A generating model for Finnish nominal inflection using distributional semantics

Alexandre Nikolaev[1], Yu-Ying Chuang[2], and R. Harald Baayen[2]
[1] University of Eastern Finland | [2] University of Tübingen

Finnish nouns are characterized by rich inflectional variation, with obligatory marking of case and number, with optional possessive suffixes and with the possibility of further cliticization. We present a model for the conceptualization of Finnish inflected nouns, using pre-compiled fasttext embeddings (300-dimensional semantic vectors that approximate words' meanings). Instead of deriving the semantic vector of an inflected word from another word in its paradigm, we propose that an inflected word is conceptualized by means of summation of latent vectors representing the meanings of its lexeme and its inflectional features. We tested this model on the 2,000 most frequent Finnish nouns and their inflected word forms from a corpus of Finnish (84 million tokens). Visualization of the semantic space of Finnish using t-SNE clarified that a 'main effects' additive model does not do justice to the semantics of inflection. In Finnish, how number is realized turns out to vary substantially with case. Further interactions emerged with the possessive suffixes and the clitics. By taking these interactions into account, the accuracy of our model, evaluated with the fasttext embeddings as gold standard, improved from 76% to 89%. Analyses of the errors made by the model clarified that 7.5% of errors are due to overabundance (and hence not true errors), and that 16.5% of the errors involved exchanges of semantically highly similar stems (lexemes). Our results indicate, first, that the semantics of Finnish noun inflection are more intricate than assumed thus far, and second, that these intricacies can be captured with surprisingly high accuracy by a simple generating model based on imputed semantic vectors for lexemes, inflectional features, and interactions of inflectional features.

Keywords: word embeddings, inflectional morphology, fasttext, word2vec, tSNE, imputed semantic vectors, Finnish language

Interactive figures available from https://doi.org/10.1075/ml.22008.nik.figures
https://doi.org/10.1075/ml.22008.nik | Published online: 17 March 2023
The Mental Lexicon ISSN 1871-1340 | E-ISSN 1871-1375

1. Introduction

This study presents a model for the conceptualization of Finnish inflected nouns. Our aim is to develop a "generating" model, i.e., a statistical model that produces quantitative predictions for the meanings of inflected nouns as represented in distributional space.[1] One way to conceptualize noun use in a language is to start with, e.g., a singular nominative form, and then to derive other noun forms from this basic form. This classical approach is often used in the description of Indo-European languages. However, in Finnish, it is not always the nominative singular from which other words are derived. As Table 1 shows, the paradigm for the word *lasi* 'glass' could be viewed as derived from the nominative singular. However, one would need to know at least several other forms (the so-called principal parts) of the word *vesi* 'water' in order to produce the other forms in the paradigm of this noun. In fact, for children acquiring Finnish, it is not the nominative singular but rather the partitive singular form of the word *vesi* that tends to be acquired first (Karlsson, 1983).

In this general approach to inflectional paradigms, the goal is to clarify how to derive the different forms of a paradigm from a small number of characteristic forms in that paradigm. However, it is not self-evident that speakers conceptualize, e.g., an allative plural by starting first conceptualizing a partitive singular and then re-conceptualizing the result as an allative plural. This study therefore explores an alternative approach to the conceptualization of inflectional semantics, using the general framework of the discriminative lexicon model (Baayen et al., 2019) in combination with a decompositional approach to the semantic vectors (word embeddings) that figure prominently in distributional semantics (Boleda, 2020; Firth, 1968; Günther et al., 2019; Harris, 1954; Landauer and Dumais, 1997; Mikolov et al., 2013; Wang et al., 2019). In distributional semantics, one way of deriving the meaning of an inflected (or derived) word form is to take the semantic vector of the base, and apply a general operation Φ that applies to any base vector and that will produce the corresponding inflected (or derived) semantics. This approach has been developed for derivation by Marelli and Baroni (2015) and is studied for English nominal pluralization by Shafaei-Bajestan et al. (2022) and Shafaei-Bajestan et al. (this volume). According to Marelli and Baroni (2015), Φ is a linear transformation. Conversely, according to Shafaei-Bajestan et al. (2022), for English noun plurals, Φ implements the addition of a (properly conditioned) plural vector to a singular vector (see also Shafaei-Bajestan et al., this volume).

1. The basic concepts underlying distributional semantics are explained in the introduction to this special issue. We also express our gratitude to two anonymous reviewers for their constructive comments and suggestions.

Table 1. Case by number paradigms for two Finnish nouns

Lexical entry: *lasi* 'glass'

	Singular	**Plural**
nominative	lasi	lasi-t
genitive	lasi-n	lasi-en
partitive	lasi-a	lase-ja
essive	lasi-na	lasei-na
translative	lasi-ksi	lasei-ksi
inessive	lasi-ssa	lasei-ssa
elative	lasi-sta	lasei-sta
illative	lasi-in	lasei-hin
adessive	lasi-lla	lasei-lla
ablative	lasi-lta	lasei-lta
allative	lasi-lle	lasei-lle
abessive	lasi-tta	lasei-tta
comitative		lasei-ne-(poss)
instructive		lasei-n

Lexical entry: *vesi* 'water'

	Singular	**Plural**
nominative	vesi	vede-t
genitive	vede-n	vesi-en
partitive	vet-tä	vesi-ä
essive	vete-nä	vesi-nä
translative	vede-ksi	vesi-ksi
inessive	vede-ssä	vesi-ssä
elative	vede-stä	vesi-stä
illative	vete-en	vesi-in
adessive	vede-llä	vesi-llä
ablative	vede-ltä	vesi-ltä
allative	vede-lle	vesi-lle
abessive	vede-ttä	vesi-ttä
comitative		vesi-ne-(poss)
instructive		vesi-n

However, for Finnish, it is far from clear what the appropriate base vector would be. We cannot simply assume that conceptualization starts with a nominative singular, leaving aside, for now, the complications that arise in Finnish due to the various suffixes and clitics that may attach to nouns inflected for case and number. In other words, following Blevins (2016), we do not assume that a useful pedagogical strategy for second language learners necessarily reflects the cognitive system of native speakers. We therefore generalize the approach pursued by Shafaei-Bajestan et al. (2022), and scrutinize the hypothesis that the semantic vector of a Finnish inflected noun is generated (by a simple additive mechanism) from the imputed semantic vector of the lexeme and the imputed semantic vectors of the inflectional features that are realized in the inflected form. In other words, our aim is to develop a quantitative (statistical) model for the conceptualization of Finnish inflectional semantics.

As shown by Shafaei-Bajestan et al. (2022) and Shafaei-Bajestan et al. (this volume) for English and Chuang et al. (this volume) for Russian, generating models that simply add vectors for lexemes and inflectional features may be very imprecise. When fitting generating statistical models to empirical, corpus-based, semantic vectors, it is necessary to consider whether interactions between inflectional features are required. Furthermore, in a statistical perspective on inflectional semantics, empirical embeddings necessarily come with measurement noise, necessitating the inclusion of an error vector in the generating model. Thus, our goal is to separate the noise in embeddings from the true signal, and to slice the true signal into a series of constituent vectors that represent both lexical and inflectional semantics. In Section 4, we show that considerable headway can be made, but that simple vector addition is only a first step.

As a consequence, if two semantic vectors of Finnish inflected nouns are similar, their similarity then should stem, ex hypothesi, at least in part from the fact that they have similar component vectors. Two inflected nouns can be similar in meaning because they share the same component vector of the stem, or because they share the same component vector for number, or for case. In addition, forms can also be similar because the vectors of their lexemes are similar, as is the case for, e.g., near synonyms.

In what follows, we propose a simple algorithm with which we can estimate the component embeddings in such a way that the error vectors in the empirical embeddings are minimized. In other words, what we are proposing is to conduct ANOVA-like decompositions on word embeddings with (as predictors) well-established semantic features from linguistics. The central goal of our study is to examine what interactions between semantic features are necessary to obtain predicted semantic vectors that are as similar as possible to the corpus-based empirical semantic vectors of inflected Finnish nouns.

2. Finnish noun inflection

Table 1 presents the case by number paradigms for two Finnish nouns. There are two numbers (singular and plural) and fourteen cases. These cases are partitioned into three grammatical cases (nominative, genitive, and partitive) and eleven locative cases. Table 1 does not list the accusative, which for nouns is homonymous with the nominative or genitive forms (and is distinct only for personal pronouns and some interrogative pronouns). Not shown in Table 1 are the five possessive suffixes (1sg., 2sg., 1pl., 2pl., and 3sg/pl.) that can follow the case/number exponents. In addition, there are several clitics (-kO, -kin, -kAAn, -hAn, -pA, -kA, and -s)[2] that realize a range of pragmatic functions. For example, the Finnish word *auto* 'car' has as one of its many inflectional variants the form *autoissanikin*, which can be translated as 'also in my cars':

auto	stem	car
i	number	plural
ssa	case	inessive
ni	possessive suffix	first person singular possessive
kin	clitic	also, too

The schema `stem` + `number` + `case` + `possessive suffix` + `clitic` generates in principle approximately 2,000 possible inflected forms for any noun (see, e.g., Karlsson and Koskenniemi, 1985).

In some nouns additional stem alternations are realized that are described as either qualitative (e.g., *luku* 'number' nom. sg., *luvun* 'number' gen. sg.) or quantitative consonant gradation (e.g., *luukku* 'hatch' nom. sg., *luukun* 'hatch' gen. sg.). The latter is phonological in nature: speakers pronounce a long consonant in open syllables and the corresponding short consonant in closed syllables. However, the process of shortening consonants in closed syllables is not entirely phonologically conditioned as in some word forms it has been partially morphologized, resulting in stem allomorphy. In addition, many inflectional paradigms include certain stem changes for certain cases, such as vowel changes (e.g., *kana* 'chicken' nom. sg., *kanoilla* 'chicken' ades. pl., and *muna* 'egg' nom. sg., *munilla* 'egg' ades. pl.), and consonant changes that are not part of qualitative consonant gradation described above (e.g., *lihas* 'muscle' nom. sg., *lihaksilla* 'muscle' ades. pl., and *patsas* 'statue' nom. sg., *patsailla* 'statue' ades. pl.). Some paradigmatic slots are overabundant, e.g., the genitive plural for the word *paperi* 'paper' is either *paperi-en* or *papere-i-den* (see, Karlsson, 2017 for a more detailed description of stem changes).

2. The capital letters O and A represent vowels the realization of which depends on vowel harmony: o/ö or a/ä.

3. Fasttext-based models of Finnish noun semantics

For this study, we examined the inflectional semantics of the 2,000 most frequent Finnish nouns (excluding compounds) taken from a frequency lexicon of Finnish newspaper language (http://urn.fi/urn:nbn:fi:lb-201405272). We retrieved all available inflected word forms of these nouns from a corpus of Finnish (84,308,641 tokens), which is based on written conversations of thousands of users in a Reddit-like internet community (http://urn.fi/urn:nbn:fi:lb-2017021505). These 2,000 nouns have in all 104,716 different word forms in this corpus, to a total of 10,427,959 word tokens (12.4% of the total number of word tokens in the corpus). For the syncretic forms (10.3% of all word forms (10,766/104,716)), the highest-frequency inflectional function was selected and included in our analyses. An example of syncretism is the word *merkityksensä*, which can be translated as (a) 'of his/her meaning' (sg. gen., corpus freq. 103), (b) 'his/her meaning' (sg. nom., corpus freq. 95), and (c) 'their meanings' (pl. nom., corpus freq. 61). Since interpretation (a) of the word *merkityksensä* has the highest corpus frequency, it was selected as an interpretation of embedding (semantic vector) for this letter string. This selection criterion is motivated by the consideration that the semantic vector of a homophone is a blend of the semantic vectors of the homophone's individual meanings that is weighted by frequency of occurrence. In these blends, the highest-frequency meaning will therefore be best represented by the embedding of the homophone. Therefore, even though the semantic vector of the word m*erkityksensä* is a mixture of three interpretations involving two different numbers (sg. and pl.) and two different cases (nom. and gen.), the sg.gen (a) will influence the semantic vector most, and therefore is the optimal choice for this syncretic form.

We then retrieved the semantic vectors for these word forms, if available, using the pre-compiled `fasttext` embeddings available at https://fasttext.cc. These embeddings are calculated from 127 mil. tokens from the Finnish Wikipedia and from 6 bill. tokens from the Common Crawl project (Grave et al., 2018). `Fasttext` semantic vectors were available for 55,271 unique word forms out of a total of 104,716 inflectionally labelled word forms in our dataset. `Fasttext` embeddings were found for all 2,000 nouns, however, in our original dataset each noun has on average 52.4 different word forms (sd 33.5), whereas in our `fasttext` dataset each noun has on average 27.6 word forms (sd 16.1). Figure 1 presents a dot plot for all the combinations of number and case in this set of 55,271 word forms. A possessive suffix is present for 30.6% of these word forms, and 7.6% of the forms carry a clitic. Figure 2 is a visualization of the number of distinct forms per lexeme in rank order.

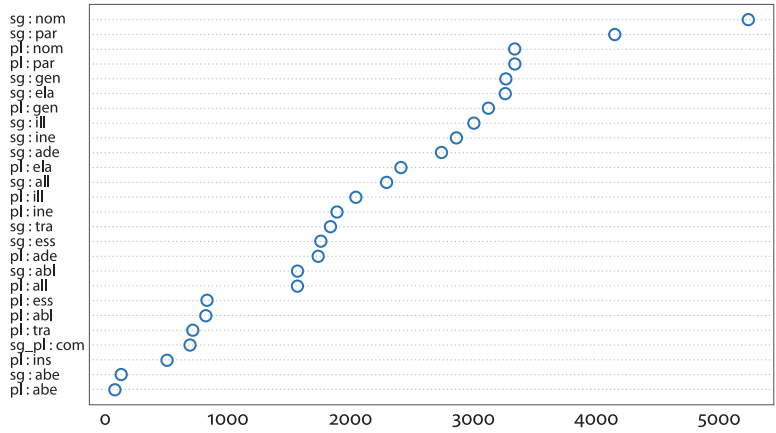

Figure 1. The number and case combinations of 55,271 word forms

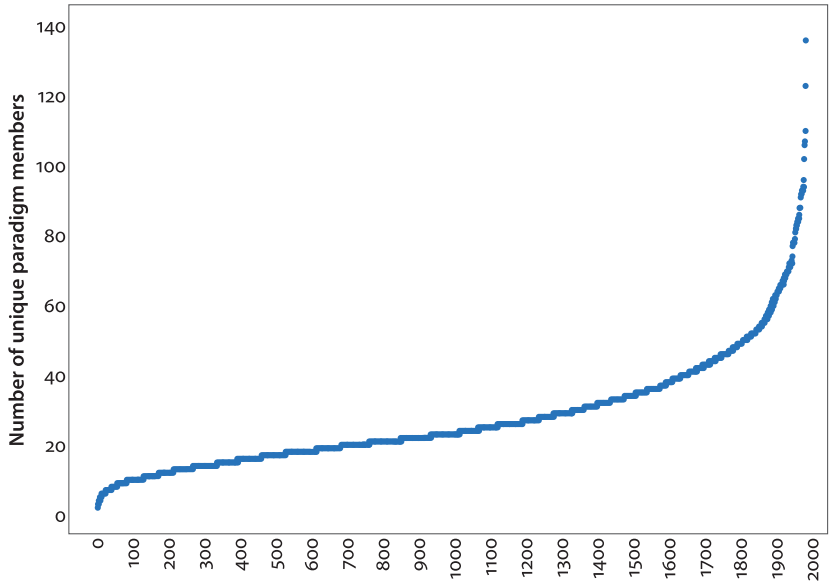

Figure 2. Rank-frequency plot for the number of different distinct forms in a paradigm counted for 2,000 lexemes in the data set of 55,271 word forms

We made use of the *t*-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction technique (van der Maaten and Hinton, 2008; van der Maaten, 2014; Krijthe, 2015), which allowed us to visualize data points in the 300-dimensional distributional vector space of `fasttext` vectors in a two-dimensional plane. We used the default parameters for the `Rtsne` function in the `Rtsne` package (Krijthe, 2015) for R (perplexity = 30, iterations = 1000). However, to

be sure that our 2D projections are robust, we replicated nearly identical figures with perplexity set to 5 or to 50 and with 500 and 3000 iterations. In what follows we first explore, step by step, how Finnish inflected words cluster in the t-SNE map.

Figure 3 presents the t-SNE map with color coding by case. The 55,271 word forms produce clear case-based clusters with in general remarkably little overlap. Grammatical cases (nominative, partitive, and genitive) are located more centrally in comparison to locative cases (e.g., inessive or adessive). One particular case, abessive, which constitutes only 0.1% of our data, does not show good clustering. By contrast, the comitative, which also constitutes only 0.1% of our data, does show clear clustering. This suggests that it is the semantics of the abessive that are less systematic, and more idiomatic. Possibly, the abessive is loosing its productivity as a case of Finnish.

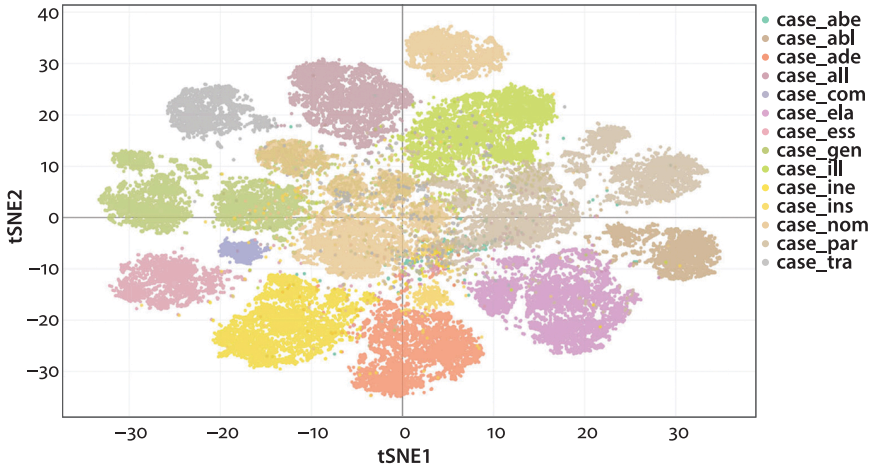

Figure 3. Finnish inflected nouns in the 2D plane constructed by the t-SNE clustering algorithm applied to the fasttext embeddings of the nouns. Colors represent cases. Nouns cluster by case in the t-SNE plane (view interactive plot – https://doi.org/10.1075/ml.22008.nik.fig3)

Figure 4 presents exactly the same t-SNE map, but now color-coded by number. The three grammatical cases are the only cases that have distinct clusters for number, one for singular forms and another for plural forms. The three completely green clusters in the upper half of Figure 4 represent the plural forms of the grammatical cases. Apart from a small group of forms that are ambiguous with respect to number (the blue cluster in the lower left quadrant), plurals (green) are found at the outer periphery of the plot. A comparison with Figure 3

clarifies that for non-grammatical cases, number is realized within the case cluster away from the origin of the map. For the grammatical cases, plurals form separate clusters that are also on the periphery. This topography of case and number may reflect differences in the type of inflection realized for these nouns: inherent inflection for number, but contextual inflection for case (see Booij, 1996, for these types of inflection). (Note that number is realized closer to the stem, cf. Bybee (1985) for detailed discussion of proximity to the stem.) Number is thus found to form sub-clusters within major case clusters. Most important, however, is the observation that how plurality is conceptualized varies with case. Chuang et al. (this volume) report a similar finding for Russian nouns.

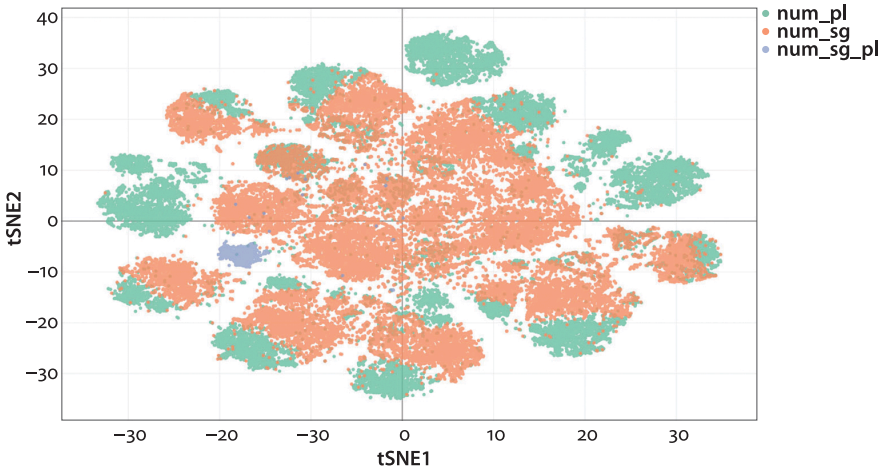

Figure 4. Finnish nouns in the t-SNE 2D map. Colors represent the category of number (sg, pl). Singular and plural word forms form distinct clusters within case clusters for all non-grammatical cases (ill, abl, etc.). The three completely green clusters in the upper half of the plot represent the plurals of the three grammatical case (nominative, partitive, and genitive) (view interactive plot – https://doi.org/10.1075/ml.22008.nik.fig4)

To further consolidate this finding, we calculated the vectors that, when added to the vector of the singular in a given case, result in the vector of the plural of the same lexeme, in the same case. We refer to these vectors as shift vectors. For any given noun, we have in principle 12 shift vectors. (Of the 14 cases, the instructive and comitative do not occur in the singular.) Given the way that plurals cluster within the different cases, we expect the shift vectors to cluster in 'shift space'. Figure 5 shows that this is indeed the case. Apparently, the semantics of number and case are not independent, but interact (for a similar conclusion, see Karlsson, 1985). The conceptualization of a plural form depends on its case.

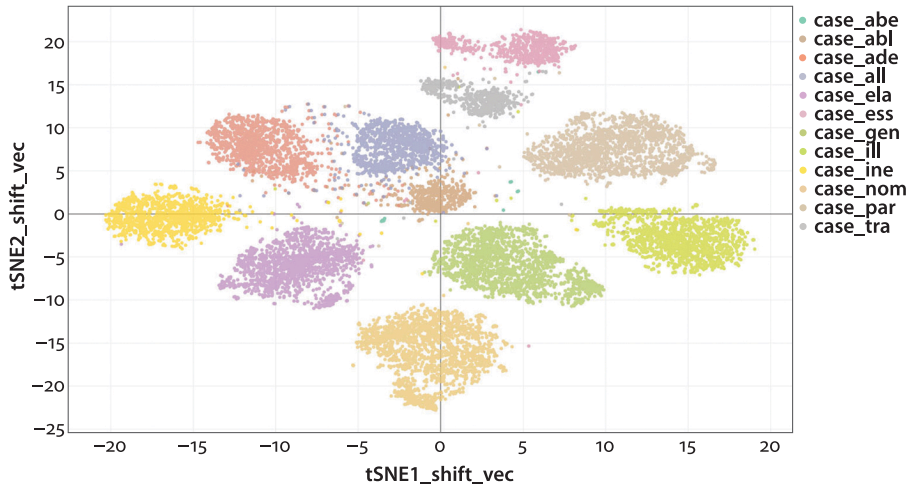

Figure 5. Shift vectors differ for number when case is fixed (view interactive plot – https://doi.org/10.1075/ml.22008.nik.fig5)

Finnish nouns also allow possessive suffixes (e.g., *-ni* 'mine', *-si* 'yours') to follow the case endings. Figure 6 shows that words with these suffixes also group into clusters. Within some case clusters, they are found closer to the origin, suggesting that they are less well differentiated than plurals, perhaps because forms with possessive suffixes are less common (in our data, suffixes are found only in 30.6% of the forms). This figure also suggests that second person singular possessive suffixes (coded in light green) have a strong preference for only a limited number of cases. Finally, Figure 7 indicates that there is some clustering, often within case, of the word-final clitics, specifically for the clitic *kin* 'also, too', presented in purple, which is generally found within case clusters closer to the origin of the plot.

4. A generating model for nominal conceptualization

In order to construct a model that generates semantic vectors for Finnish nouns, we take inspiration from the generating models in statistics that underlie analysis of variance. Following Baayen et al. (2019) (see also Boleda, 2020, for a review of studies using vector addition for derivation), we can set up a straightforward model for the Finnish noun inflections that takes the semantic vector of a word w_i to be the sum of the vectors of its stem and the vectors of its inflectional specifications for number, case, possessive suffix and clitic:

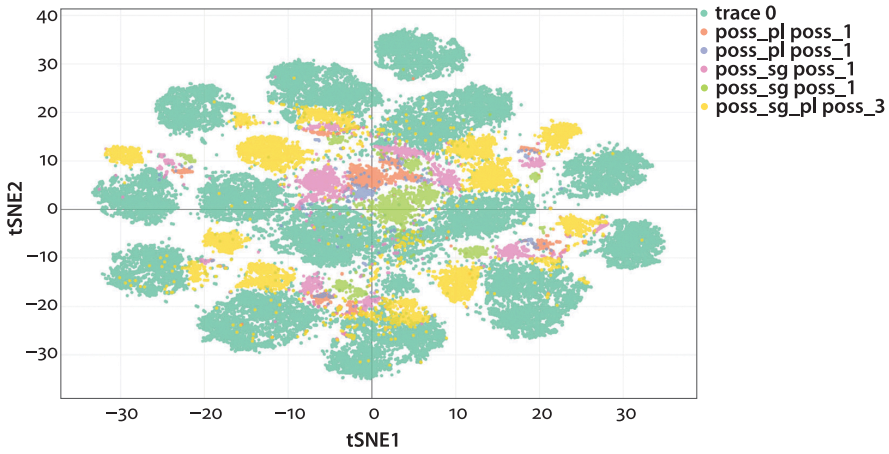

Figure 6. Finnish nouns in the t-SNE map, color-coded for possessive suffixes. The clusters indicate noun semantics reflect number and person of these suffixes, but in ways that depend on case and possibly number (view interactive plot – https://doi.org/10.1075/ml.22008.nik.fig6)

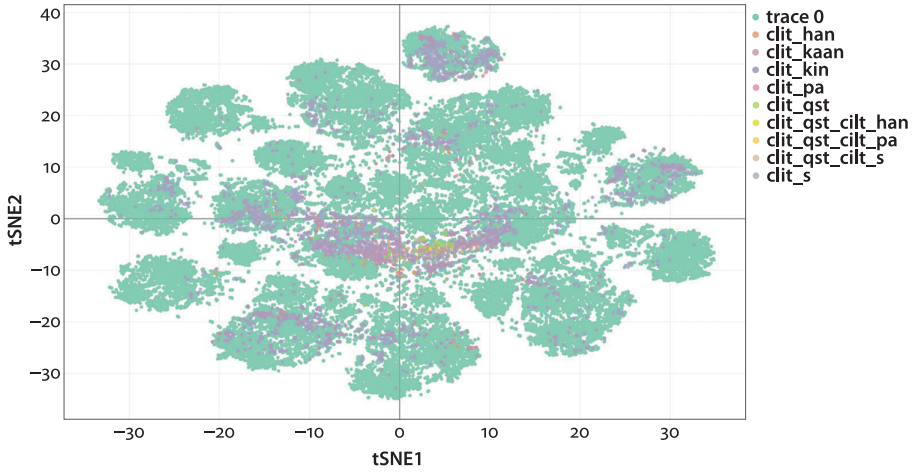

Figure 7. Finnish nouns in the t-SNE map, color-coded by clitics. Noun embeddings show clear clustering also by these extra-paradigmatic exponents (view interactive plot – https://doi.org/10.1075/ml.22008.nik.fig7)

$$\overrightarrow{w_i} = \overrightarrow{stem_i} + \overrightarrow{number_i} + \overrightarrow{case_i} + \overrightarrow{possessive_i} + \overrightarrow{clitic_i} + \vec{\varepsilon}_i. \quad (1)$$

(The details of model setup and vector estimation will be introduced below.) In (1), $\vec{\varepsilon_i}$ is the residual semantic vector that represents the combination of, on the one hand, word-specific semantics that cannot be explained by the 'main-

effects' of stem, number, case, possessiveness, and clitic, and, on the other hand, the measurement error of `fasttext` for word w_i.

The model given by Equation (1) defines a classical 'decompositional' realizational model of inflectional morphology, in which case and number are realized simultaneously on the noun. However, the t-SNE plots indicate that case and number are not independent, but interact (see also Shafaei-Bajestan et al. (this volume) and Chuang et al. (this volume)), suggesting that the above main effects model is too simple. We return to this issue below, as it is convenient to first explain how we estimate a word's component semantic vectors.

First, note that Equation (1) builds a vector representation for a word using a series of vectors that are themselves not observable: We do not have embeddings for stems or inflectional endings. We therefore have to impute these vectors. The solution that we have implemented for imputing these latent vectors proceeds as follows. We first define a matrix $\boldsymbol{L}$ that has as many rows as there are word forms, and as many columns as there are distinct lexemes and grammatical features. The entries in row i of this matrix are 1 or 0 depending on the semantic features that a given word w_i has. By way of example, consider the $\boldsymbol{L}$ matrix for just the four inflected forms *(lasi, vesi, lasit*, and *lasin)*:

$$\boldsymbol{L} = \begin{array}{c} \\ \text{lasi} \\ \text{vesi} \\ \text{lasit} \\ \text{lasin} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{glass} & \text{water} & \text{singular} & \text{plural} & \text{nominative} & \text{genitive} \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Next, we set up a matrix $\boldsymbol{S}$ that contains as its row vectors the `fasttext` embeddings of the words. We now want to find a matrix $\boldsymbol{Q}$ such that

$$\boldsymbol{LQ} = \boldsymbol{S} \tag{2}$$

Equation (2) is formally identical to a multivariate multiple regression model with predictors brought together in $\boldsymbol{L}$, multiple response variables brought together as the columns of $\boldsymbol{S}$, and $\boldsymbol{Q}$ the matrix of beta weights.[3] The row vectors of $\boldsymbol{Q}$ provide us with the imputed embeddings for stems, singulars, plurals, nominatives, and genitives:

3. For standard multiple regression, the corresponding equation takes the form $\boldsymbol{X\beta} = \boldsymbol{y}$.

$$Q = \begin{array}{c} \\ \text{glass} \\ \text{water} \\ \text{singular} \\ \text{plural} \\ \text{nominative} \\ \text{genitive} \end{array} \begin{array}{c} \begin{array}{cccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & \ldots & d_{300} \end{array} \\ \begin{pmatrix} 1.34 & 0.55 & 1.33 & 1.43 & 1.07 & -1.39 & \ldots & -0.77 \\ -0.60 & 1.54 & -0.72 & -0.02 & 1.33 & 1.02 & \ldots & 0.98 \\ -0.82 & 0.64 & 0.76 & -0.13 & 1.82 & -0.36 & \ldots & -0.54 \\ -1.45 & -0.51 & 0.72 & 0.58 & 1.44 & -0.15 & \ldots & 0.46 \\ -0.72 & -0.48 & -0.45 & -0.43 & 0.49 & 0.30 & \ldots & -0.21 \\ 1.06 & -0.19 & 0.92 & -0.35 & 1.42 & -2.09 & \ldots & 1.00 \end{pmatrix} \end{array}.$$

We solved (2) using standard methods from linear algebra (see the supplementary materials for Julia code). However, our current method is bound to be imprecise due to collinearity in $\boldsymbol{L}$ (compare, for instance, the columns for singular and plural, which are each other's mirror image). The development of a numerically more optimal estimation method is beyond the scope of the present – exploratory – study. (Alternatively, the latent vectors could be calculated using averaging of the vectors sharing combinations of features, see Chuang et al., this volume, for Russian nominal inflection.) Importantly, post-multiplication of $\boldsymbol{L}$ with $\boldsymbol{Q}$ simply amounts to adding those semantic 'primitive' vectors in $\boldsymbol{Q}$ that are relevant for the word forms (as specified on the rows of $\boldsymbol{L}$).

The multiplication $\boldsymbol{LQ}$ results in predicted semantic vectors, the row vectors of an estimated (or predicted) matrix $\hat{\boldsymbol{S}} = \boldsymbol{LQ}$. In order to evaluate how accurate the predicted semantic vectors of $\hat{\boldsymbol{S}}$ are, we calculated, for each predicted vector $\hat{\boldsymbol{s}}$ its correlations with all the 'gold standard' semantic vectors in $\boldsymbol{S}$ (i.e., the empirical `fasttext` vectors). The predicted semantic vector is considered accurate if it is better correlated with its corresponding gold standard `fasttext` embedding than all the other gold standard embeddings. (We gauge semantic similarity using the Pearson's Product-Moment Correlation measure; nearly identical results are obtained when the correlation measure is replaced by the cosine similarity measure.)

Table 2 presents the accuracies for a sequence of increasingly complex generating models. As expected, a model with only stems performs worst, at 3.6% accuracy. Adding number leads to a small increase in accuracy by 3.4%. A model with stem and case performs better (35.7%), and a full 'main effects' model achieves 75.6% accuracy. The addition of an interaction of case and number improves accuracy even further (82.4%). However, the best model requires a four-way interaction of number, case, possessive, and clitic, resulting in 89% accuracy. In other words, for every combination of number, case, possessive, and clitic, we are estimating a semantic vector that will contribute (through vector addition) to the semantic vector of the inflected noun. The substantial improvement in model fit achieved with this interaction suggests that morpho-syntactic interactions are part and parcel of Finnish inflection.

Table 2. Accuracies of the generating models for Finnish inflected nouns (based on `fasttext` vectors)

Accuracy	Model
3.6%	$\vec{w}_i = \overrightarrow{stem}_i + \vec{\varepsilon}_i$
7%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \vec{\varepsilon}_i$
35.7%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{case}_i + \vec{\varepsilon}_i$
52%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \vec{\varepsilon}_i$
75.6%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \overrightarrow{poss.}_i + \overrightarrow{clit.}_i + \vec{\varepsilon}_i$
82.4%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \overrightarrow{poss.}_i + \overrightarrow{clit.}_i + \overrightarrow{num.\ case}_i + \vec{\varepsilon}_i$
89%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \overrightarrow{poss.}_i + \overrightarrow{clit.}_i + \overrightarrow{num. : case : poss. : clit.}_i + \vec{\varepsilon}_i$

This conclusion is supported by inspection of the error vectors of this model ($\vec{\varepsilon_i}$). We are trying to find a model that generates the data apart from the measurement error that comes with the `fasttext` vectors for Finnish nouns, and possible semantic idiosyncracies of individual word forms (see Sinclair, 1991 for the specific collocational patterns that different inflected variants of the same lexeme may have). If the model fits the data, the error vectors should be random, and not reveal any structure upon inspection. As shown by Figure 8, the errors of the model with only the main effects of number, case, possessive, and clitic show considerable structure, perhaps unsurprisingly, as the model misses out on, for instance, the important interaction of case and number.

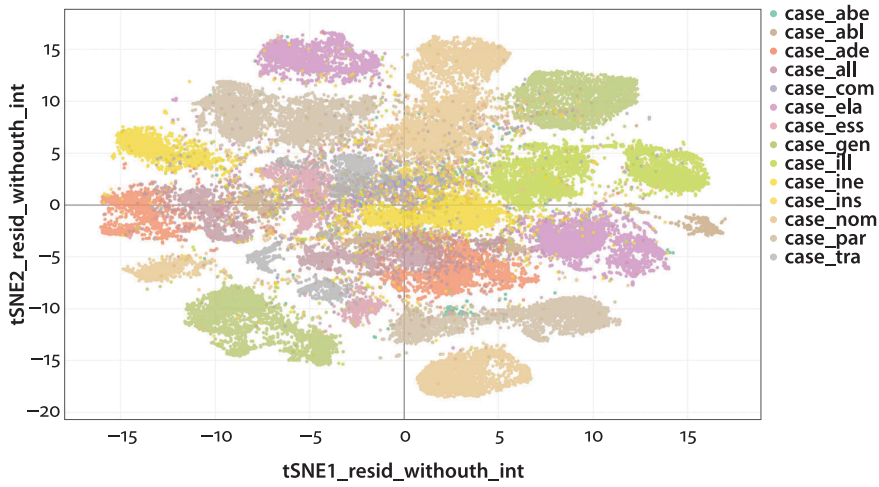

Figure 8. Visualization of residuals of the Model meaning = lexeme + case + number + possessive + clitic. Colors represent cases. Residuals form separate clusters for the category of case (view interactive plot – https://doi.org/10.1075/ml.22008.nik.fig8)

The errors of the model with the four-way interaction reveal considerably less structure (see Figure 9). We also calculated, for each word's error vector, the sum of squared errors from the two models. We found that the errors of the model with only the main effects (Figure 8) are overall larger than those of the model with interaction terms (Figure 9) ($t(55271) = 189.61$, $p < .0001$). This indicates that further subtle semantic information pertinent to inflectional features is indeed successfully accounted for by including the interaction. We leave constructing even more complex models, for instance, by allowing interactions with lexeme modeled as a random-effect factor, to follow-up research.

What we have shown is that it is possible to develop a surprisingly accurate generating model for Finnish inflected nouns. In the next section, we examine the kind of errors made by our best-performing model.

5. Error analysis

In this section we discuss what kind of errors our best generating model produced, as this contributes to a better understanding of its qualitative performance. In what follows, we discuss different types of errors according to their frequency (the most frequent type being discussed first).

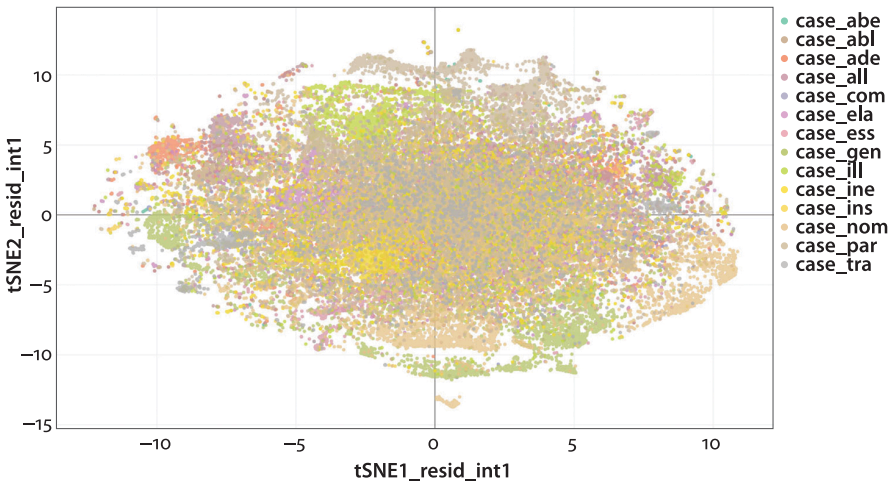

Figure 9. Visualization of residuals of the Model meaning = lexeme + case + number + possessiveness + clitic + case : number : possessive : clitic. Colors represent cases. These residuals are far less structured, but there is still some clustering, indicating this model is still missing out on inflectional structure (view interactive plot – https://doi.org/10.1075/ml.22008.nik.fig9)

The most frequent type (16.5%) was the stem exchange error, when, e.g., the model predicted *pentuun* 'into a puppy' but was evaluated on (was expected to produce) *koiraan* 'into a dog' *(koiraan* is also homonymous with the gen. sg. form of the word *koiras* 'male'). We discuss stem exchange errors in more detail later in this section.

The second most frequent type (9.9%) included a bound morpheme which expressed the category of number, e.g., *siskolta* 'from a sister' was evaluated, but *siskoilta* 'from sisters' was predicted by the model (the direction was more often sg. evaluated → pl. predicted than pl. evaluated → sg. predicted with a ratio of 5:1).

The third most frequent type (7.5%) of errors the model produced is due to overabundance. In the case of overabundance, more than one form is available for a given paradigm slot. For instance, for the noun *tulos* ('result'), the slot for the genitive plural contains two acceptable forms, *tulosten* and *tuloksien*. Even though these word forms have different stem allomorphs and different case allomorphs, they both are acceptable candidates to fill the slot of genitive plural. The model predicted *tuloksien*, but was evaluated on *tulosten*. Or, e.g., the slot for the 3rd person possessive suffix contains also two acceptable forms: *kädestään* and *kädestänsä* 'from his/her/their hand'. In other words, both of these word forms (e.g., *tulosten* and *tuloksien)* could be used in the corpus, but the model predicted not the one on which it was evaluated. Since grammatically these are not real errors, we may exclude them, in which case model accuracy increases to 92%.

Clitic errors constituted 6.4% of the errors. The most frequent error with the clitics was the clitic *-kin* 'also', which the model incorrectly predicted (somewhat counter-intuitively) to be present no less than 339 times (86% of all clitic errors).

The fifth most frequent category were the errors with possessive suffixes (4.3%). The most frequent error was for the possessive suffix for 3rd person sg./pl. (201 out of 260 errors). Most often the model predicted no suffix when it was evaluated by a word form with the suffix (156 out of 201). This suffix for 3rd person sg./pl. has the greatest allomorphy (the highest number of different forms) compared to 1st and 2nd person (sg. or pl.) possessive suffixes. Therefore, when the model produces an error with a possessive suffix, its performance can be traced to the higher complexity and greater allomorphy of this possessive suffix.

The sixth most frequent category of errors were case errors (3.7%, 228 out of 6082 errors). Almost all of them were errors in the singular form (205 out of 228). When the model prediction was an error, it was more likely to appear in more frequent (grammatical) cases (partitive (54 errors), nominative (46 errors), or genitive (35 errors)). In other words, when the model produced case errors, it tended to predict cases that were more frequent than the targeted cases. There are almost 3 times as many number errors (9.9%) than case errors (3.7%). One potential explanation of why there are more number than case errors could be that case

is meaning-wise more impactful than number. This would be in line with how the t-SNE algorithm clustered the data points (primarily according to case and secondarily according to number, see Figures 3 and 4).

So far we have explained nearly half of all errors the model produced (48.3%). Another half (51.7%) can be explained by different combinations of the categories described above. E.g., case & number errors account for 1% of all errors, and stem exchange errors in combination with clitic errors account for 0.9% of all the errors.

Another potential source of errors is the homophony that exists for some Finnish nouns. For example, the noun form *tuli* 'fire (nom.sg.)' is homophonous with the verb form *tuli* 'come (past tense)'. Likewise, out of context the word form *tuletko* can be translated either as 'fires?' or as 'will you come?'. Such homophones can lead to suboptimal `fasttext` vectors – these vectors are likely to provide some frequency-weighted average of the different meanings of the homophones. Imprecise embeddings for homophones unavoidably give rise to imprecision in the predictions of our model.

To investigate a possible influence of homonymous forms on model performance, we removed from our data set of Finnish nouns all the word forms that happen to be in use as verbs in our corpus. After removal of 1,286 homophonic nouns (2.3% of the dataset), we refitted our best generating model. Although the accuracy of the refitted model increased, it did so only by a tiny fraction (from 89% to 89.53%). Therefore, homophony can be ruled out as a major source of errors.

Above, we promised more detailed information on stem exchange errors. A closer inspection of these errors revealed that they all involve stems that are semantically related. Further inspection using a graph representation revealed that some exchange errors were clustered. We represented words as the vertices of a directed graph. Whenever a word *i* was incorrectly predicted as word *j*, we placed a directed edge from *i* to j. The resulting graph had a large number of very small components, the vast majority of which comprised only two vertices. Here, the expected word was replaced by a semantically similar word with a different stem, but with the same morphological exponents (e.g., pl. partitive). However, there was a small number of components with more edges, and their structure turned out to be informative about the nature of the errors our model makes.

In these non-trivial components of the graph, all words were semantically related and included the same morphological exponents. In these clusters, one word is selected by the model as the replacement for all other words in a cluster. In order to clarify why a particular word form was selected by the model to be the replacement target for several other word forms, we compared the embeddings of the words (using `fasttext` vectors) in a graph component. We did not select a

particular word as an example, but rather we selected a particular graph component, in which one word was erroneously predicted by the model for several other words. For instance, the word was (falsely) predicted by the model as a replacement for all of the following seven expected words:

ahdistuksia (pl.partitive)	'anxiety'
levottomuuksia (pl.partitive)	'worry'
painajaisia (pl.partitive)	'nightmare'
vihoja (pl.partitive)	'hate'
oloja (pl.partitive)	'feeling / mental state'
pelkoja (pl.partitive)	'fear'
kuolemia (pl.partitive)	'death'
harmeja (pl.partitive)	'nuisance'

To put it simply, instead of predicting *levottomuuksia* for *levottomuuksia* and *painajaisia* for *painajaisia* etc., the model predicted *ahdistuksia* for all of these eight words. Since there should be a reason for the model to favor this particular word over seven other words, our expectations were the following: (a) the semantic vectors of all these eight words must correlate with each other, and (b) the word *(ahdistuksia* 'anxiety) that was substituting in our model for all other words of this component must be central from the perspective of the network analyses. Therefore, we calculated several centrality indices, which typically quantify the relative importance of the node (the word form in our case) in relation to other nodes (other word forms) in the network.

Figure 10 depicts a network for the eight words discussed above. This network is based on the Spearman correlations between the embeddings of these eight words. We used the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996) to obtain a conservative (sparse) network model for the covariation structure in the data.[4] Figure 10 shows that, as expected, the word *ahdistuksia* 'anxiety' plays a central role in this network. It has the greatest number of edges (5), and as shown in Figure 11, *ahdistuksia* has the highest value on three

4. The tuning parameter was selected by minimizing the extended Bayesian information criterion (EBIC, Chen and Chen, 2008). The network analysis was carried out with the `bootnet` package (Epskamp et al., 2018).

centrality indices.[5] This example illustrates how semantically central words function as attractors for semantically similar words.

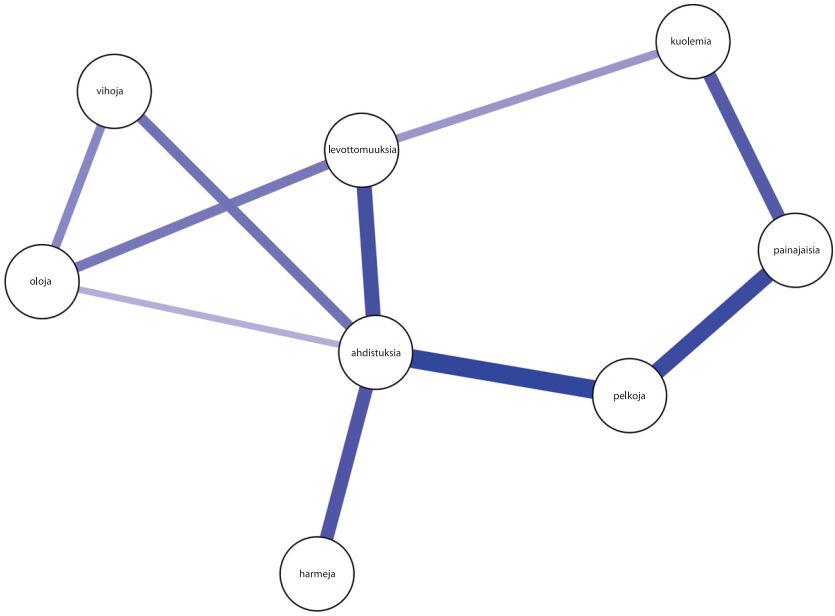

Figure 10. Estimated correlation-based network for eight words. Line width is proportional to semantic similarity. *ahdistuksia* is the most central word in this network, and it is this word that is predicted by the conceptualization model instead of the other words in the network

In clitic errors and in possessive suffix errors there was one particular exponent in each group that was the most frequent. For clitics it was the clitic *-kin* 'also' and for possessive suffixes it was the suffix for 3rd person. The question arises why exactly these exponents were particularly difficult for the model to predict correctly. To answer this question, we calculated the L2 norms (lengths) of all the vectors for stems, cases, numbers, possessives, clitics, and their interactions. Figure 12 presents the estimated density of these lengths (blue line). Superimposed on this are indicators for the lengths of the vectors for case (red), number (purple), possessives (green), clitics (black), and the interaction terms (yellow).

5. Figure 11 reports the strength of the interactions that a node has with its neighbor nodes, betweenness, i.e., the number of times a node lies on the shortest path between two other nodes, and closeness, or how well a node is indirectly connected to other nodes. In each of these definitions of centrality, a node can be somewhere on the continuum from central to peripheral. Figure 11 was produced using the `qgraph` package (Epskamp et al., 2012).

The clitic *-kin* has a relatively short vector compared to other clitics, and this may have rendered it more confusable. Also the possessive suffix for the 3rd person has a relatively short vector compared to other possessive suffixes. However, vectors for, e.g., case are also relatively short, so it remains unclear to what extent vector length as such is driving prediction errors. Alternatively, the relatively low prevalence of clitics in our dataset (only 7.6%) may have rendered our estimates of the clitics' predicted vectors imprecise especially as the range of their distances is large in comparison to, e.g., case markers (each noun in our data set had a case marker). Clitics are also the most separated from the stem, and are not part of the core inflectional paradigm: they convey some discourse-structural or speech-act-type information, which might have less to do with the lexeme in question and more with what function the surrounding phrase has.

Importantly, some of conceptualization errors that our model makes are similar to some of the semantic substitutions shown in the literature of speech errors and aphasia (e.g., Laine et al., 1992). We leave further investigations of the semantic errors made by our model to future modeling studies.

6. Models based on word2vec instead of on fasttext

`Fasttext` vectors are constructed from word co-occurrences, but boosted with sub-word co-occurrences. The developers of `fasttext` argue that this is essential for languages with more complex morphology. In such languages, the number of attested (let alone possible) words is so large that any attempt to construct semantic vectors from orthographic words will be shipwrecked on the harsh rocks of data sparsity. However, as `fasttext` vectors have access to letter substrings, the possibility cannot be ruled out completely that in part they are representing morphological form in addition to meaning.

We therefore investigated to what extent the results obtained with `fasttext` embeddings replicate when we turn to `word2vec` vectors. The `word2vec` embeddings that we used were calculated by the University of Turku NLP group (https://turkunlp.org) from 4.5 bill. tokens from the Finnish Internet Parsebank project (https://turkunlp.org/finnish_nlp.html#parsebank) and from 2 bill. tokens from the suomi24 corpus (written conversations in a Reddit-like community). Therefore, the size of the training data for `fasttext` (6.1 bill tokens) and for `word2vec` embeddings (6.5 bill tokens) is comparable. `Word2vec` embeddings were available for 88,406 out of 104,716 word forms (substantially more word forms than in the `fasttext` vectors available for 55,271 out of 104,716 word forms). `Word2vec` embeddings were found for all 2,000 nouns. Each lexeme has on average embeddings

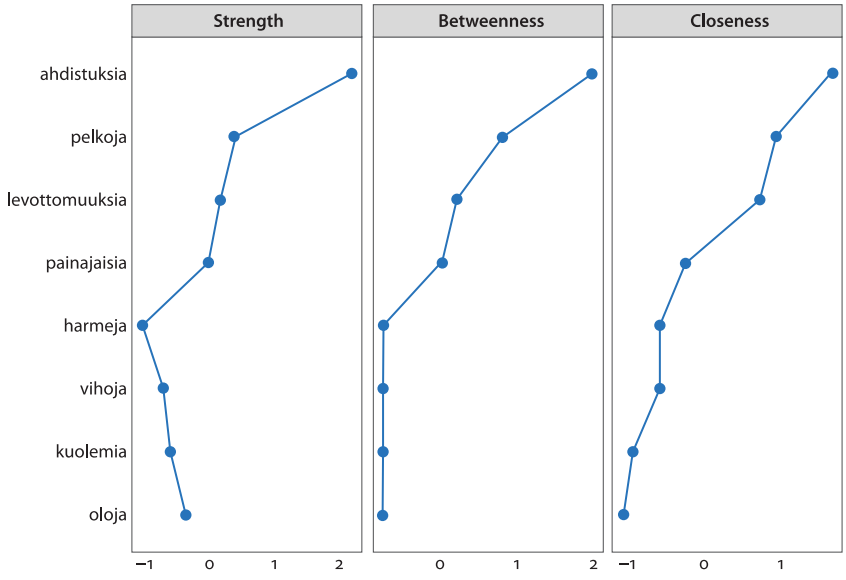

Figure 11. Centrality indices: Strength (how well a node is directly connected to other nodes); Betweenness (shows how important a node is in the average path between other nodes); Closeness (how well a node is indirectly connected to other nodes)

for 44.2 word forms (sd 27). This is less than in our original dataset, in which each noun has on average 52.4 different word forms (sd 33.5), but more than in our `fasttext` analysis, in which each noun has embeddings for 27.6 word forms (sd 16.1).

For the `word2vec` embeddings, t-SNE plots show substantially less structure than the `fasttext` embeddings (see the supplementary materials for further details). A 2D map for the shift vectors for number, conditioned on case, revealed clustering by case, but these clusters are somewhat more diffuse. However, the most striking difference between the `fasttext` vectors and the `word2vec` vectors emerged when we compared the predicted semantic vectors with the observed semantic vectors. While for the `fasttext`-based model we obtained an accuracy of 0.89, the corresponding accuracy for the `word2vec`-based model was only 0.401 (for other accuracies, see Table 3).

There could be many possible reasons for the underwhelming performance of `word2vec`. Possibly, since the dataset for which we have embeddings is substantially larger when word2vec is used, conceptualization with `word2vec` faces a more challenging task. Alternatively, `word2vec` is restricted to word contexts, which for Finnish are extremely sparse, and hence inevitably do not provide a solid foundation for generalization.

Table 3. Accuracies of the generating models for Finnish inflected nouns (based on `word2vec` vectors)

Accuracy	Model
2.2%	$\vec{w}_i = \overrightarrow{stem}_i + \vec{\varepsilon}_i$
3.6%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \vec{\varepsilon}_i$
16.5%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{case}_i + \vec{\varepsilon}_i$
21.6%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \vec{\varepsilon}_i$
30.1%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \overrightarrow{poss.}_i + \overrightarrow{clit.}_i + \vec{\varepsilon}_i$
36.3%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \overrightarrow{poss.}_i + \overrightarrow{clit.}_i + \overrightarrow{num.\ case}_i + \vec{\varepsilon}_i$
40.1%	$\vec{w}_i = \overrightarrow{stem}_i + \overrightarrow{num.}_i + \overrightarrow{case}_i + \overrightarrow{poss.}_i + \overrightarrow{clit.}_i + \overrightarrow{num.:case:poss.:clit.}_i + \vec{\varepsilon}_i$

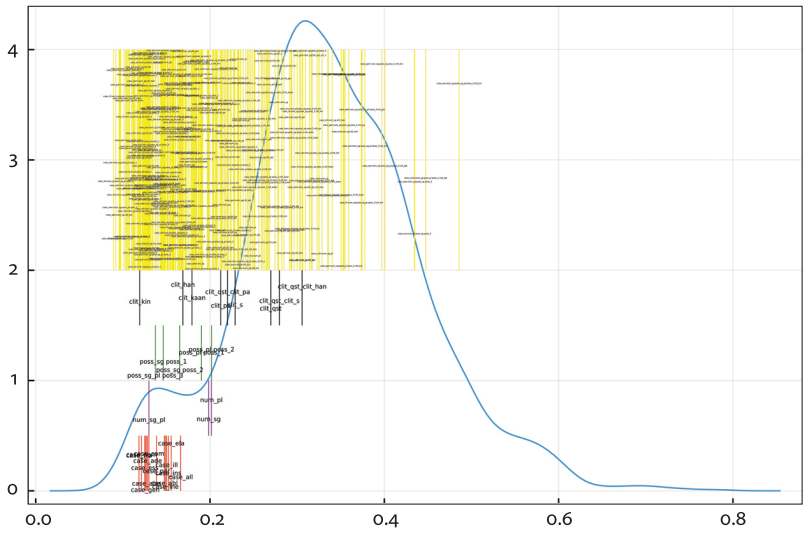

Figure 12. Estimated density for the distribution of Euclidean length (L2-norms, in blue). The positions of inflectional vectors are indicated by vertical lines. The red vertical lines represent the locations of the cases; the purple lines highlight number; the green lines locate the possessive suffixes and the black lines the clitics. The yellow lines denote the locations of the interaction terms in the model. The clitic *-kin* has the shortest vector of all clitics, which may have rendered it more error-prone

In the light of the excellent performance in general of `fasttext` vectors on a range of NLP tasks (see, e.g., Shahmohammadi et al., 2021), it seems unlikely that `fasttext` works so well primarily (or only) because it would be highly sensitive to form similarities. It is worth noting that strings similar to affixes occur in many, often highly frequent words (e.g, *un* in *uncle, er* in *her* and *beer*, as discussed in cf. Schreuder and Baayen, 1997; for Finnish, see Nikolaev et al., 2019), and

that as a consequence, it is not the substring itself that is of primary importance, but its contextual and distributional statistics. In the light of these considerations – specifically, superior clustering and substantially more precise decomposition of empirical vectors – we conclude that `fasttext` is the superior instrument for gauging inflectional semantics in Finnish.

7. Discussion

We have shown that empirical semantic vectors, obtained with `fasttext`, can be approximated with high accuracy by a generating model. This model imputes vectors for lexemes, cases, numbers, possessives, and clitics, and adds the pertinent vectors for a Finnish inflected noun to obtain its predicted semantic vector. This generating model implements core insights from realizational morphology, as building up word meanings from component vectors in our generating model provides a formalization of feature bundles in realizational morphology. Moreover, the model provides a quantitative theory for the conceptualization of inflection.

An important property of this model is that there is no need to derive the meaning of one inflected form from the meaning of another inflected form. This approach to inflectional conceptualization is independent from the possible role of principal parts when realizing semantics in form. For Finnish *lasi*, the nominative singular is the principal part, but for *vesi*, it is the partitive singular that is the principal part (see Table 1). Importantly, as pointed out by Blevins (2016), the meaning of, for instance, the allative *vedelle* is not a semantic function of the meaning of the partitive singular (*vettä)*. The model for conceptualization that we are proposing implements a straightforward additive decomposition. An alternative approach to modeling conceptualization would be to extend the model developed by Marelli and Baroni (2015) for derivation. This model takes the vectors of base words, and sets up a linear transformation that maps the base word vectors onto the vectors of the corresponding derived words. However, it is difficult to see how their approach might be generalized to Finnish nominal morphology in a straightforward way. One concern here is that in Finnish, nouns are always inflected. As a consequence, we do not have vectors for 'bare stems' (unless we impute them). Another concern is that the number of possible forms of a Finnish noun is huge, whereas the number of nouns for which all forms are actually attested in corpora is very small: even for the most frequent lexeme only a fraction of the inflectional word-forms can be observed, cf. Karlsson, 1986). A related concern is that it is unclear how to avoid some form of ordering of semantic operations while at the same time doing justice to interactions such as observed for

case and number (for further discussion in the context of English noun pluralization, see Shafaei-Bajestan, this volume).

The results that we have obtained are contingent on the validity, or perhaps, reasonableness, of the vectors that we used to represent the meanings of Finnish inflected nouns. We made use of `fasttext`, which proved to provide superior results compared to `word2vec`. We acknowledge that `fasttext` vectors are far from perfect. For instance, we are not sure that they are able to properly capture the subtle semantics and pragmatics of the clitics. As Brunila and LaViolette (2022) point out, many people working in natural language processing seem to have lost their interest in linguistic theories and compensate this by elevating the ideas of Harris and Firth (e.g., Firth, 1968; Harris, 1954) to canonical status. We, on the other hand, think that the distributional hypothesis is too narrow. For instance, `fasttext` vectors have no visual grounding, they are based only on texts and do not integrate knowledge of what things and events in the world actually look like. A technique for fusing visual information into textual embeddings (by using images and their image captions) has been found to improve the quality of embeddings, see Shahmohammadi et al. (2021) and also Shahmohammadi et al. (this volume). Hence it is worth investigating whether visual grounding will also yield improved vectors for Finnish. Possibly, visually grounded vectors will allow us to formulate not only more precise models for conceptualization, but also simpler models, i.e., models with fewer interactions.

An important challenge is to further whiten the residuals of our conceptualization model. Even for our best model, there is still some structure left. Exploratory analysis using mixed models suggests that including interactions with lexeme in a linear mixed modeling framework leads to improved prediction accuracy. We are currently researching how this framework can be used to impute the primitive semantic vectors for lexemes, inflectional meanings, and their interactions.

Placing our results in the more general perspective of construction grammar, we think that "cases" represent sets of constructions that, as such, have lexis in common. One particularly salient property of the Finnish lexicon that emerges is that these "case constructions" are pervasive in the lexicon, and provide it with its most visible similarity structure.

It is clear that many questions remain for future research, but we hope to have shown that a decompositional approach to conceptualization in a richly inflecting language such as Finnish has potential to enrich our understanding of the conceptualization processes underlying the realization of inflected words.

Funding

This research was made possible by funding from the ERC, project WIDE-742545.

References

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, *6*:213–234.

Booij, G. E. (1996). Inherent versus contextual inflection and the split morphology hypothesis. In Booij, G. E. and Marle, J. V., editors, *Yearbook of Morphology 1995*, pages 1–16. Kluwer Academic Publishers, Dordrecht.

Brunila, M. and LaViolette, J. (2022). What company do words keep? revisiting the distributional semantics of jr firth & zellig harris. *arXiv preprint arXiv:2205.07750*.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Benjamins, Amsterdam.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3):759–771.

Chuang, Y. Y., Brown, D., Evans, R. and Baayen, R. H. (2022). Paradigm gaps are associated with weird "distributional semantics". *Russian defective nouns and their case and number paradigms*.

Epskamp, S., Borsboom, D., and Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior research methods*, *50*(1):195–212.

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., and Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of statistical software*, *48*:1–18.

Firth, J. R. (1968). *Selected papers of J R Firth, 1952–59*. Indiana University Press.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Günther, F., Rinaldi, L., and Marelli, M. (2019). Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science*, *14*(6):1006–1033.

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3).

Karlsson, F. (1983). Suomen kielen äänne-ja muotorakenne [the phonological and morphological structure of finnish]. *Werner Söderström, Juva*.

Karlsson, F. (1985). Paradigms and word forms. *Studia gramatyczne*, *7*:135–154.

Karlsson, F. (1986). Frequency considerations in morphology. *STUF-Language Typology and Universals*, *39*(1–4):19–28.

Karlsson, F. (2017). *Finnish: A comprehensive grammar*. Routledge.

Karlsson, F. and Koskenniemi, K. (1985). A process model of morphology and lexicon. *Folia Linguistica*, *29*:207–231.

Krijthe, J.H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.16.

Laine, M., Kujala, P., Niemi, J., and Uusipaikka, E. (1992). On the nature of naming difficulties in aphasia. *Cortex*, *28*(4):537–554.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2):211–240.

Marelli, M. and Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, *122*(3):485–515.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 – Workshop Track Proceedings*, pages 1–12.

Nikolaev, A., Ashaie, S., Hallikainen, M., Hänninen, T., Higby, E., Hyun, J., Lehtonen, M., and Soininen, H. (2019). Effects of morphological family on word recognition in normal aging, mild cognitive impairment, and alzheimer's disease. *Cortex*, *116*:91–103.

Schreuder, R. and Baayen, R.H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*:118–139.

Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., and Baayen, R.H. (2022). Semantic properties of english nominal pluralization: Insights from word embeddings. *arXiv*.

Shafaei-Bajestan, Elnaz, Uhrig, Peter and Baayen, R.H. (2023). Making sense of spoken plurals.

Shahmohammadi, H., Lensch, H., and Baayen, R.H. (2021). Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. *CoNLL 2021*. arXiv preprint arXiv:2104.07500.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Describing English language. Oxford University Press, Oxford.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1):267–288.

van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, *15*:3221–3245.

van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, *9*:2579–2605.

Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C.C. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, *8*(May):e19.

Address for correspondence

Alexandre Nikolaev
University of Eastern Finland
Yliopistonkatu 4
PL 111
80101 Joensuu
Finland
alexandre.nikolaev@uef.fi
https://orcid.org/0000-0001-8634-5947

Co-author information

Yu-Ying Chuang
Seminar fur Sprachwissenschaft/Quantitative Linguistics
Eberhard Karls University
yu-ying.chuang@uni-tuebingen.de

R. Harald Baayen
Seminar fur Sprachwissenschaft/Quantitative Linguistics
Eberhard Karls University
harald.baayen@uni-tuebingen.de

Publication history

Date received: 30 June 2022
Date accepted: 15 November 2022
Published online: 17 March 2023