new item (word or chunk) with only one exposure.

Another point of interest is that as the rating of the less frequent form increases, the rating of its more frequent alternant decreases. At first glance, this would go against the view that acceptability is not probabilistic or a zero-sum game. In other words, it should not be the case that ratings of two competing forms add up to a constant value (1 or 100%, argued in Arppe & Järvikivi (2007a, p. 150)).[153] However, as is visible from Chart 3, the decrease in ratings of -om as we move from B1 to the

[153] Unless, of course, the rating scale is designed in such a way that they should add up to 100, as was the case in Bresnan (2007).

other bands is not as dramatic as the increase of ratings of -em between the respective bands. So the ratings are in an inversely proportional relation, but it is not a completely linear one.

6.2.4. Interim conclusions

The results of this study show that there is a strong relationship between relative proportions of competing forms in the corpus and their acceptability to native speakers. The results of the study can be summarised as follows:

- The dominant ending will always receive a high rating (> 5), a pattern also noted by Bermel & Knittl (2012, p. 21);
- Items with zero frequency in a large-scale corpus exhibit low levels of acceptability (but not absolute unacceptability). This result is similar to Keuleers, et al. (2007, p. 291), who have shown, on the example of Dutch plural doubletism, that "in most cases where one plural is preferred, speakers will not find that the other plural is *unacceptable*" (italics original);[154]
- There is a significant increase in acceptability as soon as an item moves from being infrequent to low frequent (i.e. even minimum presence in the corpus affects the rating in a positive manner);
- Forms with the 'dominant' ending are rated better than forms with the 'recessive' ending, but the acceptability of the former varies in accordance with the acceptability of the latter. As the acceptability of the recessive ending increases, the acceptability of the dominant ending decreases (i.e. the ratings tend to level out).

[154] This finding led Keuleers, et al. (2007) to claim that both endings (-s and -en in their Dutch examples) can be considered 'default', which is in direct opposition to dual-mechanism models, which claim that only one default is possible for any individual cell.

6.3. Questionnaire study 2

The most significant result that has arisen from the previous study was the psychological reality of the concept of bands, which was used as a proxy for relative proportions of the two competing forms. However, the previous study had several shortcomings, which need to be revisited in order to make our claim stronger. First of all, the number of respondents was not very high (N = 95)[155] and the tested material was of low frequency. More importantly, the material at hand involved a one-way relationship between the two endings, meaning that a single ending (-om) was dominant in all the tested lexemes.

The following two studies aimed to correct these drawbacks. For that purpose we needed to find examples of a two-way relationship between the endings, in which there was a more-or-less even distribution of items that prefer one ending and those that prefer the other. This way we could create a greater number of bands and examine whether the behaviour of an ending changes depending on whether it is the dominant ending or not. Having seen that doubletism is very widespread in Croatian, examples of such a relationship were not hard to find. The material we use in the present study are masculine plurals (see Section 4.3.1.3) and instrumental singular of masculine nouns ending in a palatal phoneme (see Section 4.3.1.2).

6.3.1. Setup of the study

In the previous study we defined three types of dominance of one doublet form over the other: *absolute*, *strong* and *weak*. A greater number of examples that we found in other instances of doubletism enables us to trace the behaviour of a single ending in various situations, ranging from its absolute dominance to its absolute recessiveness. Hence in this study we are dealing with as many as seven different bands. In the notation below, one form is taken as the orientation point based on which the terms dominance and recessiveness are defined (in the plural material, Variant_1 refers to the long plural; in Isg material, Variant_1 is the -em form). We also use

[155] Cowart (1997, p. 82) argues that the number of informants required is determined by the stability of the phenomenon itself – experiments targeted on the smallest differences will require a greater number of informants.

percentages of the two forms rather than ratios as a more intuitive and standardised measure for defining bands.

- Band 1 (named *Absolute dominance*) = $Variant_1$ appears 100% of the time : $Variant_2$ appears 0%
- Band 2 (*Extreme dominance*) = $Variant_1$ 90-99% : $Variant_2$ 1-10%
- Band 3 (*Weak dominance*) = $Variant_1$ 55-89% : $Variant_2$ 11-45%
- Band 4 (*Equiprobability*) = $Variant_1$ and $Variant_2$ appear in roughly same proportions (45-55%)
- Band 5 (*Weak recessiveness*) = $Variant_1$ 11-45% : $Variant_2$ 55-89%
- Band 6 (*Extreme recessiveness*) = $Variant_1$ 1-10% : $Variant_2$ 90-99%
- Band 7 (*Absolute recessiveness*) = $Variant_1$ 0% : $Variant_2$ 100%

Bands 1 and 7, then, contain items which appear in the corpus exclusively with a particular ending, even though the other one is also available and possible. The reasoning behind Bands 2 and 6 comes from Halliday's (1991a, p. 36) claim that the ratio 9: 1 is the dividing line between unmarked and marked forms in language as this is the point where, in Shannon & Weaver's theory of information, redundancy (R) and information (H) balance each other out (H = R = 0.5). Forms with ratios below 9: 1 represent, in Halliday's terms, competing forms proper.

The scheme as described above is the idealised scheme. Unfortunately, in the plural material we were unable to find enough examples in Band 4, hence this material only made use of six bands (the boundaries of B3 and B5 were extended to 50%). Halliday (1991b) claims that the situation of equiprobability is extremely rare in language in general. To use Shannon & Weaver's terminology again, this is the situation where information is maximum and redundancy minimum; however, minimum redundancy does not necessarily mean maximum efficiency because such a system would be too easily disrupted by noise (Halliday (1991b, p. 42). Hence it is no surprise that examples of a genuine 50: 50 distribution are hard to come across.

The null hypothesis (H_0), similar to the previous study, was defined as follows: *The ratings of the two forms in competition are not dependent on their proportional frequencies from the corpus.* We also defined alternative hypotheses that were to be tested either using parametric or post-hoc tests:

- The ratings of the dominant ending will be consistently higher than the ratings of the recessive ending;
- The ratings of the recessive ending will show more variation across bands than those of the dominant ending;
- An increase in the rating of the recessive ending will go together with a decrease in the rating of the dominant ending;
- We will not find an influence of absolute frequency (i.e. lexemes with different absolute frequencies but similar proportional distributions of endings will have similar ratings).

6.3.2. Questionnaire material

Had we followed the design of the previous questionnaire, which contained 3 items per band, the present questionnaire would be unreasonably long (more than 200 items after the inclusion of filler items). However, to ensure that as many examples as possible were included (so as to minimise lexical effects), we decided to make two versions of the questionnaires, each following the same design but using different lexemes. We present the material below.

6.3.2.1. Plural material

In Section 4.3.1.3 we have seen that the number of syllables is said to determine the behaviour of nouns in this case. Monosyllabic nouns prefer long plurals, disyllabic prefer the short form, whereas disyllabic nouns with the fleeting /a/ are somewhere in between and their behaviour is largely unpredictable. For this reason, we aimed to include lexemes from all three syllable groups in the questionnaire. After lexemes in each band were further subdivided based on the number of syllables, two lexemes were chosen from each cell to be included into the final questionnaire. As a reminder, there was a scarcity of examples in B4 (*equiprobability*), so the few examples that were found with more-or-less equal proportions were redistributed into bands 3 and 5, hence the final number of bands in this dataset is 6 rather than 7. Moreover, as this instance of doubletism spreads throughout the whole plural paradigm rather than being located in a single cell, the lexemes were used in a variety of plural cases. We present the list of lexemes used in the questionnaire in Table 26.

Table 24. Plural material for questionnaire study 2

Quest.	No. of syllables	B1	B2	B3	B4	B5	B6	B7
A	*1σ*	dug 'debt'	rog 'horn'	zvuk 'sound'	*	trak 'ray'	grijeh 'sin'	konj 'horse'
	fleeting	rit(a)m 'rhythm'	tor(a)nj 'tower'	puc(a)nj 'bang'	*	pal(a)c 'thumb'	šilj(a)k 'pike'	stal(a)k 'stand'
	2σ	dio 'part'	tečaj 'course'	pojas 'belt'	*	gavran 'raven'	greben 'reef'	badem 'almond'
B	*1σ*	bok 'flank'	ključ 'key'	znak 'sign'	*	skut 'skirt'	srh 'shudder'	keks 'biscuit'
	fleeting	troš(a)k 'expense'	manj(a)k 'deficit'	pijet(a)o 'rooster'	*	praš(a)k 'powder'	kol(a)c 'stick'	nok(a)t 'nail'
	2σ	autoput 'motorway'	golub 'pigeon'	prsten 'ring'	*	jablan 'poplar'	korijen 'root'	jelen 'deer'

An additional complication was created by the fact that some forms had not only alternative short/long forms, but their long forms had two possibilities as well. The word *pojas* (version A), for instance, can have a short plural (*pojas-i*) and two long plural forms (*pojas-ov-i* or *pojas-ev-i*). The same goes for *autoput* 'motorway' in version B. For those two lexemes we had to include all three alternatives in the questionnaire, even though we knew this would create problems for later analysis. After obtaining the results it was decided that between the two long alternatives, whichever one was rated higher would be considered as the rating of the long stem and the other rating was ignored.

6.3.2.2. Instrumental singular material

In Section 4.3.1.2 we have seen that this sub-family of nouns is quite hard to describe in a rule-based approach. For convenience's sake, we remind the reader of the rules applicable in this case:

1. Words ending in a palatal sound take -em (e.g. *ključ* > *ključem*)
    a. BUT, words where the final palatal is preceded by /e/ take -om instead due to dissimilation (e.g. *zec* > *zecom*);
        i. BUT, Rule 1a applies only to words of 1 or 2 syllables; words of 3+ syllables will take -em nevertheless (e.g. *ravnatelj* > *ravnateljem*).

However, this is a very simplistic description. In reality, there are a lot of inconsistencies not only between different grammar books but also within a single manual. On the other hand, the memory-based models that we implemented in Chapter 5 were highly successful in predicting the dominant ending for these nouns.

Our corpus analysis identified over 2,000 nouns exhibiting this kind of variation. We also included nouns and proper names of Romance origin ending in -io (alternatives in question being -ijem and -iom) and nouns ending in 'virtual palatals' /št/ and /žd/ (see footnote 89 for description of the term 'virtual palatals') into this material. After grouping the lexemes into bands, we selected four lexemes from each band to be included into the final questionnaire. These are given in Table 27 below.

Table 25. Instrumental singular material for questionnaire study 2

Quest.	B1	B2	B3	B4	B5	B6	B7
A	udarac 'hit'	Štimac (surname)	*sprej* 'spray'	*staž* 'internship'	*Mario* (name)	*Beč* 'Vienna'	nos 'nose'
	pištolj 'pistol'	redatelj 'director'	Kovač (surname)	*Sergej* (name)	*trofej* 'trophy'	Francuz 'Frenchman'	*šverc* 'smuggling'
B	sudac 'judge'	*križ* 'cross'	plašt 'cloak'	*esej* 'essay'	*Bush* (surname)	mjesec 'month'	pojas 'belt'
	nož 'knife'	ravnatelj 'principal'	*volej* 'volley'	*DiCaprio* (surname)	*muzej* 'museum'	crtež 'drawing'	*gušt* 'gusto'

The selection of this material was much simpler than the previous one as we did not have to pay attention to the number of syllables. However, there were other things that needed to be considered. As can be seen, the table contains a variety of final palatals. It was impossible to ensure an equal distribution of lexemes sharing the same palatal across the bands as some palatals have a clear preference for one ending and cluster in one or two adjacent bands. Hence the variable PHONEME could not be used in the subsequent analysis. However, post-hoc tests might reveal whether words ending in the same palatal behave in a similar way. Moreover, words of foreign origin (which are given in italics in Table 27 below) seem to be most liable to variation (as noted by Raguž (1997)) and they mostly cluster in the middle three bands. Again, we could not ensure an equal representation of foreign words across the bands, so we planned to use post-hoc tests and other methods to check whether there was any common pattern among them.

Additional filler sentence pairs were added to each questionnaire, producing two final questionnaires, each containing 48 sentence pairs/triplets. In total, each respondent had to provide 102 ratings.

6.3.3. Questionnaire results

The responses to the questionnaire were collected on paper. The author contacted a foreign language school in Zagreb asking whether they would be willing to distribute the questionnaire among their students. Foreign language courses are frequented by a cross-section of the population, hence it was a convenient way of collecting responses from different strands of society. Around 150 questionnaires were collected using this method. An additional 50 were collected by the author himself among various populations. The overall respondent structure presented in Table 28 below.

Table 26. Respondent structure for questionnaire study 2

Version A	**N** = 107 **Gender**: 39 male, 66 female, 2 missing **Age**: Mean = 29.58, SD = 10.9, Median = 25 (29 missing values) **Region**: Zagreb (50), central Croatia (12), east Croatia (9), south Croatia (14), 'mixed' regional background[156] (13) **Education**: 20 secondary school graduates, 43 currently studying, 44 university graduates
Version B	**N** = 102 **Gender**: 35 male, 67 female **Age**: Mean = 30.53, SD = 11.9, Median = 25 (26 missing values) **Region**: Zagreb (45), central Croatia (12), east Croatia (7), south Croatia (16), 'mixed' regional background (17) **Education**: 15 secondary school graduates, 51 currently studying, 36 university graduates

The first thing that needed to be determined was whether the results of the two questionnaire versions are directly comparable or whether they need to be considered as separate datasets. The latter case would mean that the obtained pattern is specific only for the lexemes in the questionnaire and not generalisable to other lexemes. The former situation would mean that whatever conclusions are drawn from the study can also be generalized to the whole inflectional cell under analysis. For that purpose, we conducted *independent samples t-tests* on the two sets of material.[157] The test results did not reach the .05 significance level, neither for the plural material (t (7,479) = 1.641, p = .101) nor for the Isg material (t (5,829) = -0.242, p = .809). Hence we decided that it was unnecessary to analyse questionnaire A separately to questionnaire B and it was justified to combine them into a single dataset. We now proceed with the analysis of the two sets of material.

[156] The respondents were asked two questions about their regional background: where they were born and where they spent the majority of their life. In cases where the two places were in dialectally different regions, they were placed in the 'mixed' category.

[157] The t-statistic is the difference between the *observed* and the *expected* mean, divided by standard error of the observed mean. In cases where the observed mean perfectly replicates the expected mean, the t-test value will be zero.

6.3.3.1. Results for plurals

To test the null hypothesis, we performed a mixed-effects ANOVA with 3 independent within-subjects variables: BAND (6 levels), SYLLABLE NUMBER (3 levels) and STEM (long/short). We also tested GENDER and REGION as between-subjects variables. The results of the ANOVA are presented below.

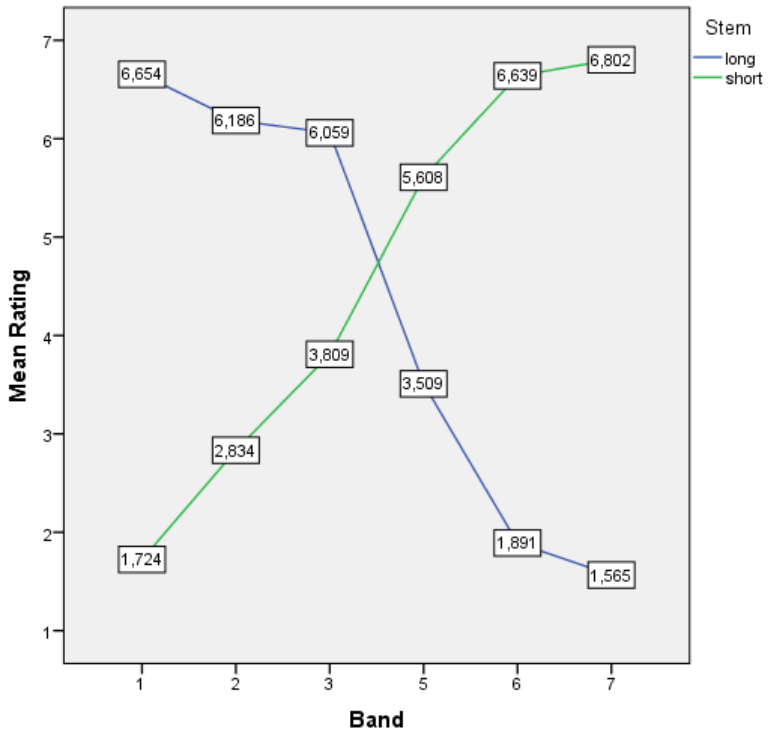
Table 27. ANOVA results for plural material (dependent variable: RATING)

Independent variable	**ANOVA result**	**Significance**	**Effect size (partial η^2)**
BAND	F (5, 905) = 24.659	p < .001	= .120
SYLLABLE	F (2, 362) = 7.715	p = .001	= .041
STEM	F (1, 181) = 1.302	p = .255	= .007
BAND*SYLLABLE	F (10, 1,810) = 2.176	p = .017	= .012
BAND*STEM	F (5, 905) = 242.062	p < .001	= .572
SYLLABLE*STEM	F (2, 362) = 3.692	p = .026	= .020
BAND*SYLLABLE*STEM	F (10, 1,810) = 7.795	p < .001	= .041
GENDER	F (1, 181) = .353	p = .703	= .004
REGION	F (8, 181) = .733	p = .662	= .031

Just like in the previous study, BAND is a significant factor, although its effect size would be considered small. An even smaller, yet still significant, effect on ratings comes from the factor SYLLABLE, which means that words with a different syllable structure show different patterns of behaviour. The insignificance of the individual effect of STEM on RATING is not surprising. Whereas the previous study contained examples where one form was consistently rated better than the other, here both forms get rated as highly acceptable and highly unacceptable roughly the same number of times, so these values cancel each other out in the analysis. It is only when STEM is combined with BAND that the biggest effect arises, which can be considered a medium-sized effect, as per Cohen (1969). The sociolinguistic factors do not have a significant effect on the ratings.

Let us see what the BAND*STEM relationship looks like in graphic form. The following figure (Figure 4) shows how mean ratings of the long and short variant change as we move between the different bands.

Figure 4. Mean ratings of the long and short plurals in different bands



We can see that the long plural is given the highest rating in B1, which is the situation of its absolute dominance. As its dominance in the corpus weakens, its ratings decrease as well. When we move from B3 to B5, which is where the proportions turn in favour of the short plural, ratings of the long plural drastically drop. The short plural shows an identical pattern as we move in the opposite way, from B7 to B1.

6.3.3.2. Results for instrumentals

The same mixed-effects ANOVA as in the previous section was performed on the Isg material. However, in this case we had one fewer variable as SYLLABLE is not a relevant factor here. The results of the test are presented in Table 30 below.

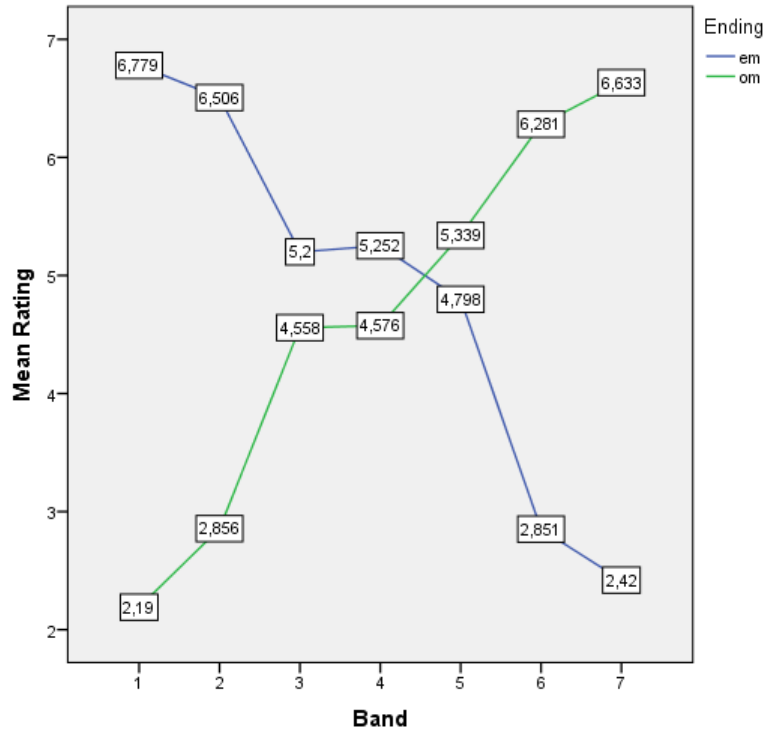
Table 28. ANOVA results for Isg material (dependent variable: RATING)

Independent variable	**ANOVA result**	**Significance**	**Effect size (partial η^2)**
BAND	F (6, 1,140) = 8.567	p < .001	= .043
ENDING	F (1, 190) = 1.561	p = .213	= .008
BAND*ENDING	F (6, 1,140) = 82.246	p < .001	= .302
GENDER	F (2, 190) = 1.681	p = .189	= .017
REGION	F (8, 190) = 2.055	p = .042	= .080

The results are similar to the plurals, in that BAND is once again a significant variable, but with a small-sized effect and ENDING (equivalent to STEM in the previous section) on its own is statistically insignificant. However, just like above, it is the combination of these two factors that has the strongest effect on rating. Similar to the previous study, we can see that the effect size of the most significant variable is not overly large (only 30%). This means that either there are some other factors that affect the rating which we did not consider or the rest of the variance is simply unexplainable.

Figure 5 below shows a similar pattern to Figure 4 – ratings of the dominant form decrease as its level of dominance decreases. However, one thing that is slightly different from the previous material is that the two lines do not intersect until after B4. Furthermore, the rating of -em is still relatively high even in B5. This suggests that this ending is more salient in speakers' mental grammars, which keeps its rating on high levels even when the relative proportions are not in its favour.

Figure 5. Mean ratings of Isg -om and -em in different bands



The ANOVA has also shown a weak effect of REGION, which did not arise in the plural material. A post-hoc analysis has revealed that respondents from the Kajkavian parts of Croatia gave significantly different ratings than respondents from other parts. Namely, it seems that in those speakers' mental grammars a different

ending has salient status. For Kajkavian speakers the two lines intersect as early as B3; in other words, -om is rated better even though it is still the recessive ending in that band and it retains higher ratings in all the subsequent bands. However, we should not take these results as highly conclusive considering the number of respondents from this region was only twelve.

6.3.4. Discussion

The present questionnaire study has provided further evidence that relative proportions do play an important role in native speakers' mental grammars, thus complementing the results from Section 6.2. Once again, zero-frequency items (populating Bands 1 and 7) received low ratings, but they were not exclusively rated as absolutely unacceptable (especially note the mean values of 2.19 and 2.42 for the Isg material). Moreover, the "levelling-out" effect from the previous study arose once again. Both competing forms in in the inside bands (Bands 3-5) received more or less similar ratings, but ratings of middling values.

However, this is not the only effect that arose from the study. One of the conclusions in Section 6.2.4 was that the dominant ending would always receive a high rating, which turned out not to be the case in the present study. Rather, we noticed instances in which the ending that was rated as better was not the more frequent ending. Since such mismatches were especially widespread in nouns that end in *-ej*, we analyse the results for this family of nouns in more detail in Table 31 below. For ease of use, we underlined the higher numerical value of both frequency and rating for each doublet pair.

Table 29. Comparison of ratings and frequencies for -ej family of lexemes

Doublet forms	Band	Corpus frequency (HJR)	Speakers' mean rating
sprejem	B3	110	5.23
sprejom		33	4.38
volejem	B3	117	5.28
volejom		38	4.43
Sergejem	B4	63	6.16
Sergejom		52	3.83
esejem	B4	20	5.36
esejom		21	4.21
trofejem	B5	26	5.08
trofejom		52	4.68
muzejem	B5	95	4.11
muzejom		285	5.31

We can see that half of the lexemes from the table above appear more frequently with -em and the other half with -om. However, speakers gave forms with -em better ratings in five out of six instances. This suggests there is some kind of a *family effect* at work. In other words, when rating words of similar phonological structure, respondents tend to give similar ratings to the same variants, even if those variants are not in the same, or even adjacent, bands. The fact that the only time this did not happen was in the instance of *muzej*, where the dominance of -om is quite strong (75%), also seems to suggest that these family effects only have a limited reach.

It is also important to note that the form that was rated better is not the one that is licensed in grammar books. We have already said that nouns whose final vowel is /e/ should undergo dissimilation and take -om; however, here we can see the -em form receiving higher ratings. This seems to suggest that the speakers' mental grammars are much simpler than formal grammars (cf. Dąbrowska (2008b, p. 948)) – instead of making various sub-rules and principles (such as dissimilation) which work against each other, they consistently use the same form of phonologically similar words. Fehringer (2004, p. 306) also noticed similar effects in her data, which she labels *gang effects*. "The emergence of these gangs reflects the way in which native speakers have come to organize their lexicon over the generations." The reader will also remember that the TiMBL models that were presented in Section 5.3 gave most weight to the final segments of the word in their analyses, thus showing that phonological families

are a cognitively plausible and realistic concept for the organisation of mental grammars.

Similar family effects, although not as widespread, could also be seen in other parts of the questionnaire. For instance, in the plural material, the short plural of *manj(a)k* was given an unexpectedly high rating (mean 3.74) when compared to other short plurals in the same band (the mean rating of all other short forms in B2 is 2.8). We argue this is so because respondents also turn to other lexemes in the same -(a)k family (*prašak, šiljak, stalak*), where the short plural is the dominant form, hence rated more favourably, which in turn increases the rating of the recessive form *manjci*. The filler material for the questionnaire also contained the superlative forms of the minimal pair of adjectives *tijesan* and *bijesan*, both of which appear as doublets. Even though the corpus distributions of the alternative forms are quite inverse ((*naj*)*tješnji* - 299 / (*naj*)*tjesniji* - 40; (*naj*)*bješnji* - 26 / (*naj*)*bjesniji* - 34), the long superlative received higher ratings for both adjectives (*najtješnji* - 3.66 / *najtjesniji* - 5.72; *najbješnji* - 3.10 / *najbjesniji* - 6.27). As noted in Section 4.3.5 the long form is the recommended form for the whole family of adjectives ending in *-(a)n*. These family effects will be revisited in the next study.

6.4. Questionnaire study 3

In Chapter 5, we applied two memory-based, or 'lazy learning', models to the Croatian data. We saw that both models were highly successful in predicting the dominant form of a lexeme in two cases, Isg and Gpl. Even though those results on their own provide quite a strong argument that these models realistically capture the processes taking place in native speakers' mental grammars, an even better test of their successfulness would be to match the performance of the models to the performance of native speakers. This is the aim of the present study. This method draws on a substantial amount of previous research demonstrating the efficiency of memory-based systems (see Divjak, Dąbrowska, & Arppe (2016) for the most up-to-date list of references). For instance, Elzinga (2006) conducted a survey to check whether speakers' responses on English adjective comparison matched the predictions made by the model. The response that was circled more by respondents was considered the desired outcome for the model. In the end, responses by the AM model and the

speakers matched in 87.9% of the cases. Similar to this, in an analysis of allomorphy in the instrumental singular case in Serbian (Milin, Keuleers, & Filipović Đurđević, 2011), the outcomes of the model also closely resembled native speakers' behaviour. Even though the authors refrain from claiming that the architecture of the model they used (TiMBL) mirrors the organization of the human cognitive system, they do state that "the patterns observed in the behaviour of native speakers can be accounted for by a very simple learning principle", which would argue against the application of rules in describing linguistic phenomena (2011, p. 78). Furthermore, their model was equally successful in predicting -om and -em, providing counterevidence to models of language that invoke the concept of 'defaults' (such as the Dual Mechanism model). Rácz, et al. (2014) compare the predictions of two alternative algorithmic models, Minimal Generalization Learner (MGL) and Generalized Context Model (GCM) to the predictions of native English speakers on the past tense of nonce verbs. Both models showed a highly significant correlation with the regularization rate of native speakers, but GCM was better at predicting the variation across items than MGL. Hence the authors conclude that "analogy to individual instances is a better approximation of the behaviour of our subjects than recourse to abstract generalizations" (2014, p. 7).

In comparisons such as above, there are three possible outcomes, as pointed out by Divjak, Dąbrowska, & Arppe (2016, p. 3): a) the model performs worse than humans, b) they perform more or less equally, or c) the model outperforms humans. According to the authors, outcome c) has not yet occurred in linguistics (although models that perform better than humans have been developed for other disciplines, such as medical diagnostics or criminology). If the first outcome is obtained, that means that the models are incomplete, i.e. are missing some information that humans make use of. Hence the researcher needs to go back to their model, recode it and test it again. The second option is the most encouraging one as it tells the researcher they have managed to pinpoint the same criteria speakers use. However, merely looking at percentages of 'correct answers' is not enough to label a model successful; one also needs to compare the variability and uncertainty in both datasets (high probability of outcome corresponding to high percentage of choice; low probability corresponding to more variability in human responses).

6.4.1. Setup of the study

In this study, we use the method of *controlled experiments*, as often used in the medical and social sciences. In such experiments, either the performance of two groups of respondents (control group vs. test group) or the performance of the same group of respondents before and after a certain intervention is compared. Our design is of the former type. However, unlike the works referenced above, where the performance of the models was directly compared to the performance of native speakers, in the present study we compare the performance of both to the corpus data, which serves as the control group.

The standard null hypothesis in controlled experiments is to assume that there will be no significant difference in performance between the compared groups or within the same group before and after a particular intervention. This would mean that there is no effect of the intervention that is being applied. Therefore, in the present study our null hypothesis is defined as:

H_0: *corpus data = model results = questionnaire results*

However, as pointed out by Cowart (1997, p. 40), "reliable differences are of little scientific interest unless we can determine how they came about in the experiment". Therefore, if our null hypothesis cannot be rejected, we believe this means that both model results and questionnaire results accurately replicate the corpus data, i.e. corpus distributions form an important part of the grammars of both the models and native speakers. We can test the null hypothesis by means of either an independent-samples t-test or a one-way ANOVA. The main difference between these two tests is the number of groups compared. The t-test is suitable for comparing only two groups – in our case this would mean performing three separate pairwise analyses: corpus vs. model, model vs. questionnaire and corpus vs. questionnaire. However, since the term 'model results' in our case actually encompasses the results of six different models (AM TYPES, AM TOKENS, AM DOUBLE ENTRIES, TIMBL TYPES, TIMBL TOKENS, TIMBL DOUBLE ENTRIES), the number of individual t-tests would be even greater. In such situations, a one-way ANOVA can replace all the individual t-tests.

6.4.1.1. Setting up the models

The modelling described in Chapter 5 was performed in artificial and idealised settings of the models. The models used the whole dataset of forms for a particular inflectional cell for making predictions, which means that were we to re-run each of them again, given precisely the same input, all the models would always return precisely the same output. However, we know from experience that speakers' performance is never that regular. "Real people sometimes respond one way and sometimes another way to the same input, reflecting at least in part the relative frequencies of those alternatives in an individual's accumulated experiences" (Chandler, 1993, p. 602). Our mental grammars are sensitive, among other things, to quantitative and qualitative differences in input, but also to noise and imperfect memory. As pointed out by Lonsdale (2002, p. 354), "to assume that input is always complete, consistent and correct would not reflect how language is actually used." In order to make the models psychologically realistic (i.e. make them resemble real speakers), we needed to include these facts into their performance.[158]

In both AM and TiMBL, the researcher can set the probability of each item of the dataset being selected for the analysis to any value between 0 and 1 (the standard being 0.5). That would mean that 50% of the data would get randomly eliminated for each individual analysis. By the very nature of random selection, none of the samples would ever be the same. This is one way of making the model imitate differences in input of individual speakers. Daelemans (2002, p. 175) claims that introducing imperfect memory can make AM equivalent to standard statistical procedures. In his experiment, forgetting 25% and 50% of the data did not significantly decrease the accuracy of the model.

For the purpose of this study, we set the imperfect memory level to 0.5 and ran each of the six models 100 times. This basically gave us six different samples of 100 'artificial' speakers of Croatian to which the performance of real speakers will be compared. An alternative way of setting up the models to replicate native speakers' behaviour would be, instead of manipulating the size and content of the training set,

[158] Note, however, an opposing view by Daelemans, van den Bosch, & Zavrel (1999, p. 38), whose empirical results "strongly suggest that keeping full memory of all training instances is at all times a good idea in language learning."

to manipulate the parameters of the model. However, that method rarely brings significant changes to the model. For instance, in Milin, Keuleers, & Filipović Đurđević (2011), the model with even a single nearest neighbour was already a very significant predictor of the participants' choice of allomorph, regardless of the similarity metric and decay weighting used.[159]

6.4.1.2. Questionnaire setup

When linguists who work with computational models such as those above want to verify the output obtained by the model against native speakers, they usually select only the material that was analysed incorrectly by the model, with the assumption that the correctly analysed items will also be correctly analysed by the speakers.[160] However, with the design we described above, it would not be methodologically correct to include only the misanalysed items into the questionnaire as this would automatically create a significant difference between the two samples. Hence in the final questionnaire we included a variety of lexemes from the 'ideal' models (described in Section 5.3) – lexemes whose dominant ending was predicted correctly by all the models, those predicted incorrectly by one of the models (AM or TiMBL), and those predicted incorrectly by both models.

Since the previous questionnaire studies already established that relative proportions are a highly significant predictor of native speakers' performance, we make use of the concept of bands in this study as well. We re-use the bands defined in the previous studies; however, we do not take the whole scale of relative proportions into account. What we focus on instead is only the behaviour of each ending in various situations of its dominance, ignoring lexemes where it is the recessive ending. The four bands thus defined are:

[159] In one part of their analysis, Divjak, Dąbrowska, & Arppe (2016) also run a certain number of 'dumb models', matching the number of participants in their study; however, their manipulation of the model setup is not in the dataset itself but rather in the variables. Hence each of those 'dumb' models uses a different subset of the 87 variables used in the full analysis.

[160] Rytting (2002), for instance, tries to establish why his respondents used different forms from those predicted by the model, thus implicitly taking a stance that the model is right and the speaker wrong – a view we find hard to agree with.

- Band 1: *absolute dominance of a variant* (100%);
- Band 2: *strong dominance of a variant* (90-99%);
- Band 3: *weak dominance of a variant* (60-90%);
- Band 4: *equiprobability* (50-60%).

We selected an equal number of lexemes with different dominant endings from each band, i.e. for Isg, two lexemes with the dominance of -om and two lexemes with the dominance of -em; for Gpl, two lexemes with the dominance of -a, two with dominance of -i and two with the dominance of a-a. Similar to questionnaire study 2, two different versions of the questionnaire were created, each using the same design but different lexemes. The total number of sentences in each version was 60 (16 Isg lexemes + 24 Gpl lexemes + 20 fillers). Respondents were assigned to a questionnaire version based on their date of birth – if they were born on an odd date, they were asked to fill out version A; if they were born on an even date, they filled out version B.

Unlike the previous two studies, which involved rating particular forms, the present questionnaire was a forced-choice task in which respondents had to indicate the form they would use in a given sentence. Even though the two methods of obtaining responses involve two different processes (*passive reception* vs. *active production*, Bermel & Knittl (2012, p. 251)), hence their results are not directly comparable, previous research using both methods (either in a single questionnaire or in separate studies) has shown that "the way respondents rate is a small contributory factor to the choices they make" (Bermel, Knittl, & Russell, 2015, p. 305).[161] We will come back to this relation in Section 6.4.2.4.

[161] Note, however, the view presented in Divjak, Dąbrowska, & Arppe (2016, p. 9) that "choosing a [word] to go in a particular sentence is a fairly artificial task; it is not what speakers do during normal language use." However, the authors nevertheless agree that it provides useful information about speakers' preferences.

Table 30. Questionnaire material (glosses provided in Appendix B)

		absolute domination (B1)		**strong domination (B2)**		**weak domination (B3)**		**equiprobability (B4)**	
		A	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Isg	*domination of -om*	Senj šverc	Polančec keš	mjesec Šprajc	crtež Poreč	hokej krpelj	šlic Bush	Rovinj kolaž	štambilj esej
	domination of -em	slučaj cilj	znanac kraj	temelj muž	tuš križ	Lošinj volej	vic roštilj	staž linč	trofej imidž
Gpl	*domination of -a*	utvrda crta	obrva utrka	pošta Ande	cesta zvijezda	nevjesta kovrča	Alpe svekrva	gesta cista	kasta bista
	domination of -i	crpka akna	stotinka milijarda	maska boljka	hrenovka njuška	dagnja praćka	ljuska vesta	fešta brazda	pasta naranča
	domination of a-a	litra zemlja	sestra pjesma	daska smokva	guska puška	školjka voćka	olovka breskva	lignja ljestve	bačva spužva

The questionnaire was designed in Qualtrics, a web-based software. Several innovations were introduced into the questionnaire, compared to the previous two studies. For instance, since we were interested in the respondents' initial, spontaneous response, they were required to provide their answers within 8 seconds (in the previous studies, respondents' time was unlimited).[162] R. Ellis (2005, p. 160) argues that tests that place time constraints on respondents tap more directly into implicit knowledge, which is the knowledge we want to get into as well.[163] This limit was also introduced in the hope that respondents would not have enough time to mull over, e.g. issues of prescriptive grammar. If the respondents did not answer within the time limit, they were automatically transferred to another question.

After every timely answer, respondents were redirected to a follow-up question, worded as *'How certain you are of this answer?'* The answer to this was provided by means of a slider scale (0-100). This part was not timed, so the respondents had time to think, which might even lead them to reconsider their previous answer. Only three points on the scale were labelled: 0 (*I should have answered differently),* 50 (*I am not sure of my answer*) and 100 (*I am certain this is the correct answer*).[164] This measure of 'confidence' will also be viewed against the concept of bands, the assumption being that the respondents will be less certain about the form of the lexemes from bands 3 and 4 as there is more interference from the recessive ending.

[162] The motivation behind this time limit comes from Gutierrez (2013) – in a task similar to this, the author set his respondents the time limit of 6-9 seconds, with the explanation that 3-6 seconds were needed for processing the sentence and another 3 seconds for indicating the answer.

[163] Hasher & Zacks (1984) claim that information about frequency of occurrence is encoded in an implicit manner. Various other authors claim that language learning itself is an implicit rather than an explicit process (Newell & Bright, 2001), (Ellis N. C., 2002).

[164] The exact Croatian wording of this question was: *Koliko ste sigurni u Vaš odgovor?* The labels were worded as: *Ipak sam trebao/la odgovoriti drugačije* (0), *Nisam pretjerano siguran/na* (50), and *U potpunosti sam siguran/na u ovaj odgovor* (100).

6.4.2. Results

The total number of respondents was 381 (193 for questionnaire A and 188 for questionnaire B). Table 33 below gives an overview of the demographic composition of the respondents.

Table 31. Respondent structure for questionnaire experiment 3

		Quest. A	Quest. B
Gender	Female	164 (85%)	157 (83.5%)
	Male	29 (15%)	31 (16.5%)
Age	Mean	26.95	26.09
	SD	5.5	4.59
	Median	26	26
Dialect	Štokavian	121 (62.7%)	122 (64.9%)
	Kajkavian	59 (30.6%)	51 (27.1%)
	Čakavian	13 (6.7%)	15 (8%)
Region	Central Croatia	38 (19.7%)	30 (16%)
	Zagreb	57 (29.5%)	63 (33.5%)
	Dalmatia	23 (11.9%)	26 (13.8%)
	Slavonija	25 (13%)	24 (12.8%)
	Istria and Kvarner	15 (7.8%)	12 (6.4%)
	Bosnia & Herzegovina	3 (1.6%)	4 (2.1%)
	Northern Croatia	22 (11.4%)	21 (11.2%)
	Dalmatinska Zagora	10 (5.2%)	8 (4.3%)
Education	Secondary school graduate	61 (31.6%)	50 (26.6%)
	Community college graduate	25 (13%)	27 (14.4%)
	University graduate	107 (55.4%)	111 (59%)
Occupation	Studying	83 (43%)	82 (43.6%)
	Working	92 (47.7%)	90 (47.9%)
	Other	18 (9.3%)	16 (8.5%)

An effect that has been persistent in all three studies presented in this chapter is that many more female than male respondents participated in the surveys. We admit to not knowing why this should be the case nor how this problem can be amended.

Prior to starting any kind of analysis, similar to the previous study, it was necessary to determine whether the results of the two questionnaire versions statistically differ in any way. We performed an independent-samples t-test, comparing the distributions of both the answers and the demographic data. Since none of the tested features approached significance, we decided that we can perform a single statistical analysis on the collated dataset.

6.4.2.1. Analysis 1: Group comparison

The predictions of all the models as well as the results of the forced-choice task are presented in Appendix 3. A variety of patterns can be seen in that table. For some lexemes (e.g. lexemes in Band 1 in both sets of material), all the models gave similar predictions, which are also in line with both the corpus and the questionnaire data. There are other lexemes for which either one or both of the models produce a great deal of 'incorrect' outcomes (e.g. all three TiMBL models were unable to predict the expected forms *Rovinj-om* or *olovak-ā*). It was noted in Chapter 4 that this usually happens with lexemes whose dominant ending is not the dominant ending for the whole family (-om in Isg and -a or a-a in Gpl).

To test the null hypothesis that the means of the 3 (8) groups are not significantly different, we performed a one-way ANOVA with a post-hoc Tukey test on variances. The results are given in Table 34 below. The asterisk marks significant differences (at .05 level).

Table 32. Results of one-way ANOVA

Control group	Test groups	Mean Difference (Control - Test)	Std. Error	p-value
Corpus (Isg)	Questionnaire	9.656	6.115	.762
	AM TYPE	-.031	6.115	1.000
	AM DOUBLE	13.094	6.115	.392
	AM TOKEN	-12.750	6.115	.428
	TiMBL TYPE	6.375	6.115	.967
	TiMBL TOKEN	12.344	6.115	.472
	TiMBL DOUBLE	21.094	6.115	.015*
Corpus (Gpl)	Questionnaire	13.604	5.615	.233
	AM TYPE	8.375	5.615	.812
	AM DOUBLE	16.854	5.615	.057
	AM TOKEN	-15.104	5.615	.129
	TiMBL TYPE	17.708	5.615	.037*
	TiMBL TOKEN	11.896	5.615	.405
	TiMBL DOUBLE	35.187	5.615	.000*

The only sample that turned out to be significantly different from the corpus data in both the Isg and Gpl datasets was the TiMBL DOUBLE ENTRIES sample. So this model does not give realistic results. Furthermore, the TiMBL TYPE sample differs significantly from the corpus sample only in Gpl, but not in Isg. Due to space

limitations, we cannot provide the full table with all the other between-sample comparisons, but we will note some other differences that were flagged as significant. Primarily, the AM TOKEN sample significantly differs from all three TiMBL models' samples as well as from the questionnaire data. AM DOUBLE ENTRIES and AM TOKEN samples also significantly differ from each other.

To conclude this part of the analysis, we will note that, even though AM and TiMBL work on slightly different principles, they still produce fairly accurate results, even when the dataset is reduced by half with the activation of the 'imperfect memory' parameter. By 'fairly accurate', we mean that the output of the models replicates the proportions found in the corpus as well as the behaviour of native speakers of Croatian, which we will analyse below.

6.4.2.2. Analysis 2: Native speakers' behaviour

The questionnaire data obtained in this study were not used only for comparison with the two other sources of data; rather, we also analyse them on their own to see whether they are comparable to the results obtained in the previous two studies described above. We have seen that BAND was a highly significant predictor in both of the previous studies, where acceptability ratings were involved. Here we are interested to see whether the position of a lexeme within a certain band can predict speakers' actual choice to the same extent. We are also testing for various social variables (gender, dialect etc.). The null hypothesis (H_0) in this case can be formulated as: *there is no relationship between the percentage of chosen forms and Bands*.

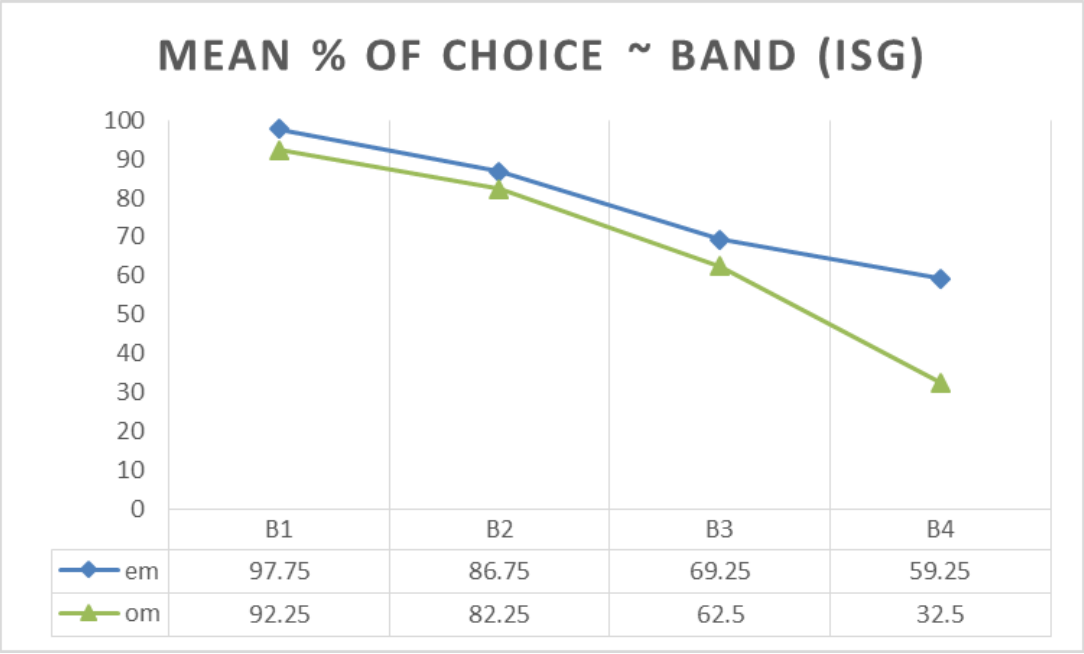
Besides speakers' CHOICE as the main dependent variable, we were able to measure two additional variables in this questionnaire, namely the time it took the respondents to answer (*response time*, coded as RT) and the level of confidence in their answer (coded as CONFIDENCE). We are interested to see whether BAND has a similar effect on these measurements as it does on CHOICE. Spearman's correlation coefficients, presented in Table 35 below, show a correlation of all three variables with BAND.

Table 33. Results of correlation testing

			Dependent variables		
Case	**Independent variable**		**CHOICE**	**RT**	**CONFIDENCE**
Isg	BAND	Spearman's rho p-value (2-tailed) N	.072 < .001 5,946	.188 < .001 6,085	-.225 < .001 3,377
Gpl	BAND	Spearman's rho p-value (2-tailed) N	-.037 < .001 8,809	.089 < .001 9,126	-.178 < .001 5,074

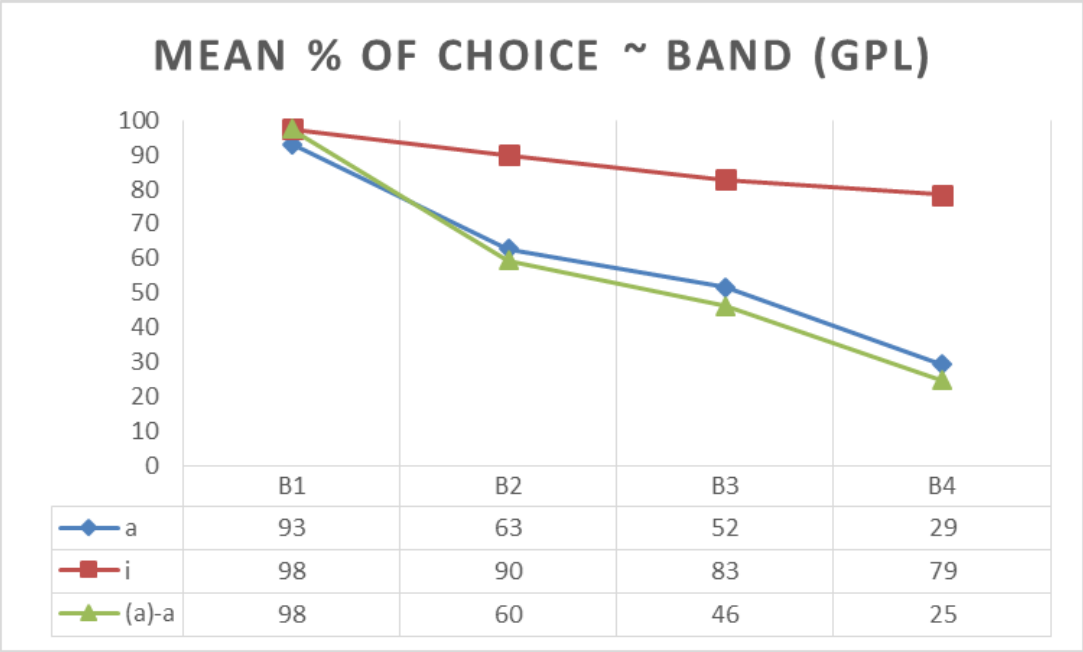
Let us first examine what kind of an effect BAND has on speakers' CHOICE. Figures 6 and 7 show how the choice of a form in the respective cases changes as we move between bands. A single line in the chart represents all lexemes with the same dominant form in the corpus (i.e. Bands 1-4).[165]

Figure 6. Mean percentages of choice of Isg -om and -em in different bands



[165] For instance, the starting point of the blue line in Figure 6 encompasses the four lexemes with absolute domination of -em (*slučaj, cilj, znanac, kraj*), whereas the starting point of the green line includes *Senj, šverc, Polančec,* and *keš*, where -om is absolutely dominant.

Figure 7. Mean percentages of choice of Gpl -a, -i, and a-a in different bands



In both charts we can see that a decrease in the percentage of a particular form in the corpus is matched by a decrease in speakers' preference for that form as well. However, it is also visible that the lines do not decrease in the same manner, especially in Gpl. Namely, if we take a look at Figure 7, we can see that in Bands 3 and 4 the Gpl endings -a and a-a have very low choice rates. In other words, even though those endings are the dominant endings for the respective lexemes, speakers still choose some other ending (usually -i) more often. Why does this happen? We offer a possible explanation below.

In the previous questionnaire study we have seen the emergence of what we termed family effects. Hoping that the same effects would arise in the present study as well, we included several examples of what we term *false pairs* into the questionnaires. These are defined as items that belong to the same phonological family (e.g. end in the same consonant cluster or same VC combination) but have inverse distributions of the respective endings. They are presented in Table 36 below. The highlighted proportions are those that do not match the corpus pattern.

Table 34. Comparison of corpus and questionnaire results for 'false pairs'

	False pairs	Corpus distributions (hrWaC14)	Questionnaire responses
Isg (% om: % em)	temelj	1: 99	16: 84
	krpelj	85: 15	68: 32
	tuš	3: 97	12: 88
	Bush	76: 24	77: 23
	Lošinj	18: 82	36: 64
	Rovinj	58: 42	**39: 61**
	staž	48: 52	48: 52
	kolaž	54: 46	**30: 70**
	vic	33: 67	46: 54
	šlic	78: 22	**41: 59**
	roštilj	13: 87	22: 78
	štambilj	51: 49	**21: 79**
Gpl (% a: % i: % a-a)	boljka	1: 99: 0	3: 96: 1
	školjka	0: 23: 77	**1: 69: 30**
	maska	2: 92: 2	3: 96: 1
	daska	1: 7: 92	1: 23: 76
	hrenovka	0: 93: 7	1: 90: 9
	olovka	0: 15: 85	1: 28: 71
	dagnja	0: 84: 16	2: 94: 4
	lignja	0: 47: 53	**3: 51: 46**
	praćka	4: 67: 29	1: 89: 10
	voćka	0: 19: 81	**1: 62: 37**
	kasta	53: 47: 0	**26: 74: 0**
	pasta	40: 60: 0	26: 74: 0

This table reveals the most likely reason why the lines in the above charts do not run parallel to each other. In situations where a certain lexeme strongly prefers one form over the other (Bands 1 and 2), the speaker will also tend to choose that form, which is a pure frequency effect. However, with lexemes where there is a lesser degree of dominance (Bands 3 and 4), the speaker will not necessarily choose the more frequent form for that particular lexeme, but rather turn to a form that is more common for the whole phonological family. We call these effects family effects. In other words, frequency effects and family effects both affect speakers' choice, but they come into play at different points of the relative proportions scale.

The correlation of the latter two variables (RT and CONFIDENCE) with BAND is graphically represented in Figures 8 and 9 below. Reaction time is often used as a

measure in psychological experiments, where it has been shown to be subject to a large number of confounding factors, including speed of reading and writing, font size and colour, manner of presentation etc. (see Schmid (2013)). Since we did not control for many of these factors in the present survey, we will not proceed with analysing this effect statistically. When it comes to the latter variable, CONFIDENCE, the questions involving it were defined as optional in the original survey, so they did not receive as many answers as the rest of the survey (around 50% of the total number of responses). Hence no reliable generalizations can be made for this variable either. Both correlations are mentioned only for illustrative purposes.

Figure 8. Mean response times in different bands

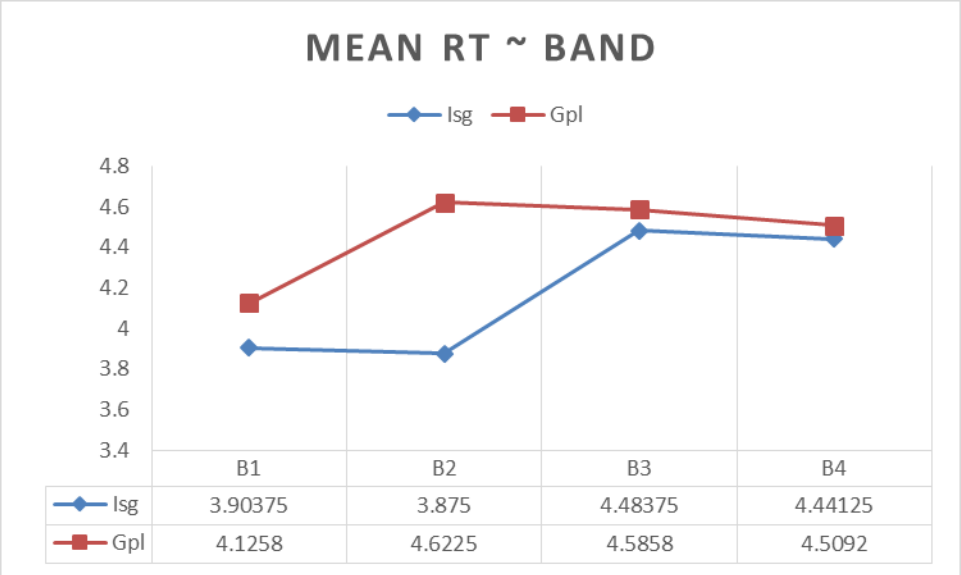
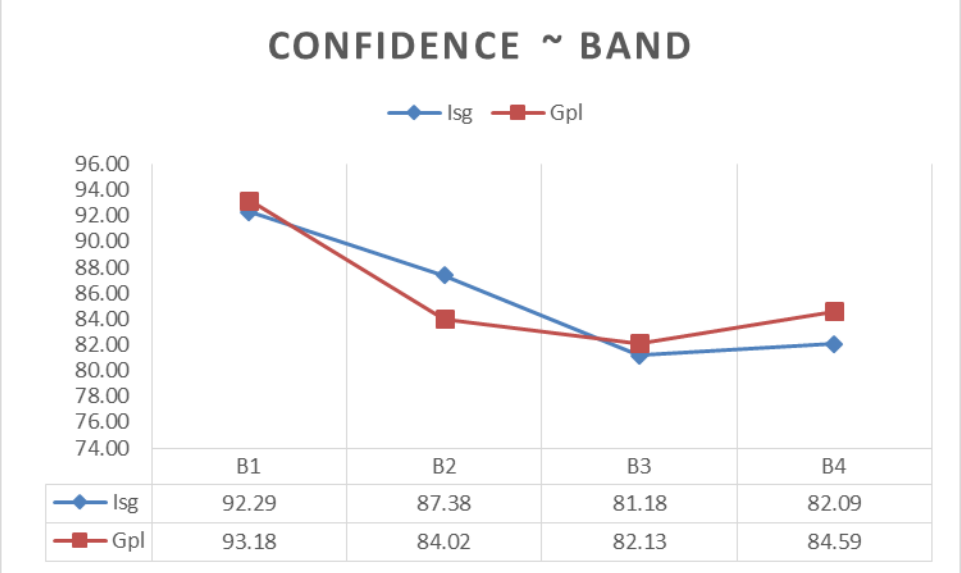


Figure 9. Speakers' mean confidence rates (%) in different bands



Similarly to the percentage of answers, we can see that the values of both of these variables also change as we move between bands. The response time of speakers increases ($\rho > 0$) as the proportions of the two endings level out, whereas their level of confidence decreases ($\rho < 0$). For instance, we can see that the RT for lexemes in Band 1 was almost half a second quicker than for lexemes in the other bands. Furthermore, Bands 3 and 4 were also the bands where the greatest number of time-outs occurred, e.g. 27 respondents did not provide the answer for the word *breskva* (B3) within the 8-second limit, just as 16 respondents failed to do for the word *praćka* (B3). Baayen, Milin & Ramscar (2016, p. 14) point out that an increase in entropy causes processing speed to decrease and we have already determined that entropy increases as we approach equiprobability. The results presented in this section all show that there is a variety of processes going on in that area, which slow down morphological processing. We explain these processes in the following section.

6.4.2.3. Proposed model

Based on the results so far, it seems that the speakers need to have the following information when making choices between doublets: 1) relative proportions of the two forms, which are expressed by the concept of bands. Based on these proportions, they are also able to derive: 2) the dominant ending for a particular lexeme and 3) the dominant ending of the whole phonological family. The choice will then depend on a combination of these three pieces of information. To check this claim, we ran a nominal regression model, which tried to predict the choice of a particular ending based on these three pieces of information as well as some demographic variables. The results of the full model are given in Table 37 below. Significant variables (at .05 level) are marked with an asterisk.

Table 35. Results of nominal regression (Dependent variable: CHOICE)

	Effect	Likelihood Ratio Tests			Model performance
		Chi-Square	Degrees of freedom	p-value	
Isg	Gender	.004	1	.949	Nagelkerke R^2 = .346
	Region	1.826	1	.177	
	Dialect	1.287	1	.257	
	Education	.013	1	.908	
	Dominant_ending*	668.631	1	.000	
	Family_dom_ending*	358.765	1	.000	
	Band*	11.723	3	.008	
Gpl	Gender	2.777	2	.249	Nagelkerke R^2 = .653
	Region	1.527	2	.466	
	Dialect	3.975	2	.137	
	Education	2.743	2	.254	
	Dominant_ending*	2881.171	4	.000	
	Family_dom_ending*	108.473	2	.000	
	Band*	944.893	6	.000	

Based on the results from the table above, which identifies three highly significant variables, we propose the following model for the processing of doublets: in cases where the dominant ending for a particular lexeme and the dominant ending of the whole phonological family are the same, the choice is simple: that ending will also be chosen by the speaker. However, in cases where they are different, the situation is much more complicated. The choice will depend on the level of dominance of one form for a particular lexeme. In situations of strong dominance (B1 and B2) the speakers are still more likely to choose the lexeme-dominant ending, but where the dominance of a particular ending weakens (B3 and B4) they will tend to choose the family-dominant ending. This model bears resemblance to Bybee's (1985) Network Model, in which she claims that in lexical access, there is competition between lexical strength and connection strength, both measures being calculated on the basis of type frequency. As we have said above, our labels LEXEME-DOMINANT ENDING and FAMILY-DOMINANT ENDING are derived from token frequencies. Racz, et al. (2014, p. 8) also argue for a hybrid model, "in which instance-based processing and reasoning sits alongside more abstract structures, and in which both types of processes may be jointly operative." In this sense, LEXEME-DOMINANT ENDING and FAMILY-DOMINANT ENDING are indeed more abstract concepts than pure frequency measures.

6.4.2.4. Analysis 3: forced choice versus ratings

We would like to make one further point about the different response methods that were used in the last two questionnaire studies. Arppe & Järvikivi (2007a) also conducted two separate questionnaire studies, one forced-choice and the other using acceptability ratings. The results of both studies converged well with the corpus data, but they concluded that the forced-choice data and the rating data do not necessarily correlate universally because the two types of tasks reflect two different processes (2007a, p. 150): forced choice measuring production (or preferences) and acceptability measuring introspection (or degree of contextual appropriateness). Their results showed that highly frequent linguistic items are likely to be both acceptable and preferable, whereas rare items might be dispreferred but they are not categorically unacceptable (2007a, p. 153). Similar to this, in Albright & Hayes' (2003) nonce word experiments on English past tense, the rate of irregular responses in an elicitation task was 8.7%, but in a separate task the mean rating of the irregulars was 4.22 (7-point scale), compared to 5.75 for regulars. Diessing, Filipović Đurđević, & Zec (2009) obtained an even better correspondence – variants that were more common in the corpus exhibited a higher percentage of participants' responses, faster processing, and higher acceptance rates; however, the authors do admit that the acceptability results are the least conclusive as the differences between the two sets were the least drastic. Haber (1976) obtained a highly significant, although not perfect, correlation ($r = .86$) between English past tense forms that speakers consider correct and the ones that they actually use.

Since one subset of the material (Isg) was the same in questionnaire studies 2 and 3, we were able to directly compare the outputs of the two methods used. Although this was not our intention, several lexemes from the second questionnaire study also appeared in the third study. Their mean results from both surveys are presented in Table 38 below.

Table 36. Comparison of results for the identical material in questionnaire studies 2 and 3

Isg forms	Corpus proportions (hrWaC14)	Q2 (Mean rating out of 7)	Q3 (% of choice)
križem	95%	6.59	80.2%
križom	5%	2.76	19.8%
volejem	67%	5.28	53.9%
volejom	33%	4.43	46.1%
trofejem	60%	5.08	50.0%
trofejom	40%	4.68	50.0%
stažem	52%	4.38	52.4%
stažom	48%	5.62	47.6%
esejem	41%	5.36	60.2%
esejom	59%	4.21	39.8%
Bushem	24%	3.77	22.9%
Bushom	76%	5.85	77.1%
mjesecem	2%	3.06	13.6%
mjesecom	98%	6.23	86.4%
crtežem	2%	2.27	8.8%
crtežom	98%	6.56	91.2%
švercem	0%	2.33	8.5%
švercom	100%	6.59	91.5%

We can see a high degree of correspondence in the results, similar to the works referred to above. In all but one instance (*stažom*), the form that was rated as more acceptable by one sample of respondents was also selected to a much greater proportion by another sample. We refrain from making any large-scale conclusions related to linguistic theory based on such a small sample, but this almost perfect correspondence is nevertheless interesting. However, as pointed out by Antti Arppe, "even if the measures do converge, this does not necessarily mean that they reflect exactly the same underlying cognitive linguistic structures or processes" (Arppe, Gilquin, Glynn, Hilpert, & Zeschel, 2010, p. 5). We leave this matter open for future debate.

6.4.3. Discussion

The goals of this study were several: first, by comparing the results of the questionnaire study to both the output of the computational models and the corpus

data, we tried to determine whether the models are psychologically realistic, which is often claimed by their authors. Neither the models (the majority of them) nor the questionnaire samples were significantly different from the control sample (the corpus data), which shows that distributions of the two forms in everyday language form an important part of the decision-making process of both the memory-based models and native speakers.

The second goal of this study was to see whether the effect that became apparent in the previous surveys would arise in the present study as well, regardless of the change in the method used (acceptability rating vs. forced-choice). Hence we re-used the concept of BAND, which expresses the level of dominance of one ending over the other. BAND was once again one of the most important predictor variables of speakers' choice. However, the results also suggest another effect coming into play in certain bands. In cases of lexemes where there is a clear dominance of one ending (i.e. in Bands 1 and 2), that ending will also be selected by the majority of speakers, i.e. individual **frequency effects** are very strong in these bands. However, as we move to other bands, where there is more fuzziness in the proportions, the frequency effects will become less strong and will instead be overruled by **family effects** – the majority of speakers will choose an ending that is dominant for the whole phonological family, even though it might not be the same ending that is dominant for that particular lexeme. Based on the results from Figure 8, it can also be argued that frequency effects are less time-consuming, or put differently, trigger a reaction more quickly than family effects. All of these results point to the conclusion that more fuzziness in the language leads to more hesitance and uncertainty in speakers. This is in line with Berg (2011, p. 42), who argues that "[morphophonological] variation incurs mental cost." However, it is important to note that this is not the same as saying that it completely stops language processing and production, which is an argument often used by Croatian linguists in order to justify their stigmatization of one of the forms.

Chapter 7. Conclusion

Having reached the end of the dissertation, in this section we will revise some of the theoretical claims presented in the first part of the work (Chapters 1 to 3) and see how well they were supported (or not) by the empirical data in the second part (Chapters 4 to 6). The methodological approach we adopted in this dissertation can be termed "pluralistic" as it involves comparing and contrasting evidence obtained by various methods instead of relying solely on one type of data. The data used in this work came from corpora, computer models as well as questionnaires.

We mentioned on numerous occasions that this work was written in the usage-based framework, which claims that language structure emerges from language use. So before starting any kind of a structural analysis of doublets, it was necessary to determine how they are used in Croatian. This was the purpose of Chapter 4. The data from this chapter also serves as the basis for following chapters since concepts used in Chapters 5 and 6, such as relative proportions or bands, are derived solely from usage data.

In Section 1.2, we mentioned R. Ellis' (1999, p. 467) views on two possible starting points in any study of free variation: 1) we can assume that the variation is systematic until shown otherwise or 2) we can assume it to be free until it is shown to be systematic. Although we have taken 2) as the starting point of our research, we have not found enough evidence to disprove either of those two hypotheses. Rather, the results presented in Chapter 4 seem to suggest that the variation present in Croatian doublets is both systematic and non-systematic, which means that part of the variation can be accounted for, but part of it cannot. This is also the position assumed by Labov (1978) and other sociolinguists following him.

The comparison of rules found in grammar books of Croatian and examples of usage from corpora presented in the same chapter has also showed that reference manuals of Croatian are severely outdated – the rules they define are full of exceptions,[166] but more importantly, the rules are not based on patterns present in real

[166] Krott, Baayen, & Schreuder (2001, p. 56) argue that when you have a great number of exceptions to a rule, you cannot talk about a real rule, but rather about a tendency.

language. The patterns present in Croatian doublets are a matter of degree rather than a question of 'regular' versus 'irregular' forms (cf. Mirković, Seidenberg, & Joanisse (2011, p. 650), who label the inflectional system of Serbian as 'quasi-regular').

In Section 1.2 we also presented Kilgariff's (2005, p. 264) view of how two phenomena in language can be associated – they can be Random, Arbitrary, Motivated, or Predictable. Kilgariff claims that nothing in language is ever random. Our corpus analysis from Chapter 4 has shown that Croatian doublets can not be placed on the other end of the scale (Predictable) either. Left with only two options, we would argue that they are in fact Motivated, by factors which we explain below.

Apart from determining a numerical distribution of the various doublet forms in present-day Croatian, this work also attempted to make some inferences about speakers' mental grammars, which was the goal of Chapters 5 and 6. First of all, Dąbrowska (2008b, p. 948) argues that "mental grammars are governed by different principles than formal linguistic grammars." One of the main arguments by Krott, Baayen, & Schreuder (2001, p. 80) is that morphological rules are grounded in analogy. The exemplar-based modelling we performed in Chapter 5 was highly successful in predicting the desired ending while using a limited number of variables and not very complicated algorithms, one of them being analogy as well. They were able to provide a form where the rules provided none, i.e. when rules from different manuals conflicted.

As a reminder, these exemplar-based models deny that speakers store any generalizations at all, but rather that grammatical knowledge is equated solely with stored instances. As pointed out by Dąbrowska (2008b, p. 933), "these claims, if true, have far-reaching implications not just for description, but also for linguistic theory, in that they significantly weaken most arguments based on the principle of economy" – a principle we have seen invoked by some Croatian linguists as well.

The main argument of this work is that, besides the word itself and its inflected forms, speakers also store frequency information in their mental grammars. Haber (1976, p. 231) concludes that each form is treated as a separate lexical item, "accompanied by a variation figure in each speaker's head." The results of our questionnaire studies, presented in Chapter 6, have shown similar results, i.e. that frequency-related knowledge about a form is indeed a part of speakers' grammatical

knowledge. This frequency-related knowledge is derived from the input speakers constantly receive. Such a view relates mental grammars to the concept of mental corpora, as introduced by Taylor (2012).

However, what kind of frequency information are we talking about? Some authors will say it is absolute numbers (i.e. token frequency), some say it is relative numbers etc. (see discussion in Section 1.2). Divjak & Arppe (2013, p. 245) believe that "it is irrelevant what speakers keep track of, as long as they keep track of something." In cases of doubletism, however, it is reasonable to say that what speakers actually store are proportions of the two forms, expressed in the form of percentages, ratios or a related number. These proportions were reflected in speakers' performance on subsequent rating and forced-choice tasks, in such a way that form X, which is less frequent than form Y, was also shown to be less acceptable than form Y. However, levels of (un)acceptability of forms X and Y varied in accordance to the relative proportions of the two forms. These results lead us the conclusion that mental grammars are gradient and that this gradience is transferred to usage.

However, frequencies (or more accurately, relative proportions) were not the only factor influencing speakers' behaviour. Rather, as argued by Roeper (2011, p. 24), "notions like similarity are where the real mental talents are hidden." Our studies also revealed what we termed family effects – speakers not necessarily opting for the more frequent form, but rather for the form that is more in line with the behaviour of phonologically similar forms. As pointed out by N. Ellis (2012, p. 18), input is "incomplete, uncertain, and noisy," so speakers need to perform some sampling and inferencing. We believe that the clustering of exemplars into phonological families is one way of compensating for the imperfections in the input.

Moreover, the emergence of these family effects also seems to suggest, as argued by Fehringer (2004, p. 306), that these forms are stored directly in the lexicon rather than derived by rule. "The fact that the preference for alternative variants is often a matter of degree argues against a 'rule-versus-list' dichotomy in which some forms are lexically listed and others generated by rule" (such as the Dual Mechanism model). Moreover, "the substantial influence of token frequency[167] on this preference

[167] Even though our studies have not revealed effects of token frequency per se, it is important to remember that the relative proportions are actually calculated from token frequencies.

is strong evidence for a model that takes full account of the concept of lexical strength" (Fehringer, 2004, p. 326). The model the author is referring to is Bybee's (1985) Network Model.

Finally, we have seen that the majority of Croatian grammarians want to eliminate doublets from the language by any means necessary, primarily by introducing a functional distinction between them, considering them, to use Babić's (1962, p. 62) term, a *deadweight* in language. However, Berg (2011, p. 42) argues that allomorphy (and we dare to say the same for doubletism) "is a tolerable, if not beneficial property of language. Otherwise, it would not occur in the first place." It is high time for linguists to stop thinking about grammar in terms of 'either-or' distinctions and adopt a more gradated view. We hope to have shown that our mental grammars are gradated, so we believe that formal descriptions of grammars of a language should accurately represent this feature.

All in all, as a general conclusion, we will use Dąbrowska's (2008b, p. 931) words, who says that "linguistic competence is shaped by performance factors such as the frequency of occurrence of a particular form and the speaker's perception of its similarity to other forms in his or her mental grammar." All of these factors show, to use a term by Milin, et al. (2009, p. 234), "structured lexical connectivity" of mental grammars.

References

Aarts, J. (1991). Intuition-based and observation-based grammars. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics: Studies in honor of Jan Svartvik* (pp. 30-43). New York and London: Longman.

Adger, D. (2006). Combinatorial variability. *Journal of Linguistics, 42*, 503-530.

Adli, A. (2005). Gradedness and Consistency in Grammaticality Judgments. In S. Kepser, & M. Reis (Eds.), *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives* (pp. 5-23). Berlin and New York: Mouton de Gruyter.

Aha, D., Kibler, D., & Albert, M. (1991). Instance-based learning algorithms. Machine Learning,. *Machine Learning, 6*, 37-66.

Albright, A. (2009). Modeling analogy as probabilistic grammar. In J. P. Blevins, & J. Blevins (Eds.), *Analogy in Grammar: Form and Acquisition* (pp. 187-212). Oxford: Oxford University Press.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90*, 119-161.

Alerić, M., & Gazdić-Alerić, T. (2013). *Hrvatski u upotrebi.* Zagreb: Profil.

Anderwald, L., & Szmrecsanyi, B. (2009). Why grammar is real: a usage-based perspective on patterns. In L. Eckstein, & C. Reinfandt (Eds.), *Anglistentag 2008 – Tübingen – Proceedings* (pp. 403-414). Trier: Wissenschaftlicher Verlag Trier.

Andrews, A. D. (1990). Case structures and control in Modern Icelandic. In J. Maling, & A. Zaenen (Eds.), *Modern Icelandic syntax* (pp. 187-234). New York: Academic Press.

Anić, V. (1973). Morfološke inovacije u hrvatskosrpskom jeziku. *Radovi Sveučilišta u Zadru, Razdio lingvističko-filološki, 7*, 7-15.

Anić, V. (2009). *Naličje kalupa: Sabrani spisi.* (I. Marković, Ed.) Zagreb: Disput.

Antilla, A. (1995). *Deriving variation from grammar: a study of Finnish genitives.* Ms, Stanford University.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*(3), 263-308.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67-82.

Aronoff, M. (1976). *Word Formation in Generative Grammar.* Cambridge, MA: MIT Press.

Aronoff, M. (1994). *Morphology by Itself* . Cambridge, MA: MIT Press.

Arppe, A. (2009). Linguistic choices vs. probabilities - how much and what can linguistic theory explain? In S. Featherston, & S. Winkler (Eds.), *The Fruits of Empiricial Linguistics. Volume 1: Process* (pp. 1-24). Berlin: Mouton de Gruyter.

Arppe, A., & Abdulrahim, D. (2013). Converging linguistic evidence on two flavors of production: The synonymy of Arabic COME verbs. *Second Workshop on Arabic Corpus Linguistics.* University of Lancaster, UK, 22-26 July, 2013.

Arppe, A., & Järvikivi, J. (2007a). Every method counts - combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory, 3*(2), 131-160.

Arppe, A., & Järvikivi, J. (2007b). Take empiricism seriously! In support of methodological diversity in linguistics. *Corpus Linguistics and Linguistic Theory, 3*(1), 99-109.

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2010). Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. *Corpora, 5*(1), 1-27.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*, 436-461.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics, 11*, 295-328.

Baayen, R. H., Endersen, A., Janda, L. A., Makarova, A., & Nesset, T. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics, 37*(3), 253-291.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*.

Baayen, R., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438-482.

Babić, S. (1962). Duljina ili dužina? *Jezik, 11*, 62-63.

Babić, S. (1979). Instrumental i pridjevi imenica na -io. *Jezik, 27*, 87-89.

Babić, S. (2006). Zatvorenost imenica i-sklonidbe. *Suvremena lingvistika, 62*, 141-150.

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics, 46*, 273-330.

Baker, M. (1985). The Mirror Principle and Morphosyntactic Explantion. *Linguistic Inquiry, 16*, 373-415.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72*, 32-68.

Bates, E., & MacWhinney, B. (1988). What is functionalism? *Papers and Reports on Child Language Development, 27*, 137-152.

Bauer, L. (1993). Progress with a corpus of New Zealand English and some early results. In C. Souter, & E. Atwell (Eds.), *Corpus-based Computational Linguistics* (pp. 1-10). Amsterdam/Atlanta: Rodopi.

Bauer, L. (2006). Competition in English word-formation. In A. van Kemenade, & B. Los (Eds.), *The Handbook of the History of English* (pp. 177-198). Malden, MA: Blackwell.

Bauer, L. (2014). Grammaticality, acceptability, possible words and large corpora. *Morphology, 24*, 83-103.

Beals, K., & al. (Eds.). (1994). *Papers from the 30th regional meeting of the Chicago Linguistic Society.* Chicago: University of Chicago Press.

Berg, T. (2011). A diachronic frequency account of the allomorphy of some grammatical markers. *Journal of Linguistics, 47*, 31-64.

Berko, J. (1958). The Child's Learning of English Morphology. *Word, 14*, 150-177.

Bermel, N. (1993). Sémantické rozdíly v tvarech českého lokálu. *Naše řeč, 76*(4), 92-98.

Bermel, N., & Knittl, L. (2012). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory, 8*(2), 241-275.

Bermel, N., Knittl, L., & Russell, J. (2015). Morphological variation and sensitivity to frequency of forms among native speakers of Czech. *Russian Linguistics, 39*, 283-308.

Berruto, G. (2004). The problem of variation. *The Linguistic Review, 21*, 293-322.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory, 9*(1), 9-37.

Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1286-1304). Berlin: Walter de Gruyter.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371-405.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics, 42*, 531-573.

Bod, R. (1995). *Enriching linguistics with statistics: Performance models of natural language. Dissertation.* Amsterdam: ILLC, Universiteit van Amsterdam.

Bod, R. (1998). *Beyond Grammar. An Experience-based Theory of Language.* Stanford, CA: CSLI publications.

Boersma, P. (1997). How we learn variation, optionality, probability. *Proceedings of the Institute of Phonetic Science, 21*, 43-58.

Boersma, P. (2004). *A stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments.* Ms, University of Amsterdam.

Bošnjak, T. (2005). Gramatička i semantička kolebljivost dvosložnih imenica sa sufiksom -ež. *Rasprave Instituta za hrvatski jezik i jezikoslovlje, 31*, 1-18.

Boyd, J. K., & Goldberg, A. (2011). Learning what NOT to say: The role of statistical preemption and categorization in A-adjective production. *Language, 87*(1), 55-83.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston, & W. Sternefeld (Eds.), *Roots: Linguistics in Search of Its Evidential Base* (pp. 77-96). Berlin: Mouton de Gruyter.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the Dative Alternation. In G. Boume, I. Kraemer, & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 69-94). Amsterdam: Royal Netherlands Academy of Science.

Brown, D., Tiberius, C., & Corbett, G. G. (2007). The alignment of form and function: Corpus-based evidence from Russian. *International Journal of Corpus Linguistics, 12*(4), 511-534.

Brozović, D. (1976). O suvremenoj morfološkoj normi hrvatskoga jezičnog standarda i o morfološkim značajkama standardne novoštokavštine uopće. *Jezik, 24*, 1-12, 41-49, 109-114.

Bybee, J. (2010). *Language, Usage and Cognition.* Cambridge: Cambridge University Press.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form.* Amsterdam: John Benjamins.

Bybee, J. L. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425-455.

Bybee, J. L. (2001). *Phonology and language use.* Cambridge: Cambridge University Press.

Bybee, J. L. (2007). *Frequency of Use and the Organization of Language.* Oxford: Oxford University Press.

Bybee, J. L., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins.

Bybee, J. L., & Newman, J. (1995). Are affixes more natural than stem changes? *Linguistics, 33*, 633-654.

Caines, A. (2012). You talking to me? Corpus and experimental data on the zero auxiliary interrogative in British English. In S. T. Gries, & D. Divjak (Eds.), *Frequency Effects in Language Learning and Processing* (pp. 177-206). Berlin: Mouton de Gruyter.

Cameron-Faulkner, T., & Carstairs-McCarthy, A. (2000). Stem Alternants as Morphological Signata: Evidence from Blur Avoidance in Polish Nouns. *Natural Language and Linguistic Theory, 18*, 813-835.

Cappellaro, C. (2013). Overabundance in diachrony: a case study. In S. Cruschina, M. Maiden, & J. C. Smith (Eds.), *The Boundaries of Pure Morphology: Diachronic and Synchronic Perspectives* (pp. 209-220). Oxford: Oxford University Press.

Carstairs, A. (1987). *Allomorphy in inflexion.* London: Croom Helm.

Cassanto, D. (2009). Embodiment of abstract concepts: good and bad in right- and left-handers. *Journal of Experimental Psychology: General, 138*, 351-367.

Chafe, W. (1992). The Importance of Corpus Linguistics to Understanding the Nature of Language. In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (pp. 79-97). Berlin: Mouton de Gruyter.

Chandler, S. (1993). Are rules and modules really necessary for explaining language? *Journal of Psycholinguistic Research, 22*, 593-606.

Chandler, S. (2002). Skousen's analogical approach as an exemplar-based model of categorization. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.).

Chandler, S. (2009). *Analogical Modeling: A unified account of the regular-irregular dissociations.* Moscow, ID: University of Idaho.

Chandler, S. (2010). The English past tense: Analogy redux. *Cognitive Linguistics, 21*(3), 371-417.

Chapman, D., & Skousen, R. (2005). Analogical Modeling and Morphological Change: The Case of the Adjectival Negative Prefix in English. *English Language and Linguistics, 9*, 333-357.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton.

Chumakina, M., Hippisley, A., & Corbett, G. G. (2004). Historical changes in the Russian lexicon: a case of alternating suppletion. *Russian Linguistics, 2004*, 281-315.

Clark, E. V. (1987). The Principle of Contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 157-193). Hillsdale, NJ: Erlbaum.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Corbett, G. G. (2009). Canonical Inflectional Classes. *Proceedings of the Decembrettes 6*, (pp. 1-11).

Corbett, G. G. (2010). Agreement in Slavic. *Glossos, 10*, 1-61.

Corbett, G. G. (2010). Canonical derivational morphology. *Word structure, 3*(2), 141-155.

Cowart, W. (1997). *Experimental syntax: Applying direct object methods to sentence judgments.* Thousand Oaks, CA: Sage Publications.

Cruse, A. D. (1986). *Lexical Semantics.* Cambridge: Cambridge University Press.

Cruse, A. D. (2000). *Meaning in Language: An Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.

Culbertson, J., & Gross, S. (2009). Are linguists better subjects? *British Journal for the Philosophy of Science, 60*, 721-736.

Cummins, G. M. (1995). Locative in Czech: -u or -e: Choosing locative singular endings in Czech nouns. *Slavic and East European Journal, 39*, 241-260.

Ćavar, D., & Brozović Rončević, D. (2012). Riznica: The Croatian Language Corpus. *Prace filologiczne, LXIII*, 51-65.

Ćorić, B. (2007). Sinonimi i dubleti. *Naučni skup slavista u Vukove dane, 35*, 59-67.

Dąbrowska, E. (2004). *Language, Mind and Brain: Some Psychological and Neurological Constraints on Theories of Grammar* (Vol. 19). Edinburgh: Edinburgh University Press.

Dąbrowska, E. (2008a). Questions with long-distance dependencies: A usage-based perspectives. *Cognitive Linguistics, 19*(3), 391-425.

Dąbrowska, E. (2008b). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language, 58*, 931-951.

Dąbrowska, E. (2010). Naive vs. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review, 27*, 1-24.

Daelemans, W. (2002). A comparison of Analogical Modeling to Memory-Based Language Processing. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.). John Benjamins.

Daelemans, W., & van den Bosch, A. (2010). Memory-based learning. In A. Clark, C. Fox, & S. Lappin (Eds.), *Handbook of Computational Linguistics and Natural Linguistic Processing* (pp. 154-179). Oxford: Wiley-Blackwell Publishers.

Daelemans, W., Gillis, S., & Durieux, G. (1997). Skousen's analogical modeling algorithm: A comparison with lazy learning. In D. Jones, & H. Somers

(Eds.), *New Methods in Language Processing* (pp. 3-15). Manchester, London: University College Press.

Daelemans, W., van den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning, 34*, 11-41.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2007). *TiMBL: Tilburg Memory-Based Learner, version 6.1, Reference Guide.* Tilburg: ILK Research Group Technical Report Series no. 07-07.

Derwing, B., & Skousen, R. (1989). Real-Time Morphology: Symbolic Ruls or Analogical Networks. In *Proceedings of the Fifteenth Annual Meeting of the Berkeley Linguistics Society* (pp. 48-62). Berkeley: Berkeley Linguistics Society.

Derwing, B., & Skousen, R. (1994). Productivity and the English past tense: Testing Skousen's analogy model. In S. Lima, R. Corrigan, & G. Iverson (Eds.), *The Reality of Linguistic Rules* (pp. 193-218). Amsterdam: John Benjamins.

Devitt, M. (2010). Linguistic intuitions revisited. *British Journal for the Philosophy of Science, 61*, 833-865.

Diesing, M., Filipović Đurđević, D., & Zec, D. (2009). Clitic placement in Serbian: Corpus and experimental evidence. In S. Featherston, & S. Walker (Eds.), *The Fruits of Empirical Linguistics* (pp. 61-75). Berlin: Mouton de Gruyter.

Divjak, D. (2006). Ways of Intending. Delineating and Structuring Near-Synonyms. In S. T. Gries, & A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics* (Vols. 2: The syntax-lexis interface, pp. 19-56). Berlin: Mouton de Gruyter.

Divjak, D. (2008). On (in)frequency and (un)acceptability. In B. Lewandowska-Tomaszczyk (Ed.), *Corpus linguistics, computer tools and applications – State of the art* (pp. 213-233). Frankfurt: Peter Lang.

Divjak, D. (2016). The Role of Lexical Frequency in the Acceptability of Syntactic Variants: Evidence From that-Clauses in Polish. *Cognitive Science*, 1-29.

Divjak, D., & Arppe, A. (2013). Extracting prototypes from exemplars. What evidence do corpora contain for the representation of linguistic categories? *Cognitive Linguistics, 24*(2), 221-274.

Divjak, D., & Gries, S. T. (2006). Ways of trying in Russian. Clustering and comparing behavioral profiles. *Corpus Linguistics and Linguistic Theory, 2*(1), 23-60.

Divjak, D., & Gries, S. T. (Eds.). (2012). *Frequency Effects in Language Representation (volume 2).* Berlin: Mouton de Gruyter.

Divjak, D., Dąbrowska, E., & Arppe, A. (2016). Machine Meets Man: Evaluating the Psychological Reality of Corpus-based Probabilistic Models. *Cognitive Linguistics, submitted*.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2016). A Functional Theory of Gender Paradigms. In F. Kiefer, J. Blevins, & H. Bartos, *Morphological Paradigms and Functions.* Leiden: Brill.

Eddington, D. (2002). A comparison of two analogical models: Tilburg Memory-Based Learner versus Analogical Modeling. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.).

Eddington, D. (2004). Issues in Modeling Language Processing Analogically. *Lingua, 114*, 849-871.

Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics, 28*, 105-144.

Ellis, N. C. (2002). Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*, 143-188.

Ellis, N. C. (2009). Optimizing the Input: Frequency and Sampling in Usage-based and Form-focussed Learning. In M. Long, & C. Doughty (Eds.), *Handbook of Language Teaching* (pp. 139-158). Oxford: Blackwell.

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition and use? In S. T. Gries, & D. Divjak (Eds.), *Frequency Effects in Language Learning and Processing* (pp. 7-34). Berlin: Mouton de Gruyter.

Ellis, R. (1999). Item versus System Learning: Explaining Free Variation. *Aplied Linguistics, 20*(4), 460-480.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141-172.

Elzinga, D. (2006). English adjective comparison and analogy. *Lingua, 116*, 757-770.

Endresen, A. (2014). *Non-Standard Allomorphy in Russian Prefixes: Corpus, Experimental, and Statistical Exploration.* PhD dissertation, University of Tromsø.

Erker, D., & Guy, G. R. (2012). The role of lexical frequency in syntactic variability: variable subject personal pronoun expression in Spanish. *Language, 88*(3), 526-558.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37-64.

Fanselow, G., & Frisch, S. (2006). Effects of processing difficulty on judgments of acceptability. In G. i. grammar, G. Fanselow, C. Fery, M. Schlesewsky, & R. Vogel (Eds.), *Gradience in grammar* (pp. 291-316). Oxford: Oxford University Press.

Featherston, S. (2005b). The Decathlon Model. In S. Kepser, & M. Reis (Eds.), *Linguistic Evidence: Theoretical, Empirical, and Computational Perspectives* (pp. 187-208). Berlin/New York: Mouton de Gruyter.

Featherston, S. (2007). Data in generative grammar. The stick and the carrot. *Theoretical Linguistics, 33*(3), 269-318.

Featherston, S. (2008). Thermometer judgements as linguistic evidence. In C. M. Riehl, & A. Rothe (Eds.), *Was is linguistische Evidenz?* (pp. 69-89). Aachen: Shaker Verlag.

Featherston, S. (2009). Relax, lean back and be a linguist. *Zeitschrift für Sprachwissenschaft, 28*(1), 127-132.

Fehringer, C. (2004). How Stable are Morphological Doublets? A Case Study of /ə/ ~ ø variants in Dutch and German. *Journal of Germanic Linguistics, 16*(4), 285-329.

Fehringer, C. (2011). Allomorphy in the German genitive: A paradigmatic account. *Zeitschrift für Germanische Linguistik, 39*, 90-112.

Feldman, L. B., & Fowler, C. A. (1987). The inflected noun system in Serbo-Croatian: Lexical representation of morphological structure. *Memory and Cognition, 15*, 1-12.

Fillmore, C. (1992). Corpus Linguistics or Computer-aided Armchair Linguistics. In J. Svartvik (Ed.), *Directions in Corpus Linguistics* (pp. 13-38). Berlin: Mouton de Gruyter.

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes, 28*(1-2), 88-124.

Giegerich, H. (2001). Synonymy blocking and the Elsewhere Condition: Lexical morphology and the speaker. *Transactions of the Philological Society, 99*, 65-98.

Gilquin, G. (2010). Language production: A window to the mind? In H. Götzsche (Ed.), *Memory, Mind and Language* (pp. 89-102). Newcastle: Cambridge Scholars Publishing.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1), 1-26.

Gordon, P., & Alegre, M. (1999). Is there a dual system for regular inflections? *Brain and Language, 68*, 212-217.

Grant, D. A., Hake, H. W., & Hornseth, J. P. (1951). Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement. *Journal of Experimental Psychology, 42*(1), 1-5.

Grčević, M. (2006). Instrumental jednine imenica i-sklonidbe u hrvatskome književnom jeziku od 16. do 19. stoljeća. *Jezik, 53*, 50-59.

Grčević, M. (2007). Instrumental imenica i-vrste na -st u suvremenome hrvatskome književnom jeziku. *Jezik, 54*, 15-22.

Greenberg, J. D. (2008). *Language and Identity in the Balkans.* Oxford: Oxford University Press.

Gries, S. T. (1999). Particle movement: A cognitive and functional approach. *Cognitive Linguistics, 10*(2), 105-145.

Gries, S. T. (2002). Evidence in linguistics: Three approaches to genitives in English. In R. M. Brend, W. J. Sullivan, & A. R. Lommel (Eds.), *LACUS Forum XXVIII: What Constitutes Evidence in Linguistics?* (pp. 17-31). Fullerton: LACUS.

Gries, S. T. (2003). Towards a corpus-based identification of prototypical instances of contructions. *Annual Review of Cognitive Linguistics, 1*, 1-27.

Gries, S. T. (2005). Syntactic priming: a corpus-based approach. *Journal of Psycholinguistic Research, 34*(4), 365-399.

Gries, S. T., & Divjak, D. (Eds.). (2012). *Frequency Effects in Language Learning and Processing (volume 1).* Berlin: Mouton de Gruyter.

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics, 9*(1), 97-129.

Grondelaers, S., Speelman, D., & Geeraerts, D. (2008). National variation in the use of er "there". Regional and diachronic constraints on cognitive explanations. In G. Kristiansen, & R. Dirven (Eds.), *Cognitive Sociolinguistics: Language variation, cultural models, social systems* (pp. 153-203). Berlin/New York: Mouton de Gruyter.

Gudkov, V. (1988). Opservacije o variranju komparativnih oblika. *Naš jezik, 27*, 332-336.

Guerrero Medina, P. (Ed.). (2011). *Morphosyntactic alternations in English: Functional and cognitive perspectives.* Sheffield: Equinox.

Gurevich, O., Johnson, M. A., & Goldberg, A. E. (2010). Incidental verbatim memory of language. *Language and Cognition, 2*, 45-78.

Gutierrez, X. (2013). The construct validity of grammaticality judgment tests as measures of explicit and implicit knowledge. *Studies in Second Language Acquisition, 35*, 423-449.

Guy, G. R. (1997). Violable is variable: Optimality theory and linguistic variation. *Language Variation and Change, 9*, 333-347.

Guzmán Naranjo, M. (2016). The se-ra alternation in Spanish subjunctive. *Corpus Linguistics and Linguistic Theory, to appear*.

Haber, L. R. (1975). The muzzy theory. In R. Grossman, J. San, & T. Vance (Eds.), *Proceedings of the Eleventh Regional Meeting of the Chicago Linguisitc Society* (pp. 240-256). Chicago: University of Chicago Press.

Haber, L. R. (1976). Leaped and leapt: A theoretical account of linguistic variation. *Foundations of Language, 14*, 211-238.

Halliday, M. A. (1991a). Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics* (pp. 30-43). New York/London: Longman.

Halliday, M. A. (1991b). Towards probabilistic interpretations. In E. Ventola (Ed.), *Functional Systemic Linguistics: Approaches and Uses* (pp. 39-61). Berlin/New York: Mouton de Gruyter.

Halliday, M. A. (1992). Language as system and language as instance. In J. Svartvig (Ed.), *The corpus as a theoretical construct: Directions in corpus linguistics* (pp. 61-77). Berlin/New York: Mouton de Gruyter.

Ham, S. (1996). O instrumentalu jednince imenica vrste i. *Jezik, 44*, 1-8.

Ham, S. (1996). Zamjeničko-pridjevna sklonidba jednine u hrvatskom književnom jeziku krajem 19.stoljeća. *Filologija, 27*, 25-41.

Ham, S. (2002). Putem ili putom, putovima ili putevima. *Jezik, 49*, 134-142, 183-190.

Ham, S. (2004). Peludne dvojbe. *Jezik, 51*, 19-31.

Hasher, L., & Zacks, R. T. (1984). Automatic Processing of Fundamental Information. The case of frequency of occurrence. *American Psychologist, 39*(12), 1372-1388.

Haspelmath, M. (2008). Creating economical morphosyntactic patterns in language change. In J. Good (Ed.), *Language Universals and Language Change* (pp. 185-214). Oxford: Oxford University Press.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative. *Linguistics, 39*(6), 1041-1070.

Heylen, K. (2005). A Quantitative Corpus Study of German Word Order Variation. In S. Kepser, & M. Reis (Eds.). Berlin/New York: Mouton de Gruyter.

HG. (2005). *Hrvatska gramatika.* (E. Barić, M. Lončarić, D. Malić, S. Pavešić, M. Peti, V. Zečević, & M. Znika, Eds.) Zagreb: Školska knjiga.

HJS. (1999). *Hrvatski jezični savjetnik.* (L. Hudeček, M. Mihaljević, & L. Vukojević, Eds.) Zagreb: Pergamena & Školske novine.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language. London/New York: Routledge.* London/New York: Routledge.

Hoffmann, T. (2006). Corpora and introspection as corroborating evidence: The case of preposition placement in English relative clauses. *Corpus Linguistics and Linguistic Theory, 2*(2), 165-195.

Hofstadter, D. R. (2001). Epilogue: Analogy as the Core of Cognition. In D. Genter, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science* (pp. 499-538). Cambridge, MA: The MIT Press/Bradford Book.

Hopper, P. J., & Traugott, E. C. (1993). *Grammaticalization.* Cambridge: Cambridge University Press.

Householder, F. W. (1966). Phonological theory: a brief comment. *Journal of Linguistics, 2*, 99-100.

HP. (2013). *Hrvatski pravopis.* (Ž. Jozić, Ed.) Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Ignjatović, D. (1960). O rodu imenice bol. *Naš jezik, 10*, 55-59.

Itkonen, E. (2005). *Analogy as structure and process : approaches in linguistics, cognitive psychology and philosophy of science.* Amsterdam: John Benjamins.

Ivić, M. (1965). Problem norme u književnom jeziku. *Jezik, 13*, 1-8.

Jakobson, R. (1932 [1974]). O dnešní brusičství českém [reprinted as O današnjem češkom jezičnom čistunstvu].

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1217-1218.

Janda, L. A. (2014). Introduction to Slavic historical morphology: Slavic noun classes. In K. Gutschmidt, S. Kempgen, T. Berger, & P. Kosta (Eds.), *The Slavic Languages. An International Handbook of their Structure, their History and their Investigation. Volume 2.* (pp. 1565-1582). Berlin/Munich/Boston: Mouton de Gruyter.

Jonke, L. (1965a). *Književni jezik u teoriji i praksi.* Zagreb: Znanje.

Jonke, L. (1965b). Problem norme u hrvatskosrpskom književnom jeziku. *Jezik, 13*, 8-15.

Jusczyk, P., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159-207.

Kačić, M. (2001). *Jezikoslovna promišljanja.* Zagreb: Pergamena.

Kager, R. (1999). *Optimality Theory.* Cambridge: Cambridge University Press.

Kapatsinski, V. (2014). What is grammar like? A usage-based constructionist perspective. In B. Cartoni, D. Bernhard, & D. Tribout (Eds.), *Theoretical and computational approaches to morphology.*

Kapović, M. (2011). Language, Ideology and Politics in Croatia. *Slavia Centralis, IV*(2), 45-56.

Kapović, M. (2012). *Čiji je jezik?* Zagreb: Algoritam.

Katičić, R. (1974). Nešto napomena o postanku složenoga suvremenog jezičnog standarda hrvatskoga ili srpskoga. *Zbornik zagrebačke slavističke škole, 2*, 225-257.

Katičić, R. (2001). Croatian lingustic loyalty. *International Journal of Sociology of Language, 147*, 17-29.

Kempen, G., & Harbusch, K. (2005). The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In S. Kepser, & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp. 329-349). Berlin/New York: Mouton de Gruyter.

Kempen, G., & Harbusch, K. (2008). Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses. In A. Steube (Ed.), *The discourse potential of underspecified structures* (pp. 179-192). Berlin: Walter de Gruyter.

Kepser, S., & Reis, M. (Eds.). (2005). *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives.* Berlin/New York: Mouton de Gruyter.

Keuleers, E., & Daelemans, W. (2007). Memory-Based Learning Models of Inflectional Morphology: A Methodological Case Study. *Lingue e Linguaggio, 6*, 151-174.

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G., & Martens, E. (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive psychology, 54*(4), 283-318.

Kidd, E., Lieven, E., & Tomasello, M. (2010). Lexical frequency and exemplar-based learning effects in language acquisition: evidence from sentential complements. *Language Sciences, 32*, 132-142.

Kilgariff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory, 1*(2), 263-276.

Kiparsky, P. (2005). Blocking and periphrasis in inflectional paradigms. In G. Booij, & J. van Marle (Eds.), *Yearbook of morphology 2004* (pp. 113-135). Dordrecht: Springer.

Klavan, J. (2012). *Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy (PhD dissertation)*. University of Tartu.

Korecky-Kröll, K., Libben, G. S., Wiesinger, J., Reinisch, E., Bertl, J., & Dressler, W. U. (2012). Helping a crocodile to learn German plurals: children's online judgment of actual, potential and illegal plural forms. *Morphology, 22*, 35-65.

Kottum, S. E. (1981). The genitive singular form of masculine nouns in Polish. *Scando-Slavica, 27*(1), 179-186.

Králík, J., & Šulc, M. (2005). The Representativeness of Czech Corpora. *International Journal of Corpus Linguistics, 10*(3), 357-366.

Kroch, A. (1994). Morphosyntactic variation. In K. Beals, & al. (Eds.). Chicago: University of Chicago Press.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: modelling the choice of linking morphemes in Dutch. *Linguistics, 39*(1), 51-93.

Krug, M. (1998). String frequency. *Journal of English Linguistics, 26*, 286-320.

Labov, W. (1969). Contraction, deletion and inherent variability of the English copula. *Language, 45*, 715-762.

Labov, W. (1978). *Sociolinguistic Patterns.* Oxford: Blackwell.

Labov, W. (1994). *Principles of linguistic change, vol. 1: Internal factors.* Oxford: Blackwell.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar (Volume I: Theoretical Prerequisites).* Stanford, CA: Stanford University Press.

Langacker, R. W. (1991). *Concept, Image, and Symbol.* Berlin/New York: Mouton de Gruyter.

Langacker, R. W. (2010). How not to disagree: The emergence of structure from usage. In K. Boye, & E. Engberg-Pedersen (Eds.), *Language Usage and Language Structure* (pp. 107-143). Berlin: Walter de Gruyter.

Lehmann, C. (2007). Linguistic competence: Theory and empiry. *Folia Linguistica, 41*(2), 223-278.

Longo, N. (1964). Probability-learning and habit-reversal in the cockroach. *The American Journal of Psychology, 77*, 29-41.

Lonsdale, D. (2002). Data files for Analogical Modeling. In R. Skousen, D. Lonsdale, & D. B. Parkinson (Eds.).

Luka, B. J., & Barsalou, L. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language, 52*, 436-459.

Lukatela, G., Gligorijević, B., Kostić, A., & Turvey, M. T. (1980). Representation of inflected nouns in the internal lexicon. *Memory and Cognition, 8*, 415-423.

Ljubešić, N., & Klubička, F. (2014). {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In F. Bildhauer, & R. Schäfer (Ed.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, (pp. 29-35). Gothenburg.

MacWhinney, B. (2000). Connectionism and Language Learning. In M. Barlow, & S. Kemmer (Eds.), *Usage-based models of language* (pp. 121-149). Cambridge: Cambridge University Press.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the Verb Learning Model. *Cognition, 29*, 121-157.

Maiden, M. (2004). Morphological autonomy and diachrony. *Yearbook of Morphology*, 137-175.

Maretić, T. (1899). *Gramatika i stilistika hrvatskoga ili srpskoga književnog jezika.* Zagreb: Knjižare L. Hartmana.

Marković, I. (2012). *Uvod u jezičnu morfologiju.* Zagreb: Disput.

Matasović, R. (2008). *Kratka poredbenopovijesna gramatika hrvatskoga jezika.* Zagreb: Matica hrvatska.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics. An Introduction.* Edinburgh: Edinburgh University Press.

Meyer, C. F. (2002). *English Corpus Linguistics: An introduction.* Cambridge: Cambridge University Press.

Milin, P., Keuleers, E., & Filipović Đurđević, D. (2011). Allomorphic Responses in Serbian Pseudo-Nouns as a Result of Analogical Learning. *Acta Linguistica Hungarica, 58*, 65-84.

Milin, P., Kuperman, V., Kostić, A., & Baayen, R. H. (2009). Paradigms bit by bit: an information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J. P. Blevins, & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition* (pp. 214-252). Oxford: Oxford University Press.

Mirković, J., Seidenberg, M. S., & Joanisse, M. F. (2011). Rules versus Statistics: Insights from a Highly Inflected Language. *Cognitive Science, 35*, 638-681.

Mlađenović, A. (1977). Neka pitanja varijantnosti norme u savremenom srpskohrvatskom književnom jeziku. In S. Urbanczyk (Ed.), *Warjancja normy we wspolczesnych slowianskich jezykach literackich* (pp. 51-56). Krakow: Polska akademia nauk.

Moguš, M. (1970). Postoji li danas u pridjeva komparativni nastavak -ji? *Jezik, 18*, 150-152.

Moguš, M. (1977). Još o neujednačenosti komparativnih oblika. *Jezik, 25*, 146-148.

Moguš, M. (1993). *Povijest hrvatskoga književnog jezika.* Zagreb: Nakladni zavod Globus.

Moguš, M., Bratanić, M., & Tadić, M. (Eds.). (1999). *Hrvatski čestotni rječnik.* Zagreb: Školska knjiga.

Morth, K., & Dressler, W. U. (2014). German plural doublets with and without meaning differentation. In F. Reiner, F. Gardani, H. C. Luschutzky, & W. U. Dressler (Eds.), *Morphology and Meaning* (pp. 249-258). Amsterdam: John Benjamins.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition, 94*, 1-18.

Myers, J. (2009). The design and analysis of small-scale syntactic judgment experiments. *Lingua, 119*, 425-444.

Nagata, H. (1988). The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research, 17*(1), 1-17.

Nakisa, R. C., & Hahn, U. (1996). Where defaults don't help: the case of the German plural system. In G. Cotrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 177-182). Mahwah, NJ: Lawrence Erlbaum Associates.

Nevins, A. (2012). Haplological dissimilation at distinct stages of exponence. In J. Trommer (Ed.), *The Morphology and Phonology of Exponence* (pp. 84-116). Oxford: Oxford University Press.

Newell, B. R., & Bright, J. E. (2001). The relationship between the structural mere exposure effect and the implicit learning process. *The Quarterly Journal of Experimental Psychology, 54A*(4), 1087-1104.

Newmeyer, F. (2003). Grammar is grammar and usage is usage. *Language, 79*(4), 682-707.

Nida, E. (1948). The Identification of Morphemes. *Language, 24*, 414-441.

O'Grady, W. (2008). The emergentist program. *Lingua, 118*, 447-464.

Okuka, M. (1982). I stroži i strožiji. *Književni jezik, 11*, 93-97.

Penke, M., & Rosenbach, A. (2004). What counts as evidence in linguistics: An introduction. *Studies in Language, 28*(3), 480-526.

Pennington, J. J. (2013). Wackernagel's Looking Glass: The Mirror Principle and BCS Adjectival Long-Form Allomorphy. *Scando-Slavica, 59*(2), 234-249.

Peti-Stantić, A. (2008). *Jezik naš i/ili njihov: Vježbe iz poredbene povijesti južnoslavenskih standardizacijskuh procesa.* Zagreb: Srednja Europa.

Pinker, S., & Prince, A. (1988). On language and connectionism. *Cognition, 28*, 73-193.

Poplack, S. (2001). Variability, frequency, and productivity in the irrealis domain in French. In J. L. Bybee, & P. Hopper (Eds.).

Pullum, G. K. (2007). Ungrammaticality, Rarity and Corpus Use. *Corpus Linguistics and Linguistic Theory, 3*(1), 33-48.

Quine, W. V. (1951). Two dogmas of empiricism. *Philosophical Review, 60*, 20-43.

Quirk, R. (1970). Aspect and varient variation in English verbs. *Language, 46*(2), 300-317.

Rácz, P., Beckner, C., Hay, J. B., & Pierrehumbert, J. B. (2014). Rules, Analogy, and Social Factors codetermine past-tense formation patterns in English. *52nd Annual Meeting of the Association for Computational Linguistics.* Baltimore, Maryland.

Raguž, D. (1997). *Praktična hrvatska gramatika.* Zagreb: Medicinska naklada.

Rasinger, J. M. (2008). *Quantitative Research in Linguistics.* London: Bloomsbury Academic.

Raymond, W. D., & Brown, E. L. (2012). Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In S. T. Gries, & D. Divjak (Eds.).

Riđanović, M. (2005). O žalosnom stanju naše nauke o jeziku. *Sarajevske sveske, 10*, 46-62.

Roeper, T. (2011). What frequency can do and what it can't. In I. Gülzow, & N. Gagarina (Eds.), *Frequency Effects in Language Acquisition* (pp. 23-49). Berlin: Mouton de Gruyter.

Roland, D., & Jurafsky, D. (2002). Verb sense and verb subcategorization probabilities. In P. Merlo, & S. Stevenson (Eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues* (pp. 325-345). Amsterdam: John Benjamins.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. F. Lloyd (Eds.), *Cognition and Categorization* (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum.

Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In G. Rohdenburg, & B. Mondorf (Eds.), *Determinants of Grammatical Variation in English* (pp. 379-411). Berlin: Mouton de Gruyter.

Rosenbach, A. (2013). Combining elicitation data with corpus data. In M. Krug, & J. Schlüter (Eds.), *Research Methods in Language Variation and Change* (pp. 278-294). Cambridge: Cambridge University Press.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In D. Rumelhart, & J. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (pp. 216-71). Cambridge, MA: MIT Press.

Rytting, C. A. (2002). Testing Analogical Modeling: The /k/~Ø alternation in Turkish. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.). John Benjamins.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110-114.

Samardžija, M. (1987). Duga i kratka množina u hrvatskom književnom jeziku. *Jezik, 35*(5), 129-136.

Samardžija, M. (Ed.). (1999). *Norme i normiranje hrvatskoga standardnoga jezika.* Zagreb: Matica hrvatska.

Sampson, G. (2005). Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics, 10*(1), 15-36.

Sampson, G. (2007). Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory, 3*(1), 1-32.

Schmid, H.-J. (2010). Does frequency in text really instantiate entrenchment in the cognitive system? In D. Glynn, & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics* (Vol. 51, pp. 101-136). Berlin/New York: Mouton de Gruyter.

Schmid, H.-J. (2013). Is usage more than usage after all? The case of English not that. *Linguistics, 51*(1), 75-116.

Schönefeld, D. (Ed.). (2011). *Converging Evidence: Methodological and Theoretical issues for Linguistic Research.* Amsterdam: John Benjamins.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Schütze, C. T. (2005). Thinking about what we are asking speakers to do. In S. Kepser, & M. Reis (Eds.), *Linguistic Evidence: Theoretical, Empirical, and Computational Perspectives* (pp. 457-485). Berlin: Mouton de Gruyter.

Schütze, C. T. (2011). Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(2), 206-221.

Schütze, C. T., & Sprouse, J. (2014). Judgment Data. In D. Sharma, & R. J. Podesva (Eds.), *Researche Methods in Linguistics* (pp. 27-50). Cambridge: Cambridge University Press.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379-423.

Silić, J. (1998). Leksik i norma. *Radovi Zavoda za slavensku filologiju, 32*, 169-175.

Silić, J., & Pranjković, I. (2005). *Gramatika hrvatskog jezika (za gimnazije i visoka učilišta).* Zagreb: Školska knjiga.

Skousen, R. (1989). *Analogical Modeling of Language.* Dordrecht: Kluwer Academic Publishers.

Skousen, R. (2002). Issues in Analogical Modeling. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.).

Skousen, R., Lonsdale, D., & Parkinson, D. B. (Eds.). (2002). *Analogical Modeling: An exemplar-based approach to language.* Amsterdam/Philadelphia: John Benjamins.

Snow, C., & Meijer, G. (1977). On the secondary nature of syntactic intuitions. In S. Greenbaum (Ed.), *Acceptability in Language* (pp. 163-177). The Hague: Mouton.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry, 31*, 575-582.

Sokolova, S., Lyashevskaya, O., & Janda, L. A. (2012). The Locative Alternation and the Russian 'empty' prefixes: A case study of the verb gruzit 'load'. In D. Divjak, & S. T. Gries (Eds.), *Frequency Effects in Language Representation* (pp. 51-85). Berlin: Mouton de Gruyter.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua, 115*, 1497-1524.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics, 1*, 123-134.

Sprouse, J. (2008). The differential sensitivity of acceptability judgments to processing effects. *Linguistic Inquiry, 39*(4), 686-694.

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry, 40*, 329-341.

Sprouse, J. (2011). A Test of the Cognitive Assumptions of Magnitude Estimation: Commutativity does not Hold for Acceptability Judgments. *Language, 87*(2), 274-288.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*.

Sprouse, J., & Almeida, D. (submitted). *Power in acceptability judgment experiments.*

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua, 134*, 219-248.

Stefanowitsch, A. (2005). New York, Dayton (Ohio), and the Raw Frequency Fallacy. *Corpus Linguistics and Linguistic Theory, 1*(2), 295-301.

Stefanowitsch, A. (2006). Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory, 2*(1), 61-77.

Stefanowitsch, A. (2008). Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics, 19*(3), 513-531.

Stefanowitsch, A. (2011). Cognitive linguistics meets the corpus. In M. Brdar, S. T. Gries, & M. Žic Fuchs (Eds.), *Cognitive Linguistics: Convergence and Expansion* (pp. 257-290). Amsterdam: John Benjamins.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209-243.

Stojanov, T., & Vučić, Z. (2012). Korpusnojezikoslovna obradba tekstova Sportskih novosti. *Filologija, 59*, 103-130.

Svartvik, J. (1966). *On voice in the English verb.* The Hague: Mouton.

Szmrecsanyi, B. (2005a). Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory, 1*(1), 113-150.

Šimandl, J. (2008). *Dnešní stav skloňování substantiv takzvaných typů kámen a břímě.* PhD dissertation, Charles University in Prague.

Šipka, M. (1957). Vokativ ličnih imena i prezimena. *Jezik, 6*.

Šipka, M. (1987). *Književnojezička politika i jezička kultura.* Sarajevo: Izdavačka djelatnost.

Tadić, M. (1998). Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika. *Filologija, 30-31*, 337-347.

Tafra, B. (1981). Vrste imeničke deklinacije. *Jezik, 29*, 44-47.

Taylor, J. R. (1989). *Linguistic categorization.* Oxford: Oxford University Press.

Taylor, J. R. (2012). *The Mental Corpus: How Language is Represented in the Mind.* Oxford: Oxford University Press.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10*(1), 1-13.

Težak, S. (1980). Genitiv množine vrste e s višesuglasničkim osnovnim završetkom. *Jezik, 28*, 1-15.

Težak, S. (1982). Vokativ jednine imenica e vrste. *Radovi Zavoda za slavensku filologiju, 17*, 31-40.

Težak, S. (1984). Pridjevsko-zamjenički alomorfi. *Jezik, 32*, 33-43.

Težak, S. (1986). Sibilarizacija u suvremenomu hrvatskom književnom jeziku. *Filologija, 14*, 395-402.

Težak, S. (1989). Instrumental jednine imenica i-vrste. *Jezik, 37*, 33-39.

Težak, S. (1991). *Hrvatski naš svagda(š)nji.* Zagreb: Školske novine.

Težak, S., & Babić, S. (2004). *Gramatika hrvatskoga jezika: Priručnik za osnovno jezično obrazovanje.* Zagreb: Školska knjiga.

Theijssen, D., ten Bosch, L., Boves, L., Cranen, B., & van Halteren, H. (2013). Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory, 9*(2), 227-262.

Thornton, A. M. (2011). Overabundance (multiple forms realizing the same cell): a non-canonical phenomenon in Italian verb morphology. In M. Maiden, & al. (Eds.), *Morphological Autonomy: Perspectives from Romance Inflectional Morphology* (pp. 358-381). Oxford: Oxford University Press.

Thornton, A. M. (2012). Reduction and maintenance of overabundance: A case study on Italian verb paradigms. *Word Structure, 5*(2), 183-207.

Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Pronunciation reflects syntactic probabilities: evidence from spontaneous speech. *Language and Cognition, 1*(2), 147-165.

Tomasello, M. (2003). *Constructing a Language: a Usage-based Theory of Language Acquisition.* Cambridge, MA: Harvard University Press.

Trask, R. L. (1999). *Key Concepts in Language and Linguistics.* London: Routledge.

Ullman, M. T. (1999). Acceptability Ratings of Regular and Irregular Past-tense Forms: Evidence for a Dual-system Model of Language from Word Frequency and Phonological Neighbourhood Effects. *Language and Cognitive Processes, 14*(1), 47-67.

Vennemann, T. (1969). *Zur synchronischen und diachronischen Phonologie des Deutsches.* Munich: Max Hueber.

VHG. (2007). *Velika hrvatska gramatika, knjiga prva: Glasovi i oblici hrvatskoga književnoga jezika.* (S. Babić, D. Brozović, I. Škarić, & S. Težak, Eds.) Zagreb: Nakladni zavod Globus.

Vince, Z. (1977). Neujednačenost u komparativnim oblicima nekih pridjeva. *Jezik, 25*, 142-145.

Vukušić, S. (1990). Sklonidba imenica vuk, vrag, rog, bog u kratkoj množini. *Jezik, 38*, 70-71.

Wasow, T., & Arnold, J. (2004). Intuitions in linguistic argumentation. *Lingua, 115*, 1481-1496.

Weisser, P. (2006). A distributed morphology analysis of Croatian noun inflection. In G. Müller, & J. Trommer (Eds.), *Subanalysis of argument encoding in distributed morphology* (pp. 131-142). Leipzig: Universität Leipzig.

Weskott, T., & Fanselow, G. (2009). Scaling issues in the measurement of linguistic acceptability. In S. Featherston, & S. Winkler (Eds.), *The Fruits of Empirical Linguistics. Volume I: Process* (pp. 229-246). Berlin/New York: Mouton de Gruyter.

Weskott, T., & Fanselow, G. (2011). On the Informativity of Different Measures of Linguistic Acceptability. *Language, 87*(2), 249-273.

Wulf, D. J. (2002). Applying Analogical Modeling to the German Plural. In R. Skousen, D. Lonsdale, & D. B. Parkinson (Eds.). Amsterdam/Philadelphia: John Benjamins.

Yang, C. D. (2004). Universal grammar, statistics, or both. *Trends in Cognitive Sciences, 8*(10), 451-456.

Appendix A. Results of questionnaire study 2

Quest. A	Mean rating	Freq.	Quest. B	Mean rating	Freq.
dugovi	6.93	6,000	bokovi	6.60	900
duzi	1.36	0	boci	1.23	0
ritmovi	6.82	260	troškovi	6.90	12,800
ritmi	1.77	0	trošci	1.28	1
dijelovi	6.92	45,000	autoputo/evi	5.74	29
dijeli	1.58	0	autoputi	3.16	0
rogovi	6.87	800	ključevi	6.89	1,650
rozi	1.92	70	ključi	2.10	120
tornjevi	5.97	670	manjkovi	4.78	300
tornji	2.55	20	manjci	3.74	15
tečajevi	6.42	2,100	golubovi	6.14	770
tečaji	3.08	25	golubi	3.67	50
zvukovi	6.28	1,400	znakovi	6.77	11,000
zvuci	4.36	850	znaci	2.69	1,450
pucnjevi	6.07	200	pijetlovi	6.10	180
pucnji	3.26	160	pijetli	3.13	80
pojase/ovi	5.93	790	prstenovi	5.20	445
pojasi	4.52	700	prsteni	4.88	165
trakovi	2.83	110	skutovi	1.86	35
traci	5.56	280	skuti	6.54	170
palčevi	5.66	75	praškovi	3.37	65
palci	4.00	110	prašci	5.58	195
gavranovi	3.91	42	jablanovi	3.41	95
gavrani	5.84	125	jablani	6.14	145
gr(i)jehovi	1.53	60	srhovi	2.04	10
grijesi	6.92	1,500	srsi	5.54	150
šiljkovi	1.66	3	kolčevi	2.63	2
šiljci	6.91	90	kolci	6.67	130
grebenovi	2.08	5	korjenovi	1.43	40
grebeni	6.81	250	korijeni	6.94	1,860
konjevi	1.45	2	keksovi	1.15	0
konji	6.90	1,600	keksi	6.96	280
stalkovi	1.53	0	noktovi	1.07	0
stalci	6.87	145	nokti	6.91	450
bademovi	2.08	0	jelenovi	2.10	0
bademi	6.54	110	jeleni	6.63	100

Quest. A	Mean rating	Freq.	Quest. B	Mean rating	Freq.
udarcem	6.74	1,300+	sucem	6.76	1,000+
udarcom	2.45	1	sucom	2.18	1
pištoljem	6.73	1,820	nožem	6.85	2,400+
pištoljom	2.21	0	nožom	1.91	0
Štimcem	6.40	150	križem	6.59	450
Štimcom	3.02	12	križom	2.76	25
redateljem	6.43	565	ravnateljem	6.61	990
redateljom	2.81	26	ravnateljom	2.82	20
sprejem	5.23	110	volejem	5.28	117
sprejom	4.38	33	volejom	4.43	38
Kovačem	5.55	150	plaštem	4.72	230
Kovačom	4.04	27	plaštom	5.41	105
stažem	4.38	315	esejem	5.36	20
stažom	5.62	225	esejom	4.21	21
Sergejem	6.16	63	diCaprijem	5.10	15
Sergejom	3.83	52	diCapriom	4.63	18
Marijem	3.82	52	Bushem	3.77	190
Mariom	5.53	160	Bushom	5.85	375
trofejem	5.08	26	muzejem	4.11	95
trofejom	4.68	52	muzejom	5.31	285
Bečem	1.89	4	mjesecem	3.06	11
Bečom	6.62	95	mjesecom	6.23	180
Francuzem	4.17	9	crtežem	2.27	9
Francuzom	5.73	82	crtežom	6.56	260
nosem	2.15	5	pojasem	3.28	4
nosom	6.62	630	pojasom	6.51	590
švercem	2.33	0	guštem	1.93	0
švercom	6.59	130	gušťom	6.81	64

Appendix B. Results of questionnaire study 3 (group comparison)

Green cells indicate matching answers (i.e. the form chosen by the majority of respondents/models matches the dominant ending for a particular lexeme), red cells indicate non-matching answers, yellow cells indicate a 50: 50 decision.

Case	Lexeme	Gloss	Dom. ending	Corpus	Ques.	AM type	AM double	AM token	TIMBL type	TiMBL double	TIMBL token
Isg	slučaj	case	em	100	98	100	91	100	100	97	100
	cilj	goal, aim	em	100	96	100	98	100	100	91	100
	znanac	acquaintance	em	100	99	96	92	100	100	88	100
	kraj	end	em	100	98	100	90	100	99	87	100
	muž	husband	em	98	95	95	62	100	94	51	100
	temelj	foundation	em	99	84	98	70	100	9	37	95
	tuš	shower	em	97	88	82	61	100	83	71	65
	križ	cross	em	95	80	92	58	100	85	55	100
	Lošinj	name of an island	em	82	64	100	56	100	97	73	99
	volej	volley	em	67	54	97	72	100	66	58	0
	roštilj	barbecue	em	87	78	100	79	100	97	73	100
	vic	joke	em	69	81	68	65	100	79	67	95
	linč	lynch	em	52	73	100	64	67	100	67	100
	staž	work experience	em	52	52	54	60	73	65	40	0
	trofej	trophy	em	60	50	77	68	98	37	30	2
	imidž	public image	em	54	62	100	68	76	60	60	50
	Senj	town name	om	100	84	67	77	100	94	80	100
	šverc	smuggling	om	100	92	99	98	100	100	99	100
	Polančec	surname	om	100	95	45	59	100	38	40	19
	keš	cash	om	100	98	99	85	100	100	69	100
	mjesec	moon, month	om	98	87	74	67	100	94	71	100
	Šprajc	surname	om	93	64	98	69	100	80	50	100

	crtež	drawing	om	98	91	100	80	100	100	84	100
	Poreč	town name	om	96	87	62	43	100	95	61	100
	hokej	hockey	om	81	64	73	60	100	94	67	100
	krpelj	tick (animal)	om	85	68	93	59	100	99	59	100
	Bush	surname	om	76	77	52	48	100	19	36	5
	šlic	zipper	om	78	41	96	71	100	90	55	100
	Rovinj	town name	om	58	39	47	60	93	0	2	1
	kolaž	collage	om	54	30	67	72	79	85	31	7
	esej	essay	om	59	40	55	63	98	63	60	6
	štambilj	stamp, seal	om	51	21	54	55	63	13	55	0
Gpl	utvrda	fortress	a	100	91	98	100	81	100	100	78
	crta	line	a	100	95	100	100	84	100	100	62
	obrva	eyebrow	a	100	99	99	100	88	100	100	88
	utrka	race	a	100	87	71	100	72	100	100	42
	pošta	post office	a	94	45	48	100	64	36	100	52
	Ande	mountain range	a	91	63	46	100	53	0	0	15
	zvijezda	star	a	98	55	72	100	68	19	100	39
	cesta	road	a	98	88	84	100	62	96	99	44
	nevjesta	bride	a	78	36	46	100	60	80	100	30
	kovrča	curly piece of hair	a	67	71	78	100	66	100	100	34
	Alpe	mountain range	a	82	58	57	100	48	8	0	7
	svekrva	mother-in-law	a	75	42	97	100	74	100	100	90
	gesta	gesture	a	54	22	59	79	61	87	67	30
	cista	cyst	a	54	43	87	80	66	96	100	40
	kasta	caste	a	53	26	53	72	52	77	100	30
	bista	bust (statue)	a	54	26	65	84	59	88	100	32
	crpka	pump	i	100	97	100	100	99	100	100	100
	akna	acne	i	100	97	100	100	95	100	100	86
	stotinka	hundredth of a second	i	100	99	100	100	100	100	100	98

	milijarda	billion	i	100	97	100	100	85	100	100	90
	maska	mask	i	96	96	99	100	66	99	100	84
	boljka	complaint	i	99	96	92	100	73	98	90	89
	hrenovka	hot dog	i	93	90	100	100	73	100	100	87
	njuška	muzzle	i	90	78	82	100	71	68	100	97
	dagnja	mussel	i	84	94	100	100	68	94	100	92
	praćka	slingshot	i	67	89	100	100	59	98	100	89
	ljuska	shell	i	64	63	90	100	59	95	100	82
	vesta	sweater	i	66	85	84	100	76	22	56	60
	fešta	festivity	i	60	89	74	99	69	74	31	56
	brazda	furrow	i	60	78	100	97	67	99	100	61
	pasta	paste	i	60	74	75	100	73	26	0	63
	naranča	orange	i	51	73	100	81	55	100	100	77
	litra	litre	a-a	100	97	54	100	53	51	82	5
	zemlja	land, earth	a-a	100	100	54	100	51	14	0	23
	sestra	sister	a-a	100	96	47	100	54	30	7	5
	pjesma	song	a-a	100	97	54	100	57	38	97	30
	daska	wooden board	a-a	92	76	56	100	55	5	0	12
	smokva	fig	a-a	93	47	82	100	55	78	100	22
	guska	goose	a-a	92	65	58	100	58	39	32	19