Article:
Nikolaev, A. and Bermel, N. (2023) Studying negative evidence in Finnish language corpora. Word Structure, 16 (2-3). pp. 206-232. ISSN: 1750-1245

https://doi.org/10.3366/word.2023.0229

Studying negative evidence in Finnish language corpora[1]

Alexandre Nikolaev (University of Eastern Finland)
Neil Bermel (University of Sheffield)

Abstract

This study explores the relationship between lower-than-expected frequencies of word forms and inherent gaps in Finnish inflectional paradigms. The research aims to determine whether it is possible to predict paradigmatic gaps from lower-than-expected frequencies of word forms. We examined Finnish nouns inflected in a marginal case (instructive) and hypothesized that some of these nouns may potentially have gaps in their inflectional paradigms. However, we found that such gaps are contingent and do not cause uncertainty when filled. We find that the correlation between inherent gaps and lower frequencies is one-directional: predicting inherent gaps from lower-than-expected frequencies is problematic. The results suggest that any paradigmatic gap suggested by corpus frequency is more likely to be contingent than inherent, and that the less semantic need there is for a particular word form, the more likely it will be unattested even in a large corpus. The research highlights the importance of considering semantic profiles when analyzing the grammaticality of word forms and suggests that statistical tests like Fisher's exact are not necessarily the right approach to tackle the problem of negative evidence in corpus studies.

[1] This research was supported by grant AH/T002859/1 from UK Arts and Humanities Research Council.

1 Introduction

In morphologically complex languages, such as Finnish, the inflectional paradigm of a noun can include up to two thousand inflected word forms. However, this upper limit can only be achieved in practice when a noun displays all possible combinations of its stem, number and case suffixes, all optional possessive suffixes, and clitics. Therefore, the difference between these idealized (canonical) inflectional noun paradigms and attested noun paradigms (even in large corpora) is almost always enormous. Janda and Tyers (2018) express a strong form of this observation in deducing from the facts at hand that the inflectional paradigms of all lexemes are defective. The authors make their case primarily on Russian data, so it stands to reason that Finnish, with its even greater inflectional complexity, would be an even better subject. This state of affairs means that native speakers may not have heard, read, or produced all possible inflected word forms, even for the most common lexemes. However, this broad view of paradigmatic defectivity differs from usual definitions in that potential paradigmatic gaps can be purely accidental, and in most instances, we expect speakers to have no difficulty producing or interpreting the missing word forms.

According to the opposing narrow (but also in its way radical) view of inflectional defectivity, only those word forms that grammars and dictionaries mark as missing or not in use are considered defective. Often, this is based on unspecified criteria, most likely on the authors' introspection or sometimes on searches in corpora. This type of paradigmatic gap typically manifests as uncertainty in native speakers. For example, Finnish speakers have no difficulty producing the phrase *parane pian* "get better soon" or *potilas parani* "a patient got better". However, there seems to be no infinitive form for this verb, such as *?parata* "to get better", which is acknowledged by the Grammar of Finnish (Hakulinen et al., 2004: § 348). This does not mean that Finnish speakers do not know how to get better; most of them just do not know how to express it using the infinitive form of the verb, due to the gap in this inflectional paradigm.

Nikolaev and Bermel (2022) take a hybrid view on defectivity. The authors do not restrict defectivity only to those word forms that are mentioned in handbooks as defective, nor do they claim that all paradigms are defective. Instead, the authors propose the so-called "dimmer" model of inflectional defectivity. They describe defectivity as a continuum of paradigmatic gaps ranging from the most contingent (with no uncertainty) to the most inherent (with a lot of uncertainty), meaning that there must be cases of less contingent and less inherent gaps. The authors administered language production and evaluation tasks to native speakers of Finnish. They

conceptualized uncertainty in the word production task as different avoidance strategies applied by native speakers. In this approach, defectivity in inflectional paradigms is linked to the production or evaluation results of native speakers. However, the authors found that the level of uncertainty is strongly (and negatively) linked to the corpus frequencies of the lexemes, especially lemma frequency: the lower the frequency, the higher the uncertainty.

1.1 Present study

The findings in our previous study (Nikolaev & Bermel, 2022) suggest that smaller-than-expected corpus frequencies of Finnish word forms (or the absence of a form in a corpus) might be considered as indicators of their defectivity. The present study discusses the usefulness of corpora as a tool for detecting paradigmatic gaps in a language.

Language corpora are often used as an alternative for studying paradigmatic defectivity because no sample of participants can compete with the amount and variety of data in large corpora. We selected the 2,000 most frequent Finnish nouns from a list of the 10,000 most frequent Finnish words (CSC – Tieteen tietotekniikan keskus, 2004). We then retrieved frequencies of all inflected word forms for these nouns from a corpus of Finnish (84,308,641 tokens) based on written conversations of thousands of users in a Reddit-like internet community (Aller Media ltd., 2014). These 2,000 nouns have 104,783 different word forms, and their summed frequency is 10,427,959, which accounts for 12.4% of the corpus tokens.

We chose one particular case, the instructive case, because it is one of the least frequently used cases in Finnish. There is a chance that because of its overall low frequency, this case will be more easily realized as a gap in inflectional paradigms. However, unlike the similarly rare abessive case, the instructive still displays usage patterns similar to those of other, more frequently-used cases. Nikolaev, Chuang, and Baayen (2023) used an unsupervised data clustering algorithm (tSNE) to cluster Finnish cases based on *fasttext* embeddings (300-dimensional semantic vectors that approximate words' meanings; Grave et al., 2018). The authors found that all Finnish cases show clear clustering except one: the abessive. The authors propose that the semantics of the abessive are less systematic and hence more idiomatic, but the instructive's more case-like clusterings make it a suitable candidate for casehood.

1.2 Finnish nouns and their distribution

Nouns in Finnish can be inflected in the singular and plural number and in 15 cases (or 14, if we exclude the accusative, which typically is homonymous with the nominative or genitive forms: a non-autonomous case in the sense of Zaliznjak, 2002: 629). These inflected word forms can then include one of the 5 possessive suffixes (1sg., 2sg., 1pl., 2pl., and 3sg/pl.) followed by a pragmatic clitic or a combination of clitics (-kO, -kin, -kAAn, -hAn, -pA, and the two rare clitics -kA and -s; the capital letters O and A indicate vowel harmony: o/ö, a/ä). An example of one inflected word form is the following:

(1) auto-i-ssa-ni-kin
car-PL-INESS-1SG.POSS-ALSO
'also in my cars'

All the type counts presented in this section (e.g., in Table 1) refer to the number of distinct word forms, not the frequency of occurrence of these word forms. E.g., the word *auto* 'car' had 193 different attested word forms in the corpus (e.g., *auto*, *auton, autoa*, *autoja*, *autolla*, *autot*, etc.). In this section thus we ignore for the moment how many times each of these forms was met in the corpus (e.g., *auto* 11,973, *auton* 11,303, *autoa* 4,917, *autoja* 3,536, *autolla* 3,497, *autot* 3,017, etc. with the total lemma frequency for *auto* being 55,940). To put it simply, 2,000 different nouns had 104,783 distinct word forms, which had 10,427,959 occurrences in the corpus. In this section, we will discuss only those 104,783 distinct word forms; their lemma frequencies (10,427,959) will be discussed in the next section (see section 1.3).

On average, each of our nouns had 52 distinct word forms. Word forms that did not have possessive suffixes and clitics constituted 36.6% of the data (38,334 / 104,783). Word forms that did not have clitics constituted 76.6% of the data (80,235 / 104,783).

The 104,783 word forms were distributed among 699 different combinations of 2,000 possible outcomes manifested by the schema *stem* + *number* + *case* + *possessive suffix* + *clitic* (Karlsson & Koskenniemi, 1985). Table 1 shows the top 30 and the bottom 30 of these combinations sorted by the frequency of each combination[2].

[2] The frequencies of combinations at the bottom are all '1', as expected. However, this does not mean that only combinations 670-699 have a frequency of 1; rather, they represent a sample from the tail.

The top 30	Word form	Freq.	The bottom 30	Word form	Freq.
1	num pl case gen	2492	670	num sg case tra clit foc pa	1
2	num sg case nom	2000	671	num sg case tra clit qst+foc s	1
3	num sg case par	1998	672	num sg case tra poss px3 clit qst	1
4	num sg case gen	1996	673	num sg case tra poss pxpl1 clit foc kaan	1
5	num pl case par	1963	674	num sg case tra poss pxpl1 clit qst+foc han	1
6	num sg case ela	1943	675	num sg case tra poss pxpl1 clit qst+foc kaan	1
7	num pl case nom	1932	676	num sg case tra poss pxpl2 clit foc kaan	1
8	num sg case ill	1914	677	num sg case tra poss pxpl2 clit qst	1
9	num sg case ade	1868	678	num sg case tra poss pxpl2 clit qst+foc kaan	1
10	num sg case ine	1772	679	num sg case tra poss pxsg1 clit qst+foc kaan	1
11	num sg case all	1768	680	num sg case tra poss pxsg2 clit foc kaan	1
12	num pl case ela	1721	681	num sg case tra poss pxsg2 clit qst+foc kaan	1
13	num pl case ill	1655	682	num sg pl case com poss px3 clit foc kaan+qst	1
14	num sg case tra	1596	683	num sg pl case com poss px3 clit qst+foc kaan	1
15	num sg case ess	1592	684	num sg pl case com poss pxpl1 clit foc kaan	1
16	num pl case ade	1510	685	num sg pl case com poss pxpl1 clit foc kaan+qst	1
17	num pl case ine	1510	686	num sg pl case com poss pxpl1 clit qst	1
18	num sg case nom clit foc kin	1492	687	num sg pl case com poss pxpl1 clit qst+foc kaan	1
19	num sg case abl	1455	688	num sg pl case com poss pxpl2 clit foc kaan	1
20	num pl case nom poss px3	1390	689	num sg pl case com poss pxpl2 clit foc kaan+qst	1
21	num pl case all	1358	690	num sg pl case com poss pxpl2 clit qst	1
22	num sg case gen poss px3	1291	691	num sg pl case com poss pxpl2 clit qst+foc kaan	1
23	num sg case par poss px3	1270	692	num sg pl case com poss pxsg1 clit foc kaan	1
24	num sg case nom poss pxsg2	1245	693	num sg pl case com poss pxsg1 clit foc kaan+qst	1
25	num pl case par poss px3	1223	694	num sg pl case com poss pxsg1 clit qst	1
26	num sg pl case com poss px3	1222	695	num sg pl case com poss pxsg1 clit qst+foc kaan	1
27	num sg case nom clit foc han	1155	696	num sg pl case com poss pxsg2 clit foc kaan	1
28	num pl case nom clit foc kin	1123	697	num sg pl case com poss pxsg2 clit foc kaan+qst	1
29	num sg case nom poss pxsg1	1101	698	num sg pl case com poss pxsg2 clit qst	1
30	num sg case ela poss px3	1001	699	num sg pl case com poss pxsg2 clit qst+foc kaan	1

Table 1. The top 30 and bottom 30 combinations of number, case, possessive suffix, and clitic sorted by their frequency.

As we can see from Table 1, each of the 2,000 nouns had their nominative singular form (the so-called "dictionary form") represented in this corpus. One might wonder how it is possible that 2,000 nouns had more than 2,000 genitive plural word forms. In order for this to be possible, all or almost all nouns have to be represented by the genitive plural form and some of the nouns have to be represented by more than one genitive plural form. This is evidence of overabundance: e.g., the word *asia* 'matter, question, issue, thing, subject' has three genitive plural forms attested in the corpus: *asioiden*, *asioitten*, and *asiain*. Other cases too can demonstrate overabundance, e.g., partitive plural forms for the word *keskustelu* 'conversation' are *keskusteluja* and *keskusteluita*, or, e.g., the word *omena* 'apple' has two partitive plural forms, *omenoita* and *omenia*.

Karlsson and Koskenniemi (1985: 210) calculated that each Finnish nominal paradigm can have approximately 150 core slots (by which they mean word forms that do not have clitics). Indeed, after excluding all word forms with clitics and recalculating Table 1, we found 151 different inflectional slots. Therefore, all core paradigmatic slots are present in our 84.3-million-word corpus. Karlsson and Koskenniemi (1985: 210) claim that there should be 1,850 additional extra-

paradigmatic cliticized forms; however, in our corpus we only found 548 extra-paradigmatic slots (699 inflectional slots minus 151 core paradigmatic slots). *Summa summarum*, the core forms constituted 76.6% of our data and they were distributed among 151 different forms. In addition, 23.4% of the word forms in our data had clitics, and these extra-paradigmatic forms were distributed among 548 different extra-paradigmatic slots. These numbers suggest that in nouns, the core paradigmatic forms prevail in the lexicon and their paradigm size is large, but not as astonishingly large as many articles discussing the inflectional morphology of Finnish nouns claim. Larger sets of word forms per lemma (2,000 [150+1,850]) are achieved mostly in theory and mostly because of those extra-paradigmatic members that have clitics.

Table 1 is presented as a plot in Figure 1.

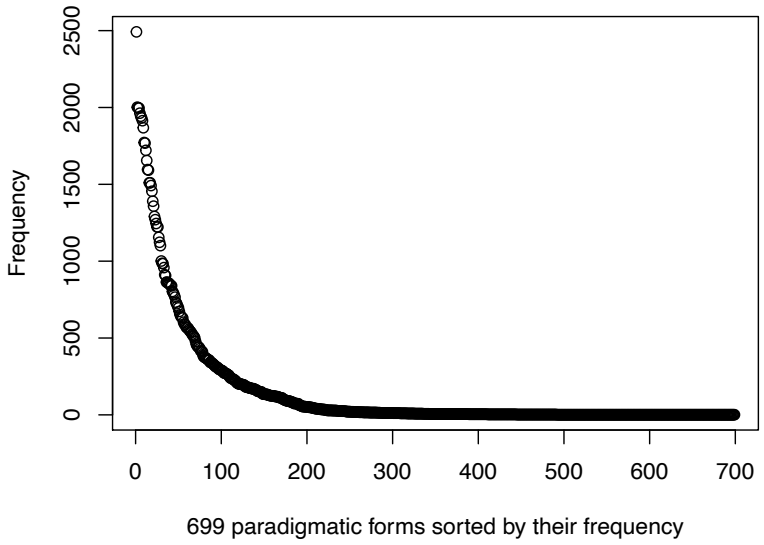

Figure 1. 699 attested combinations of 2,000 possible outcomes by which the schema *stem* + *number* + *case* + *possessive suffix* + *clitic* can manifest itself in Finnish nouns. Combinations (inflectional slots) are sorted according to their frequency in the corpus.

1.3. The instructive case in Finnish nouns and its distribution

As signalled in the previous section, in this section we discuss corpus frequencies – in other words, not only how many different inflectional forms the word *auto* 'car' has (193), but how many times we meet each of these 193 forms of the word *auto* in the corpus (e.g., *auto* 11,973, *auton* 11,303, *autoa* 4,917, *autoja* 3,536, *autolla* 3,497, *autot* 3,017, etc.). We will then discuss the frequency and distribution of forms for one specific case: the instructive.

To begin with, we present the lemma frequencies of our 2,000 highest-frequency nouns in Figure 2.

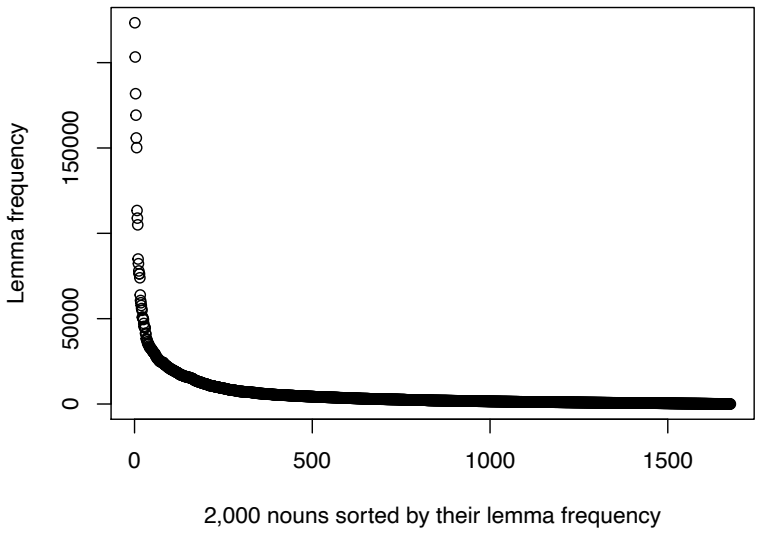

Figure 2. Lemma frequencies retrieved for 2,000 nouns from the corpus of 84.3 million words.

Both figures 1 and 2 show a Zipfian distribution that is one of the power law probability distributions; however, Figure 2 is a more typical Zipfian distribution, with frequency dropping off more sharply at the high-frequency end of the graph. This is expected, since each data point in Figure 2 is a lexeme (e.g., *auto* 'car') which obviously carries its own autosemantic meaning, and thus lexemes form a range from the least used to the most used. However, the data points in Figure 1 are inflectional forms, in other words functional concepts, e.g., *stem* + *nominative singular*, which are placeholders for word forms such as *auto* or any other noun in the nominative singular. These placeholders also form a range from the least used to the most used; however, the second most used (in this case it happens to be the functional slot for the nominative singular) is not that different in frequency from the third most used (partitive singular).

Case	Frequency	Percentage
Nominative	3204389	30.7
Partitive	2078001	19.9
Genitive	1804387	17.3
Inessive	756853	7.3
Illative	696812	6.7
Elative	593981	5.7
Adessive	527436	5.1
Allative	262657	2.5
Essive	214560	2.1
Ablative	117882	1.1
Translative	103888	1.0
Instructive	46133	0.4
Comitative	10804	0.1
Abessive	10176	0.1

Table 2. Finnish cases and their frequency distribution in the corpus.

As an example of paradigmatic defectivity, we use Finnish nouns inflected in the instructive case. The instructive is a marginal case in Finnish. It accounts for only 0.3–2% of all nominal word forms (Hakulinen et al., 2004). In our data (see Table 2) it accounts for 0.4% of noun forms (46,133 out of 10,427,959). The instructive case expresses manner, means, instrument, location or time, and thus its syntactic function is primarily adverbial (Karlsson, 2018; Hakulinen et al., 2004). Like another marginal case, the comitative, the instructive is defective in that it only can be used in the plural (with a few rare exceptions for singular forms, e.g., *jalan* 'by foot', in which instances the instructive is homonymous with the genitive singular *jalan* 'of the foot / foot's' and the accusative singular *jalan* 'foot').

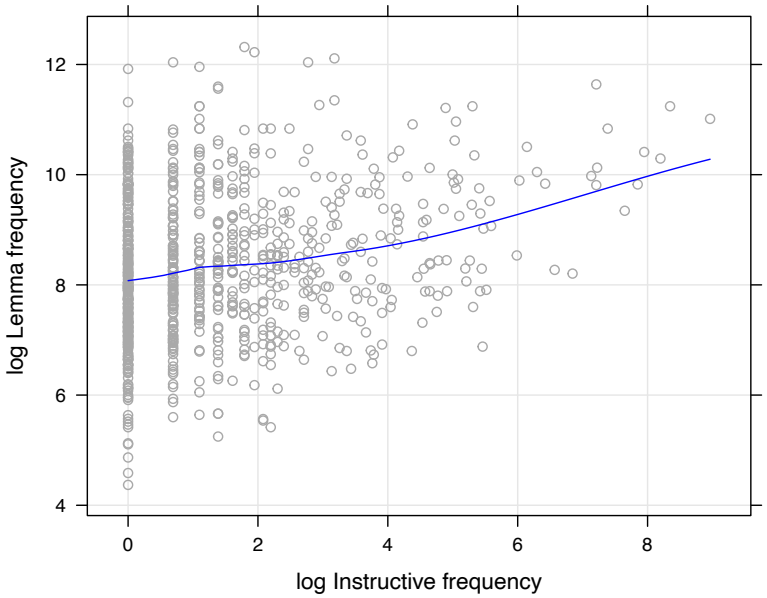

Figure 3. The relationship between the logarithmically transformed lemma frequency of the 827 (out of 2000) most frequent nouns that have at least one occurrence in the instructive case. The blue line is a smoother fitted to the data.

However, the distribution of these 46,133 instructive case nouns seems not to be random. On the one hand, the higher the instructive case frequency of a lexeme, the higher the probability that this lexeme will have a higher lemma frequency (B = 0.304, t = 7.167, p < 0.001; see Figure 3). However, the opposite statement (the higher the lemma frequency, the higher the instructive case frequency) is not necessarily correct: far more high-frequency lemmas have few or no instructive case forms in the corpus than have high numbers of instructive case forms. The relationship between the number of instructive case forms and lemma frequencies does not seem to be equally distributed at all points along the regression line: variability in lemma frequency is high when instructive case forms are few, and it is much narrower when instructive case forms are many in number. This is a clear case of heteroscedasticity, when the variability of the instructive occurrences is unequal across the range of values of the lemma frequency. If we sort nouns in our data according to their lemma frequency, we can see this pattern in Table 3.

Noun	Gloss	Instructive freq	Lemma freq
asia	matter, question, issue, thing, subject	5	223365
ihminen	human	2	203309
aika	time	18	181805
mies	man	12	169245
vuosi	year	11	155868
nainen	woman	0	150204
mieli	mind	1353	113378
maa	land, earth, soil, ground	142	108904
lapsi	child	0	104970
työ	work	22	84948
elämä	life	0	82208
raha	money	19	77963
osa	part	4216	76258
päivä	day	201	76245
kerta	time, occasion, turn	51	73943
paikka	place	603	63880
tapa	way, manner, means, mode, habit, custom	7808	60658
maailma	world	0	59326
syy	reason	156	57725
auto	car	0	55940

Table 3. List of the first 20 most frequent nouns, their lemma frequencies and frequencies in the instructive case.

The bottom right corner in Figure 3 is almost empty of data points, showing that the higher the frequency of the instructive word form, the more likely it is that its lemma frequency will be also high. Hence, the relation between the lemma and instructive case frequencies seems to be rather one-directional. How can this non-mutual relation be explained? Words that are frequently present in language do not necessarily have semantic features that would fit well with the functions of the instructive case (manner, means, instrument, location or time). Hence, the mere fact that a noun has a high lemma frequency does not entail the noun's applicability to the constructions requiring the instructive case. Some instructive constructions are indisputably productive (e.g., those that contain the lexemes *mieli* 'mind' or *osa* 'part') whereas some other instructive constructions show low productivity (e.g., those containing the lexeme *mies* 'man') and still other constructions are in between those poles (e.g., the lexeme *syy* 'reason'). We thus consider productivity of instructive constructions to form a spectrum from the least productive to the most productive. Constructions such as those containing the noun *lapsi* 'child' or *nainen* 'women' are the least productive constructions: both, *lapsi* and *nainen* lack instructive case forms in our data. If we agree with one of the basic tenets of cognitive linguistics – that grammatical constructions are learned pairings of a form and a function (e.g., see Goldberg, 2019) – then there should be certain semantic components that make certain nouns more likely to have frequently-used instructive case forms.

However, for these nouns, their varying degree of association with an instrument-type meaning suggests that if the pairings of a form and a function are learned, there need not be anything inherent about them; the pairings are simply results of one of several possible variants that became most frequent, which might or might not fit with some inherent features that make instructive constructions less or more productive (e.g., compare the floor of a building: English *on the first floor*, Russian *на первом этаже* 'on the first floor', vs. Finnish *ensimmäisessä kerroksessa* 'in the first floor', and Czech *v prvním patře* 'in the first floor'; with at least two ways of saying it that are semantically/functionally plausible, one becomes conventionalized and can eventually exclude all others).

Not all these instructive constructions have the same form: *stem* + *instructive case*. Some of them require a premodifier, which does not always obey the rules of agreement (Karlsson, 2018) as in the phrase *tällä tavoin* 'in this way' (adessive singular *tällä* 'this' + instructive *tavoin* 'way'). Karlsson (2018) states that the instructive plural forms of some words have broken away from their original nominals and become independent adverbs: *käsin* (*käsi* 'hand' + instructive plural) 'by hand', *paikoin* (*paikka* 'place' + instructive plural) 'in (some) places'. Regardless of the lemma frequency, some nouns in the instructive case are used in a context that suggests a rather different interpretation from the one reserved for a noun. These word forms have a different syntactic (and hence semantic) function and can be labeled as adverbs, preposition, postpositions, or particles. This can explain why some words in Table 3, such as the noun *tapa* 'way' are much more frequent in the instructive case than some other nouns, e.g., *päivä* 'day', which otherwise is more frequent in the corpus than *tapa*.

There seem thus to be semantic and syntactic reasons behind the defectivity of words such as *lapsi*, *nainen*, *elämä* vs. *mieli*, *osa*, *tapa* (see Table 3). However, how "deeply" defective are the paradigms of the words *lapsi*, *nainen*, *elämä* regarding the absence of instructive forms? Is their absence *contingent* (i.e., on corpus size) and might be reversed if our sample was larger, or is it *inherent* (i.e., structurally problematic in some way) and thus part of the description of the lexeme (on these terms, see Chuang et al., 2022)? One way to determine the depth of defectivity is to increase the corpus size and investigate whether these words' paradigms remain defective. In what follows, we analyze separately each of the five words from Table 3 that have zero occurrences of the instructive case and hence are good candidates for studying negative evidence in Finnish corpora.

2 Can corpus frequencies indicate defectivity?

2.1 The word elämä 'life' as a case study and application of Fisher's exact test

Stefanowitsch (2020) claims that zero frequency does not mean a lack of information and zero can be as informative as any non-zero number denoting frequency. He discusses the problem of negative evidence in a given corpus: namely, claims that corpora do not contain negative evidence. As his response to a critique that some constructions in a corpus do not provide evidence of their ungrammaticality because they "simply haven't occurred yet", he suggests quantifying these constructions as zero and treating them as if they are as informative as any other numeric variable. As a method, he suggests using Fisher's exact test[3] to explore the reliability and usefulness of negative evidence in corpora.

Following Stefanowitsch (2020) we calculated how unlikely it is that the word *elämä* 'life' has zero occurrences in the instructive case in the corpus given the word's lemma frequency and given the frequency of other words inflected in the instructive case in the corpus. Fisher's exact test produced a *p*-value of 2.198244e-187 meaning that it has 186 zeros after the decimal point: 0.000[...]0002198. There is thus basically a zero probability that we obtained this contingency table (see Table 4) simply by chance. Hence, the instructive form of the word *elämä* is expected to have many more than zero occurrences in the given corpus. The fact that it does not makes the instructive case of this word (*elämin*) a good candidate for a paradigmatic gap.

	Instructive	Other cases
elämä	0	82208
other nouns	46132	8740698

Table 4. The non-occurrence of the noun *elämä* 'life' in the instructive case in a corpus of 84.3 million tokens

[3] Fisher's exact test is a statistical test used to determine the probability of a particular distribution of frequencies within a contingency table, given the marginal totals of that table. It is commonly used in linguistics to analyze the association between two categorical variables. For instance, a researcher is interested in analyzing the relationship between say, a speaker's gender and their use of a particular grammatical structure. The researcher might collect data on a sample of speakers and record the number of male and female speakers who use the structure and the number who do not. The resulting data can be arranged in a contingency table, where the rows represent the gender of the speakers and the columns represent their use of the grammatical structure. Fisher's exact test can then be used to determine the probability of observing the distribution of frequencies in the contingency table, assuming that there is no real association between the two variables (i.e., the observed frequencies are simply due to chance). If the probability is sufficiently low, the researcher can conclude that the observed association between the two variables is unlikely to have occurred by chance alone, and therefore, that there is likely a real association between the two variables.

We then increased the corpus size by a factor of 31, from 84.3 million words to 2.6 billion words (a corpus based on written conversations of thousands of users in a Reddit-like internet community, Aller Media ltd., 2014). However improbable it was according to Fisher's exact test (with its *p*-value with 186 zeros after the point) to witness zero instructive case forms for the word *elämä* in corpora based on data from a corpus of 84.3 million tokens, it turned out that this form was not impossible: we found two occurrences of the word *elämä* in the instructive case in our larger corpus (see sentences in Appendix A, part I). We repeated this procedure for other words such as *nainen,* 'women', *lapsi* 'child', *maailma* 'world', and *auto* 'car'. Details can be found in the next section and in the Appendix A, parts II-V.

The fact that the predictions of Fisher's exact test did not hold after we increased the corpus size suggests that it may not be an effective tool for linguists to identify defective paradigms based on lower-than-expected corpus frequencies. This possibility highlights the need for alternative methods to be considered in future research. The major issue with this approach is as follows: Fisher's exact test was designed to test whether two different conditions found in a fixed dataset (for example, profession: manual/office and gender: male/female in a certain group of people; or in our study a specific lexeme vs. other lexemes and one case vs. other cases in a corpus) are independent of each other or linked in their reflection in the data. As usual, a small p-value suggests that the chances of independence (randomness) are small, i.e., the linkage we notice is significant. Thus in our current situation, a small p-value of the test should indicate that a word form or construction occurs in a corpus less frequently than expected (which could indicate a paradigmatic gap). Alternatively, a small p-value could indicate that a word form or construction occurs in a corpus far more frequently than expected, as described in Stefanowitsch and Gries (2003). However, Fischer's exact test was developed based on assumptions that the data come from an otherwise homogeneous population. Corpora, on the other hand, are far from homogeneous: they sample different registers and texts from different speakers of various ages that address different topics. Many aspects of language structure have probability distributions that can be characterized as Large Number of Rare Event distributions (see Baayen, 2001). Large numbers of linguistic *events* have extremely small probabilities that are difficult or even impossible to estimate based on small corpora (e.g., 84.3 million tokens). Even in large corpora (e.g., 2.6 billion tokens), probabilities will fluctuate depending on the kinds of registers and topics sampled[4].

[4] We thank Harald Baayen for suggesting this explanation.

2.2 Case studies of the words nainen, 'women', lapsi 'child', maailma 'world', and auto 'car'

We proposed that increasing the corpus size could provide evidence against the hypothesis that a missing word form was defective. However, what if this is not the case, and Fisher's exact test with an extremely small p-value is actually a reliable indicator of defectivity? In other words, is it possible that the word *elämä* discussed in the previous subsection is an outlier or exception to the general pattern?

The world *nainen* 'woman' also has zero occurrences in the instructive case in our original corpus, even though the word's lemma frequency (150,204) is greater than that of the word *elämä* 'life'. Fisher's exact test calculated in R using the function *fisher.test*() assigns zero probability (p-value = 0; no decimal points) that this paradigmatic gap could be observed only by chance. As with the word *elämä*, again, we increased the corpus size from 84.3 million to 2.6 billion words. We found one occurrence of the word *nainen* in the instructive case (*naisin*) (see Appendix A, part II).

The world *lapsi* 'child' has zero occurrences in the instructive case in our original corpus. The word's lemma frequency (104,970) is greater than that of the word *elämä* 'life' but smaller than that of *nainen* (150,204). As with the words *elämä* and *nainen*, Fisher's exact assigns 6.929628e-239 probability that this paradigmatic gap could be observed only by chance. However, after we increased the corpus size from 84.3 million to 2.6 billion words, we found 10 occurrences of the word *lapsi* in the instructive case (*lapsin*, Appendix A, part III).

The world *maailma* 'world' is technically a compound. Its first constituent is the word *maa* 'earth' *ilma* 'air'. However, this compound is not transparent to most naïve language users. The word has zero occurrences in the instructive case in our original corpus. Its lemma frequency (59,326) is smaller than that of the words *elämä*, *nainen*, or *lapsi*. However, Fisher's exact test assigns 1.14487e-135 probability that this paradigmatic gap is observed by chance. After we increased the corpus size from 84.3 million to 2.6 billion words, we still found zero occurrences of the word *maailma* in the instructive case. We then increased the corpus size from 2.6 billion words to 13.2 billion words by including all possible corpora from historic to newspaper and to fiction books. As a result, we found one instructive-case occurrence of the word *maailma*, but only as a constituent in another compound: *mielikuvitusmaailma* 'imaginary world' (in the sentence we found it was written separately which is an error; compounds in Finnish should be written without a space between constituents, see Appendix A, part IV).

The word *auto* 'car' has zero occurrences in the instructive case in our original corpus. Its lemma frequency (55,940) is smaller than that of the other words we analyzed in the previous sections. However, Fisher's exact test assigns 7.705405e-128 probability that this paradigmatic gap could be observed by chance. After we increased the corpus size from 84.3 million to 2.6 billion words, we found 26 occurrences of the word *auto* in the instructive case (*autoin,* see Appendix A, part V)[5].

How interpretable (grammatical) are these sentences? We asked three Ph.D. students in linguistics who are native speakers of Finnish to provide their grammaticality judgments for all the sentences we found in a 2.6-billion-word corpus. Our informants provided their subjective grammaticality ratings using a four-point Likert scale, where 0 represented "totally ungrammatical" and 3 represented "absolutely grammatical." We did not anticipate receiving ratings of 3 or near 3, given the infrequent use of these words in the instructive case. However, we also did not expect all the sentences to be rated as 0 or near 0. We also asked our informants whether the target word was actually in the instructive case for each sentence. While their responses were not always in agreement, they often indicated that the word form was in the instructive case for many sentences. However, for some other sentences, they agreed that the word form was not in the instructive case. Interestingly, when we presented several examples where all informants agreed that the word form was not in the instructive case (e.g., those with the words *nainen* or *maailma*) and asked them independently to specify which case it belonged to, they revised their original choice and indicated that it was, in fact, the instructive case. We assume that the lower grammatical acceptability of these words is also related to how we meta-linguistically perceive them with respect to their case. Nonetheless, in Appendix A, we report our informants' original choices.

2.3 Other candidates besides Fisher's exact test to link negative corpus evidence to defectivity

The case studies presented in sections 2.1 and 2.2 (above) show that relying on a purely frequency-based analysis to spot paradigmatic gaps could lead to many false positive findings that are not confirmed by usage or by speakers' intuition. In a morphologically complex language such as Finnish, one could find, using an even bigger corpus and a test such as Fisher's exact test, many

[5] In some sentences a corpus parser treats noun forms *autoin* as inflected verb forms *autoin* 'I helped'. Therefore, it is likely that the word auto has some instructive case forms even in our original 84.3 million-token corpus; they are simply parsed as verbs.

contingent paradigmatic gaps which, according to Fisher's test, should not be considered contingent.

A Bayesian approach to studying paradigmatic gaps may have more potential to be informative than the frequentist approach when the goal is to distinguish purely accidental (contingent) gaps in inflectional paradigms from those that cause uncertainty in language users (inherent gaps). In the Bayesian approach, we are allowed to use background information while setting our priors, and this is where subjective grammaticality judgments could be used to set priors and then be updated by corpus data. However, developing a Bayesian counterpart to Fisher's exact test that would be tuned to put more weight on prior probabilities is not a definitive solution, for the same reason mentioned in section 2.1: many aspects of language structure have probability distributions that can be characterized as Large Number of Rare Event distributions (Baayen, 2001). When linguistic events have extremely small probabilities, they are difficult, if not impossible, to estimate based on corpora.

Zero-inflated models (see, e.g., Dalrymple, Hudson, & Ford 2003) are a type of statistical model used to analyze data where there are an excessive number of zero values. They are often used in instances where the data generating process has two components: one generates zero values, and the other generates non-zero values. Zero-inflated models have been used in linguistics to model frequency data in which many words appear only once or twice in a large corpus, leading to a high number of zero values. These models can account for the fact that some words in a language are very rare and occur infrequently, while others are much more common. However, there are reasons to believe that no statistical test or model can distinguish between contingent gaps and inherent gaps. We discuss these reasons in section 3.

2.4 Hierarchies

If we follow Janda and Tyers (2018), all inflectional paradigms in Finnish are defective. However, if this is the case, then it would be appropriate to substitute the term *inflectional paradigm* for *defective inflectional paradigm*. If, on the other hand, we follow the traditional approach found in handbooks, all Finnish paradigms are full/canonical, except for a few that are defective for several reasons. We have decided to use a hybrid approach. Throughout this paper, we refer to paradigmatic gaps as either contingent (an adaptation of Janda & Tyers, 2018's ideas) or inherent (an adaptation of the traditional approach). However, this hybrid (binary) approach is

a simplification that is primarily used for better readability. In fact, Nikolaev and Bermel (2022) showed that, for native Finnish speakers, paradigmatic gaps can be located on a spectrum from the most contingent to the least contingent (most inherent). If our article aims to find a statistical test that can differentiate between contingent and inherent gaps based on corpus frequencies, then this would be the first problem we need to address. Any border line that this test would draw would be too conventional and not helpful when used as such. Two word forms placed close to each other but on different sides of this conventional border according to some statistical test/model would not be that different from each other regarding their defectivity. The question then arises: why would we need this conventional border line on the spectrum from the most contingent to the most inherent gap? Indeed, we argue that we do not need this line. Therefore, we do not need a test that can help us draw this line based on corpus frequencies (whether it is Fisher's exact or zero-inflated or any other test/model). The concept of two points being similar to each other but divided by some conventional line (cf. significant and non-significant p-values, e.g., 0.06 vs. 0.04) is the target of critique presented by proponents of the Bayesian approach. However, even in studies using the frequentist approach, such as Stefanowitsch and Gries (2003), the authors argue against fixating on this conventional line and for using a hierarchical approach instead. They use p-values against frequentist basic tenets (which are that, e.g., a p-value of 0.02 is no more significant than a p-value of 0.04) as a weight to create a hierarchical structure of words' collostructions. In other words, Stefanowitsch and Gries (2003) use the frequentist Fisher's exact test almost as if it were a Bayesian test.

If we adopt the approach of Stefanowitsch and Gries (2003) to construct a hierarchy of gaps based on the p-values obtained from Fisher's exact test, we can arrange them in the following order, from more inherent to less inherent:

1) *nainen* 'woman' (the most inherent, p-value = 0)
2) *lapsi* ‘child’ (6.929628e-239)
3) *elämä* ‘life’ (2.198244e-187)
4) *maailma* ‘world’ (1.14487e-135)
5) *auto* ‘car’ (the least inherent, p-value = 7.705405e-128).

Nevertheless, this hierarchy does not correspond to a hierarchy based on the number of instructive cases that we identified for these words after expanding the corpus size from 84.3 million to 2.6 billion words (with the exception of the least inherent gap):

1) *maailma* (0 sentences, or 1 sentence after the corpus size is increased from 2.6 to 13.2 billion; the most inherent gap)

2) *nainen* (1)

3) *elämä* (2)

4) *lapsi* (10)

5) *auto* (26 sentences, the least inherent gap).

The following is a hierarchy of grammaticality that we constructed based on our informants' judgments (for words with more than one sentence in the 2.6-billion-word corpus, we selected the most grammatical score as the representative score):

1) *maailma* (0.67 out of 3, the most inherent gap)

2) *nainen* (1 out of 3)

3) *elämä* (1.33 out of 3)

4) *lapsi* (1.67 out of 3)

5) *auto* (2.67 out of 3, the least inherent gap).

The order of the latter hierarchy is the same as the one based on the number of hits when the corpus size is increased. This suggests that p-values from Fisher's exact test, used as weights, are not good predictors of how grammatically acceptable the inflected words are perceived to be or how likely they are to appear in a larger corpus. However, the last two measures seem to correlate well, which in a way links higher corpus frequencies to better acceptability. The question remains whether better grammatical acceptability aligns with the spectrum of gaps from the most inherent to the most contingent.

3 Discussion

We presented different measures for the hierarchy of paradigmatic gaps, including p-values obtained from Fisher's exact test, the number of hits in a corpus, and judgments of grammaticality by informants. While the order of these measures is not consistent, there are reasons to believe that lower corpus frequencies correlate with lower acceptability. We added a third component to this interaction: paradigmatic defectivity. However, the results do not provide conclusive evidence to support or refute the hypothesis that grammatical acceptability aligns with the spectrum of paradigmatic gaps, from the most inherent to the most contingent gap in word inflection.

According to Nikolaev and Bermel (2022), inherent paradigmatic gaps in inflectional paradigms in Finnish are linked to lower corpus frequencies or the absence of word forms in corpora. The causal relation discovered in Nikolaev and Bermel's study, where inherent gaps lead to lower frequencies and lower acceptability, appears to be more robust than the causal relation found in the present study, where lower frequencies lead to lower acceptability and inherent gaps.

In the current study, we examined Finnish nouns inflected in a marginal case (instructive). We assumed that some of these nouns potentially have gaps in their inflectional paradigms due to, for example, semantic reasons (cf. the word *clothes*, whose singular form *cloth* has developed as a separate semantic unit). Contrary to this assumption, these gaps seem to be contingent and, when filled, do not cause uncertainty. One of the reasons why an attempt to predict paradigmatic gaps from lower-than-expected frequencies seems to fail is that inherent paradigmatic defectivity in Finnish is a marginal phenomenon. Any paradigmatic gap suggested by corpus frequency is orders of magnitude more likely to be contingent (accidental) than inherent. If, on average, 1,948 word forms of one noun (2,000 potential forms minus 52 forms that were attested in a corpus of 84.3 million tokens) have zero frequency, this means we have 1,948 potential gaps (using a corpus of 84.3 million tokens) for each noun. However, any of these word forms might be encountered in other circumstances (e.g., if we enlarge the corpus size).

Another problem if the aim of our article were to find a statistical test that (based on the corpus frequencies) would differentiate between contingent and inherent gaps is that inherent paradigmatic gaps are moving targets. What does make a gap inherent (or more inherent than some other gap)? To be able correctly decode some word forms one has never encountered before, one needs to match them with other stored/familiar word forms based on formational, referential and constructional similarities between them (see, e.g., Blevins 2006 for the Word and Paradigm model of language). However, an absence of some word forms that are candidates for paradigmatic gaps seems to be only weakly linked to their ungrammaticality (if at all). The less semantic need there is for a particular word form, e.g., for an instructive case form of the words *elämä* 'life' or *maailma* 'world', the more likely these forms will be unattested even in a large corpus. Consider Table 3, in which the word form *elämin* 'life-PL-INSTR' was witnessed zero times in a corpus. Therefore, Fisher's exact test states that the chances that we will observe this particular distribution only by accident are extremely small. However, the whole assumption of chance/randomness ignores the fact that there might be less semantic need for the instructive case of the word *elämä* compared to

other nouns. And yet Table 3 does compare this particular zero frequency of the instructive word form *elämin* to frequencies of other nouns in the instructive case. If some of these other nouns are more likely to be met in the instructive case because of their semantic idiosyncrasies, then the assumption of a random distribution of their forms does not hold and hence should not be tested. Stefanowitsch (2006; 2020) suggests that the knowledge that a particular word form is significantly less frequent than expected should remain relevant even when faced with apparent counterexamples. Can the statistical concept *being present significantly less than expected* be straightforwardly linked to the notion of ungrammaticality of a particular word form? We are afraid that it cannot, and one of the reasons is the semantic idiosyncrasy of some word forms as revealed in the current study by the instructive case. Statistical tests like Fisher's exact are useful for detecting anomalies in contingency tables. However, some of these anomalies may not be due to grammaticality but to varying semantic needs for certain word forms. Therefore, some of the variability in our informants' judgments may also result from their assessment of the likelihood of the context, in addition to the likelihood that a form can fill that context. For example, in Russian, the dative plural is the least frequent case. If we use the overall frequencies to evaluate the case, many animate nouns that refer to humans will outscore the mean, while many inanimate nouns will fall below or close to the mean. This could lead us to wrongly propose that most inanimate nouns are defective, even though native speakers can easily predict the dat. pl. of, e.g., the word диск 'disc' (дискам), despite it being found in only 0.3% of the 6052 examples of диск in the 300m-word Russian National Corpus.

After a corpus reaches a certain size, it exceeds the total amount of language data that any individual will encounter. As it gets larger, a corpus linguist starts seeing more positive evidence for some word forms or constructions which for a large number of native speakers are unacceptable, in part because large corpora are richer in genres, styles, vocabulary, etc. than what the average person encounters in the course of a lifetime. This could call into question the idea that a corpus can represent acceptability. Finding boundaries for what would be considered grammatical/acceptable in language seems to require an algorithm or set of algorithms. Various models of language have pursued this kind of mathematical elegance (e.g., Chomsky, 1957). However, these attempts seem more successful on a theoretical level than a practical one (it is easier to posit a mathematically elegant model than to prove its relationship to reality). Instead, in recent years the capacity of language models has increased from 100 million parameters to 175

billion parameters (GPT-3, an autoregressive language model[6] with 96 layers trained for a total of 300 billion tokens: Brown et al., 2020), and this is a strong trend in language modeling.

We assume that one of the reasons why pursuing mathematical elegance in language and developing rule-based artificial intelligence has fallen short so far is that language is embedded in living organisms, complex biological systems that went through millions of years of evolution (see, e.g., Dennett, 1996). Hence, language was shaped by these pressures/restrictions. It had to adapt to articulatory and respiratory organs and to brain circuits that originally had different functions than language (e.g., Lieberman, 2015). Therefore, language, as any biological system which was developed through the process of adaptation, is messier than anyone seeking mathematical elegance would wish it to be.

4 Conclusion

This study explored the relationship between the frequency of word forms and paradigmatic gaps in Finnish inflectional paradigms. The results suggest that lower-than-expected frequencies of word forms do not necessarily indicate inherent paradigmatic gaps in Finnish. Instead, the absence of some word forms may be due to semantic idiosyncrasies, which weakly link to their ungrammaticality. It is noteworthy that paradigmatic gaps in Finnish are epiphenomena of its morphological complexity. As the size of paradigms expands, the usefulness of corpora to predict inherent versus accidental paradigmatic gaps declines.

Although there is no definitive way to identify inherently defective items using only corpus data, it is possible to detect potential defectivity that can then be empirically verified elsewhere, albeit with some degree of imprecision. This method aligns with an original goal of corpus linguistics, which is to identify patterns of contrast not visible to the naked eye or to intuition. Further research is needed to determine the significance of such patterns.

Acknowledgement

This research was supported by grant AH/T002859/1 from the UK Arts and Humanities Research Council.

[6] Newcomers, such as spiking neuronal network models (Traub et al., 2020), are not that different from deep learning models, since they are mostly concerned with problems of carbon footprint and thus their goal is to reduce energy consumption (which would be biologically more plausible since our brain does not need the energy of several power plants in order to function).

References

Aller Media ltd. (2014). The Suomi 24 Sentences Corpus (2016H2) [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017021505

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Basic Dictionary of Finnish '*Suomen kielen perussanakirja*' (1990–1994). Heksinki: Edita Oyj.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics, 42*, 531–573.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chuang, Y. Y., Brown, D., Baayen, H., & Evans, R. P. (2023). Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigms. *The Mental Lexicon*.

CSC – Tieteen tietotekniikan keskus (2004). Frequency Lexicon of the Finnish Newspaper Language. Retrieved from http://urn.fi/urn:nbn:fi:lb-201405272

Dalrymple, M. L., Hudson, I. L., & Ford, R. P. K. (2003). Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, 41(3-4), 491-504.

Dennett, D. C. (1996). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Penguin Books: London.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions.* Princeton University Press.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R. & Alho, I. (2004). *Iso suomen kielioppi.* SKS:n toimituksia 950. Helsinki: Suomalaisen Kirjallisuuden Seura.

Janda, A. L., & Tyers, M. F. (2018). Less is more: why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory, (published online ahead of print 2018), doi: https://doi.org/10.1515/cllt-2018-0031*

Karlsson, F. (2018). *Finnish: A comprehensive grammar*. Routledge: London and New York.

Karlsson, F., & Koskenniemi, K. (1985). A process model of morphology and lexicon. *Folia Linguistica, 19(1-2)*, 207–232.

Lieberman, P. (2015). The evolution of language. In *Handbook of Intelligence* (pp. 47-64). Springer, New York, NY.

Nikolaev, A,, & Bermel, N. (2022). Explaining uncertainty and defectivity of inflectional paradigms. Cognitive Linguistics, 33(3), 585-621.

Nikolaev, A., Chuang, Y. Y., & Baayen, R. H. (2023). A generating model for Finnish nominal inflection using distributional semantics. *The Mental Lexicon*. Doi: 10.1075/ml.22008.nik

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, 8(2), 209-243.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.

Traub, M., Butz, M. V., Baayen, R. H., & Otte, S. (2020). Learning Precise Spike Timings with Eligibility Traces. *arXiv* preprint arXiv:2006.09988.

Zaliznjak, A. A. (2002). O ponimanii termina ’padezh’ vi lingvisticheskix opisanijax. In Russkoe slovoizmenenie. Moscow: Jazyki slavjanskoj kul’tury. 613–647.

Appendix A

I) The word *elämä* 'life'

(1) *Peliä kun ei voi reaalimaailmassa aloittaa alusta uusin **elämin**.*

"*You can't start a game from the beginning using new lives in a real world.*"

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(2) *Todellakin, yksiöhän ON arvoasunto. Omin intiimein **elämin** ja minimoiduin hajuhaitoin.*

"*Indeed, a one-bedroom flat is an apartment. With its own private lives and minimal smell downsides*."

a) Case: instructive 1 out of 3

b) Number: plural 0 out of 3

c) Grammaticality: 1 out of 3

II) The word *nainen* 'woman'

(1) *On jo ihmeellistä, ettei ole tuolla päättävissä elimissä muuta tekemistä. Monen miehin ja **naisin** jauhetaan tälläista aivan Suomen "peruskallioon" kuuluvaa kevätvirttä, lauletaanko eikö lauleta.*

"*I wonder if this deciding organ has anything else to do. Through its many men and women they harp on the same old song that belongs to "the bedrock" of Finland, should we be singing or shouldn't we.*"

a) Case: instructive 0 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1 out of 3

III) The word *lapsi* 'child'

(1) *Toiset naiset eivät suinkaan vihaa näitä petettyjä vaimoja vaan säälivät ja ovat lopulta helpottuneita, että itse pääsevät eroon miehestä eivätkä ole sidottuja näihin avioliiton sitein, yhteisin taloin, omaisuuksin ja **lapsin**.*

"Other women certainly don't hate these betrayed wives but rather pity and are ultimately relieved to be able to get rid of their husbands and not be bound by marriage ties, joint houses, property, and children."

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(2) *Kun ei itestä viitinyt huolta pitää, olihan mies jo kahlittu **lapsin** ja omaisuuksin*.

"When one didn't bother to take care of themselves, the man was already tied up with children and property."

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 0.67 out of 3

(3) *Kerjääviä äitejä pakkosynnytetyin vammasin **lapsin**? niin raamatullista?*

"Begging mothers with forcibly born disabled children? so biblical?"

a) Case: instructive 1 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 0.67 out of 3

(4) *Haluat torit täyteen ei toivottuja lapsia ja äitejä vammaisin **lapsin** kerjäämään?*

"Do you want squares full of unwanted children and mothers with disabled children begging?"

a) Case: instructive 0 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(5) *Se, että anoppi elää omaa elämäänsä, ehkä uutta avioliittoa, uusin **lapsin**, työelämässä vielä 15 v. ei ole mikään este niille vaatimuksille, joita suositellaan.*

"The fact that the mother-in-law lives her own life, maybe a new marriage, new children, and is still working for 15 years, is no obstacle to the demands recommended."

a) Case: instructive 1 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(6) *Ämmillä on taipumusta aloittaa lihominen kunhan ensin ovat saaneet miehen sidottua **lapsin** itseensä.*

"Women tend to start gaining weight once they've tied a man to themselves with children."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(7) *Toista ei kuitenkaan voi sitoa itseensä lainoin, **lapsin** tai avioliitolla, kyllä se täytyy tulla tunteesta ja halusta olla yhdessä ja jakaa kaikki.*

"However, one cannot tie oneself to another through loans, children, or marriage; it must come from the feeling and desire to be together and share everything."

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1 out of 3

(8) *Tajuan myös nuorison, aika ankeata on aloittaa työelämä, jonka tulevaisuudessa ei voi nähdä omaa asuntoa, perhettä, hyvin koulutetuin* ***lapsin*** *ja autoa.*

"I also understand the youth; it's pretty bleak to start a career when you cannot see your own home, family, well-educated children, and a car in the future."

a) Case: instructive 0 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 0.33 out of 3

(9) *Ahneus voi potkaista omaan nilkkaan koska tiedän että periaatteesta ihmisiä jotka ei maksa moisesta ja onhan se toki esim 4h perheeltä, puoliaikuisin* ***lapsin****, 20e tyhjästä.*

"Greed can backfire because I know that, as a matter of principle, there are people who won't pay for that, and of course, it's, for example, 4 person family with part-time children, 20 eur for nothing."

a) Case: instructive 1 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(10) *Ensin sitoudutaan toiseen valoin, veloin,* ***lapsin*** *jne. Ja sitten huomataan ettei homma toimikaan.*

"First, one commits to the other with loans, children, etc. And then you realize that it doesn't work."

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1 out of 3

IV) The word *maailma* 'world'

(1) *Minusta tälläinen ketju on aivan paikallaan ja tänne kirjoitellaan omia tapahtumia ja sattumia,huumorilla ,mielikuvitus* ***maailmoin****,ja joskus ihan asiaakin kerrotaan.*

"*I think this thread is valid and one can write here stories and coincidences with humor, with imaginary worlds [world-PL-INSTR], and sometimes seriously.*"

a) Case: instructive 0 out of 3

b) Number: plural 0 out of 3

c) Grammaticality: 0.67 out of 3

V) The word *auto* 'car'

(1) *Kyseessä on keskustan liike-elämän ns. kestävä kehittäminen: ihmiset halutaan omin* ***autoin*** *ostosparatiiseihin, myllyihin ja muihin kauppakeskuksiin, viihtymään ja rahojaan tuhlaamaan.*

"This is about sustainable development in the center's business world: people want to go to shopping paradises, mills, and other shopping centers with their own cars, to enjoy themselves and spend their money."

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(2) *Eivät uskalla ajella teipatuin **autoin**, asuvat postilokerossa, että silleen.*

"They don't dare to drive around in wrapped cars and live in a post office box, like that."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 2.33 out of 3

(3) *Kokonaan toinen juttu on lähtiä vyöräytymään paikalle omin **autoin** ja kalustein.*

"It's a completely different thing to start rolling to the site with your own cars and equipment."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(4) *Varsinkin kaupungin itäisistä naapurikunnista Turkuun omin **autoin** tuleville on asetettava ruuhkien pienentämiseksi reilusti bussilipun hintainen tullimaksu!*

"Especially for those coming to Turku from the neighboring municipalities in the east of the city with their own cars, a toll fee equivalent to the cost of a bus ticket must be charged to reduce traffic congestion!"

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 2 out of 3

(5) *Entäpä he, jotka lähtevät **autoin**, lentokonein ja laivoin ties minne?*

"What about those who go by car, plane, and ship anywhere?"

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(6) *tervetuloa runsain **autoin** mukaan!*

"Welcome along with plenty of cars!"

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(7) *Mut eiköhän tämä ole monen miehen "kohtalo" tavata persaukinen ikuinen opiskelija/siivooja/kotiäiti jolla on kovat halut perustaa perhe taloin mökein uusin **autoin** ja ulkomaanmatkoin sisutaa jatkuvasti ja haluta aina lahjoja yms.!*

“But isn't it the "destiny" of many men to meet a perpetually impoverished student/cleaner/housewife with strong desires to start a family with houses, cottages, new cars, and foreign trips and always want gifts, etc.!”

a) Case: instructive 1 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1 out of 3

(8) *Päätös tarkoittaa sitä, että uusien asuinalueiden liikenne tapahtuu vain ja ainoastaan omin* ***autoin****.*

“The decision means that the traffic on new residential areas will only take place by private cars.”

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(9) *Mitenkäs ne noin viinanpersoja ovat kuskeiksi ottaneet. Ovatko renkeinä vai ihan omin* ***autoin*** *liikkuvat.*

“How have they taken such drunken people as drivers? Are they as hired workers, or do they drive their cars?”

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(10) *Mun puolestani systeemin sais romuttaa kyllä sellaiseksi että vajaasti toimintakunnossa olevat ihmiset ei todellakaan liikennöis omin* ***autoin*** *metriäkään.*

“As for me, the system could be destroyed so that people who are slightly dysfunctional wouldn't be able to drive a meter by their cars.”

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(11) *Tuo taasen johtaa siihen että enenevissä määrin ihmiset siirtyvät omin* ***autoin*** *liikenteeseen, koska eivät viitsi maksaa "kallista" lipunhintaa (tosin bussilla matkustaminen on edelleen edullista autoon verrattuna ja suh'koht sujuvaa).*

“This, in turn, leads to an increasing number of people using private cars because they do not want to pay the "expensive" fare (although traveling by bus is still cheap and relatively smooth compared to a car).”

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 2 out of 3

(12) *kokemuksia on myös virolaisista -vuokralainen piti maksullisia naisia asunossa, sitten lapsiperheen joka hajosi ..sitten yksinhuoltaja muksujen kanssa johon sitten mies monin hienoin* ***autoin*** *-itä rekkareinkin asuttui...*

"There are also experiences with Estonians - a tenant kept paid women in the apartment, then a family with children that broke up... then a single parent with children whom a man came with many fancy cars..."

a) Case: instructive 1 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1 out of 3

(13) *Ei Saksaa pelkin* ***autoin*** *ja lentokonein olisi voitettu, tarvittiin raakaa ja raskasta tulivoimaa ja sitä neukulla oli ihan omasta takaa.*

"Germany couldn't have been defeated with just cars and airplanes, raw and heavy firepower was needed and the Soviet Union had it in abundance."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 2.67 out of 3

(14) *Ei ollut rahaa ja politikointia, vaan kaikki oli puhdasta kilpailua paremmuudesta tasaväkisin* ***autoin****.*

"There was no money or politicking, everything was pure competition for superiority with evenly matched cars."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(15) *Invapaikkojen ollessa täynnä sinne kuulumattomin* ***autoin*** *voi invalidi itse tehdä aloitteen city-marketin henkilökunnalle parkkipaikkojen valvonnan tehostamisesta.*

"When disabled parking spots are full of cars that don't belong there, the disabled person can initiate the reinforcement of parking monitoring to the staff of the city market."

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 0.67 out of 3

(16) *Näiltä nurmijärviltä ei kuitenkaan aina kannata järjestää joukkoliikenneyhteyksiä, koska kaikki ajavat omin* ***autoin****.*

"However, it's not always worth organizing public transportation from these Nurmijärvi areas because everyone drives their own cars."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(17) *Mitä omin* ***autoin*** *aina ajavata omakotiasukkaatb siihen vastaavat?*

"What do detached house residents answer to the question of why they always drive by their own cars?"

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(18) *Parhaaseen lintujen pesimäaikaan mennään mellastamaan tuhansin **autoin** ja ämyrein sinne ja samalla kuitenkin kehdataan pitää paitaa, jossa lukee isolla TURKIS ON MURHAA.*

"During the best nesting time for birds, we go there with thousands of cars and speakers blaring, and yet dare to wear a shirt that says FUR IS MURDER in big letters."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3

(19) *Mainonnan ja ohjelman raja on häilyvä, ja sitä viranomaiset seuraavat: Hyvin tamanomaista on esim. salkkareissa ilmoitus "ohjelma sisältää tuotesijoittelua" ja se tarkoittaa sitä, että ohjelmassa ajetaan tietyn merkkisin **autoin**, taustalla olevaan kaupan hyllyyn on ladattu tietyn firman tuotteita, jne.*

"The line between advertising and programming is blurred, and authorities monitor it: for example, in Salatut Elämät, there is an announcement "the program contains product placement" which means that a certain brand of car is driven in the program, a certain company's products are loaded onto the store shelf in the background, etc."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 2.33 out of 3

(20) *Mihin kiskoja enää tarvitaan, jos kaikki ajaa tuhatta ja sataa moottoritieluokkaisilla teillä keskustaan omin **autoin**?*

"What use are rails anymore if everyone drives their own cars at full speed on highway-class roads to the city center?"

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(21) *Siihen on mukava vastata haalarein ja nelivetoisin **autoin**.*

"It's nice to answer that in overalls and with four-wheel drive cars."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 2 out of 3

(22) *Itse kun noudatan käskyä ja lähes pysähdyn joka töyssyyn kun perkeleistä ei pääse ehjin **autoin** yli niin johan on takapuskurissa ylipainoisia perheenpäitä ruotsintraktoreillaan huudattamassa tööttiä.*

"When I follow the order and almost come to a stop at every bump because the damn things can't be crossed over intact, there are overweight family members honking their horns at me with their Swedish tractors on the rear bumper."

a) Case: instructive 2 out of 3

b) Number: plural 0 out of 3

c) Grammaticality: 2.67 out of 3

(23) *Berner kertoi, että nyt saa jopa luvan ruveta liftaamaan* ***autoin*** *ilmaiseksi.*

"Berner said that now you can even get permission to hitchhike for free with cars."

a) Case: instructive 2 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 0.67 out of 3

(24) *Pitkäkilpisin* ***autoin*** *liikkuvat rosvot ovat taas tulleet venelaitureiden tuntumaan suunnittelemaan kesän varkauksiaan.*

"Long-plated robbers moving by car have come back to the vicinity of the boat piers to plan their summer thefts."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(25) *Kun vielä otetaan huomioon, että henkilöautojen määrä kasvaa hurjaa vauhtia, kohta omin* ***autoin*** *torilla kävijät olettekin ihmeessä?*

"Considering that the number of passenger cars is growing at a rapid pace, soon those who visit the square by their own cars will be in a bit of a predicament."

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.33 out of 3

(26) *Missä käytte töissä, työmatkat/muut matkat omin* ***autoin*** *vai julkisilla?*

"Where do you work, do you commute/ travel by your own car or by public transportation?"

a) Case: instructive 3 out of 3

b) Number: plural 3 out of 3

c) Grammaticality: 1.67 out of 3