

Historical and psycholinguistic perspectives on morphological productivity: A sketch of an integrative approach

Harald Baayen¹, Kristian Berg², and Mirrah Maziyah Mohamed¹

University of Tübingen (1) and University of Oldenburg (2)

May 20, 2025

1 Introduction

In this study, we approach morphological productivity from two perspectives: a cognitive-computational perspective, and a diachronic perspective zooming in on an actual speaker, Thomas Mann. For developing the first perspective, we make use of a cognitive computational model of the mental lexicon, the discriminative lexicon model (DLM). For developing the second perspective, we study how the intake and output of one prolific writer changes over time.

The DLM models implements mappings between numeric representations for words’ forms, and numeric representations for their meanings, using embeddings from distributional semantics. Model mappings can be evaluated not only on how precise their predictions are for the data on which they have been trained, but also how well they predict novel, unseen data. We will illustrate the promise of using model accuracy on held-out data as a novel fine-grained measure of productivity. To do so, we present case studies of nominal inflection in Finnish, derivational prefixation in Malay, and compounding in English.

Although the DLM is a model for a language user’s mental lexicon, a DLM model is typically trained using words as observed in corpora. Typically, these corpora sample language from many different speakers. Because individual speakers have their own areas of interest and expertise, they will know the vocabulary specific to these interests, but will be less familiar or unfamiliar with the vocabulary of other areas of specialization. As a consequence, corpora overestimate the vocabularies of individual speakers, and the ecological validity of computational models targeting an individual’s mental lexicon are inevitably negatively affected when trained on aggregated community data.

We therefore complement this cognitive-computational perspective with a socio-historical perspective, investigating the productivity of selected German derivational suffixes in the writings of Thomas Mann, comparing what he is known to have read (community input) with his own writings (personal output). We will sketch how a DLM model could be set up for Thomas Mann, and present some first findings.

A common thread across all our case studies is the use of embeddings to represent words’ meanings. We will show that for inflection and derivation, the DLM model links n-grams that are co-extensive with inflectional or derivational exponents with the most prototypical embeddings of the words that use these exponents. We will also show that Thomas Mann is more likely to produce a novel German derived word with a given suffix when the embeddings of the words with that suffix cluster more closely around the prototypical embedding.

In what follows, we first report our computational experiments with the DLM as a tool for gauging productivity. We then turn to our study of Thomas Mann and the productivity of German suffixes in his writings. We conclude with a discussion of our findings.

2 Productivity: approximations with the DLM

2.1 General considerations

The discriminative lexicon model (DLM) is a computational tool for, on the one hand, assessing how well mappings between form and meaning can be learned, and on the other hand, for tracing the consequences of learnability for lexical processing. The model requires the analyst to define mappings between form and meaning for comprehension. Both words’ forms and words’ meanings are ‘embedded’ in high-dimensional vector spaces. The numeric vectors for words’ forms are brought together as the row vectors of a matrix (table) \mathbf{C} , and the numeric vectors for words’ meanings are brought together as the row vectors of a matrix \mathbf{S} . For comprehension a mapping f takes the form vectors and transforms them into the corresponding meaning vectors, as precisely as possible:

$$f(\mathbf{C}) = \mathbf{S}.$$

For comprehension, a mapping g takes the semantic embeddings in \mathbf{S} and transforms these into their form embeddings \mathbf{C} :

$$g(\mathbf{S}) = \mathbf{C}.$$

As is customary in machine learning, model parameters (which define the mappings f and g) are estimated for a training dataset, and the quality of the model is evaluated on held-out test data. This provides a natural framework for evaluating morphological productivity, namely, by asking the question how well the model is able to understand and produce novel complex words in the held-out dataset, i.e., words that it has not encountered before during training.

For mappings between form and meaning to be productive, in the sense that novel, previously unencountered words, can be understood and produced, there must be systematicities between the form space and the semantic space. If the relation between form and meaning would be truly arbitrary, a model could memorize form and meaning pairings, but there is no way in which the model would be able to generalize to novel test data. Thus, the arbitrariness of the linguistic sign (De Saussure, 1966) is a worst-case scenario for any learning algorithm. Morphology, however, breaks this arbitrariness. What is of interest to us here is that the way in which morphology breaks this arbitrariness differs substantially between inflection, compounding, and derivation, and that the consequences for generalization are profound.

In the literature on systemic productivity, restrictions (on the form and meaning of base words, and the form and meaning of complex words) have played a foundational role (see, e.g., Booij, 1977). We maintain that such restrictions enable generalization. Without such restrictions, generalization is impossible. We will argue, and present modeling evidence, that compounding, lacking such restrictions, does not enable solid generalization, whereas inflection and derivation, thanks to being constrained by restrictions, do enable generalization.

In what follows, we present a series of computational experiments that clarify this perspective on morphological productivity. We first zoom in on inflection, using nominal inflection in Finnish as a non-trivial example of the issues that any theory of morphology seeking to predict inflectional productivity has to solve. We then discuss the very opposite kind of word formation, compounding, using datasets from English, but briefly venturing into compounding in Mandarin Chinese. Following this, we target derivational morphology, using prefixal morphology in Malay as our working example.

2.2 Inflection

Inflectional morphology is where regularity tends to be most clearly visible. Yet, inflectional systems often have their own pockets of irregularity. Past-tense inflection in English provides a well-known example. Regular past-tense inflection with the dental suffix has been described (and hotly debated) as fully regular and productive, and the irregular past tense forms as being unproductive and stored in memory. In this section, we consider a very different inflectional system: nominal inflection in Finnish. Finnish nouns are inflected for two numbers and fourteen cases. At the form side, some 49 inflectional classes are distinguished (see Nikolaev et al., 2025, for detailed discussion). At the semantic side, Nikolaev et al. (2023) have shown that the change in semantic space from singular to plural varies systematically by case. In other words, number and case do not have straightforward simple semantic main effects, instead, number and case enter into a semantic interaction.

Can the DLM master this complex inflectional system, without having recourse to stems, features for stem allomorphy, features for inflectional class, and rules of referral (most plural forms are based on the partitive singular)? Nikolaev et al. (2025) tested the DLM on 55,271 nominal forms associated with the 2000 most frequent Finnish nouns. Using fasttext embeddings to represent words’ meanings, 4-gram vectors to represent words’ forms, and frequency-informed learning, they observed that on the training data, the model correctly understood 75.96% of the types and 95.76% of the 40,694 tokens of the dataset on which the model was trained. For held-out data, the 14,577 words with a frequency less than or equal to five, 64.29 were correctly understood. For 98.29% of the held-out tokens, the predicted vector was among the top 10 closest neighbors of the targeted ‘gold standard’ embedding. These results indicate that the DLM comprehension model is productive, in that it can understand many inflected forms that it has not encountered during training.

However, this overall assessment of the DLM’s performance (and by implication, the productivity of the Finnish noun system) is not fair, as the 49 inflectional classes differ substantially in size and productivity. Some inflectional classes have many types, many hapax-legomena, and low median word frequency. Others comprise only small numbers of types, few if any hapax legomena, and many high-frequency words. Performance of the DLM on the less productive and unproductive inflectional classes should be worse than its performance on more productive inflectional classes. This is exactly what Nikolaev et al. (2025) observed, as illustrated in Figure 1. As the number of types in an inflectional class increases, the accuracy of the DLM on held-out data increases as well (left panel). Similarly, inflectional classes with more hapax legomena afford greater DLM accuracy, again on held-out data (center panel). Furthermore, as the median word frequency in an inflectional class increases, prediction accuracy for held-out data decreases.

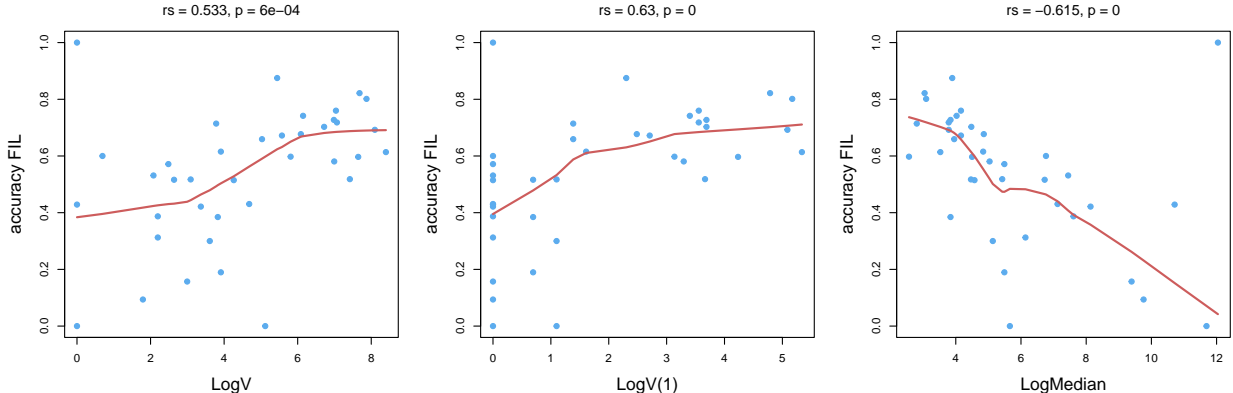


Figure 1: By-class accuracy of a frequency-informed comprehension model on held-out data as a function of three productivity measures. Dots represent inflectional classes.

We propose that the prediction accuracies of the DLM for the different inflectional classes provide insight

into the degrees of productivity of these classes. These usage based, learning-driven accuracies integrate the joint effects of type and token frequencies, as well as the effects of idiosyncracies at the level of word form and the level of word meaning.

Thus far, we have used the DLM as a black box. We have shown that the black box exhibits the theoretically expected behavior. But why does it work so well? In what follows, we will pry open the black box, and show that what the model does is remarkably well-interpretable, and is surprisingly similar to realizational theories of morphology.

Recall that the DLM sets up a matrix \mathbf{C} of form vectors. Each row of such a table specifies which n-grams are present in a word’s form. For the Finnish words *vuonna* (‘year’, a lexicalized form of the singular essive) and *kello* (‘clock’), the ‘form embeddings’ based on letter 4-grams could look like this:

	#vuo	vuon	uonn	onna	nna#	#kel	kell	ello	llo#	...
vuonna	1	1	1	1	1	0	0	0	0	...
kello	0	0	0	0	0	1	1	1	1	...

Only a small part of these embeddings are shown. There are 13,364 different 4-grams in the dataset of Finnish analyzed by Nikolaev et al. (2025), for *vuonna*, $13,364 - 5 = 13,359$ 4-grams, the value in the row vector is zero. The DLM assumes that there is a function f that takes these form embeddings and transforms them into semantic embeddings. In other words, f applied to the row vectors of the form embeddings \mathbf{C} returns the corresponding set of semantic embeddings, which we bring together as the row vectors of a table (matrix) \mathbf{S} :

$$f(\mathbf{C}) = \mathbf{S}.$$

In what follows, we assume that this function f is very simple and implements a linear transformation:

$$\mathbf{CF} = \mathbf{S}.$$

How linear transformations work is explained in detail in Heitmeier et al. (2025). Important for the present discussion is that the row vectors of the table (matrix) \mathbf{F} specify 4-gram specific embeddings. The first five rows of the following display show the first 6 cells of these 300-element long 4-gram embeddings.

#vuo	-0.007273	-0.014577	-0.026716	0.014613	-0.004810	-0.019703	...
vuon	-0.032234	-0.023539	0.003362	0.001231	0.008429	0.022490	...
uonn	0.030358	-0.015991	-0.010521	0.022888	-0.001061	0.032784	...
onna	-0.000150	0.007794	0.011207	0.003758	-0.009883	-0.002486	...
nna#	0.007692	0.010210	-0.007754	-0.006202	-0.037584	-0.018848	...
SUM	-0.001606	-0.036103	-0.030423	0.036290	-0.044910	0.014237	...

What the linear mapping \mathbf{F} does is look up in table \mathbf{C} which 4-grams occur in say *vuonna*, retrieve the corresponding 4-gram specific embeddings, and sum these column wise. The first 6 elements of the resulting semantic embedding predicted for *vuonna* are shown in the last row of the above display. The mapping \mathbf{F} faces a hard task: it has to stay faithful to the meaning of the lexeme, and not confuse the meaning of *vuosi* (‘year’) with the meaning of *kello* (‘clock’). It has to realize essive singular for *vuonna*, and nominative singular for *kello*. And it has to make peace with the extensive stem allomorphy that characterizes Finnish nouns (for *vuosi*, *vuo-*, *vuon-*, *vuot-*). The 4-grams that are part of the stem have to take care of implementing the meaning “year”. The initial 4-gram *#vuo* is an excellent cue for “year”, all forms of *vuosi* start with *#vuo*, and there are only two other lexemes in the dataset that share this initial trigram: *vuokra* ‘rent’ and *vuokralainen* ‘tenant’. The next 4-gram, *vuon*, is unique to *vuosi*, but occurs in only four forms. The remaining 4-grams incorporate part of the inflectional endings, which are shared with many other words. What semantic embeddings does the linear mapping estimate for these trigrams? To address this question, we inspected the final 4-grams, which consist of three letters and the word boundary marker *#*. We limited our exploration to the 34,262 words in our dataset that have no possessive clitics nor discourse clitics.

Given that in the semantic system of Finnish, case and number interact (Nikolaev et al., 2023), we collected all word-final 4-grams (e.g, *nna#* for *vuonna*) and sorted them by the case+number combinations.

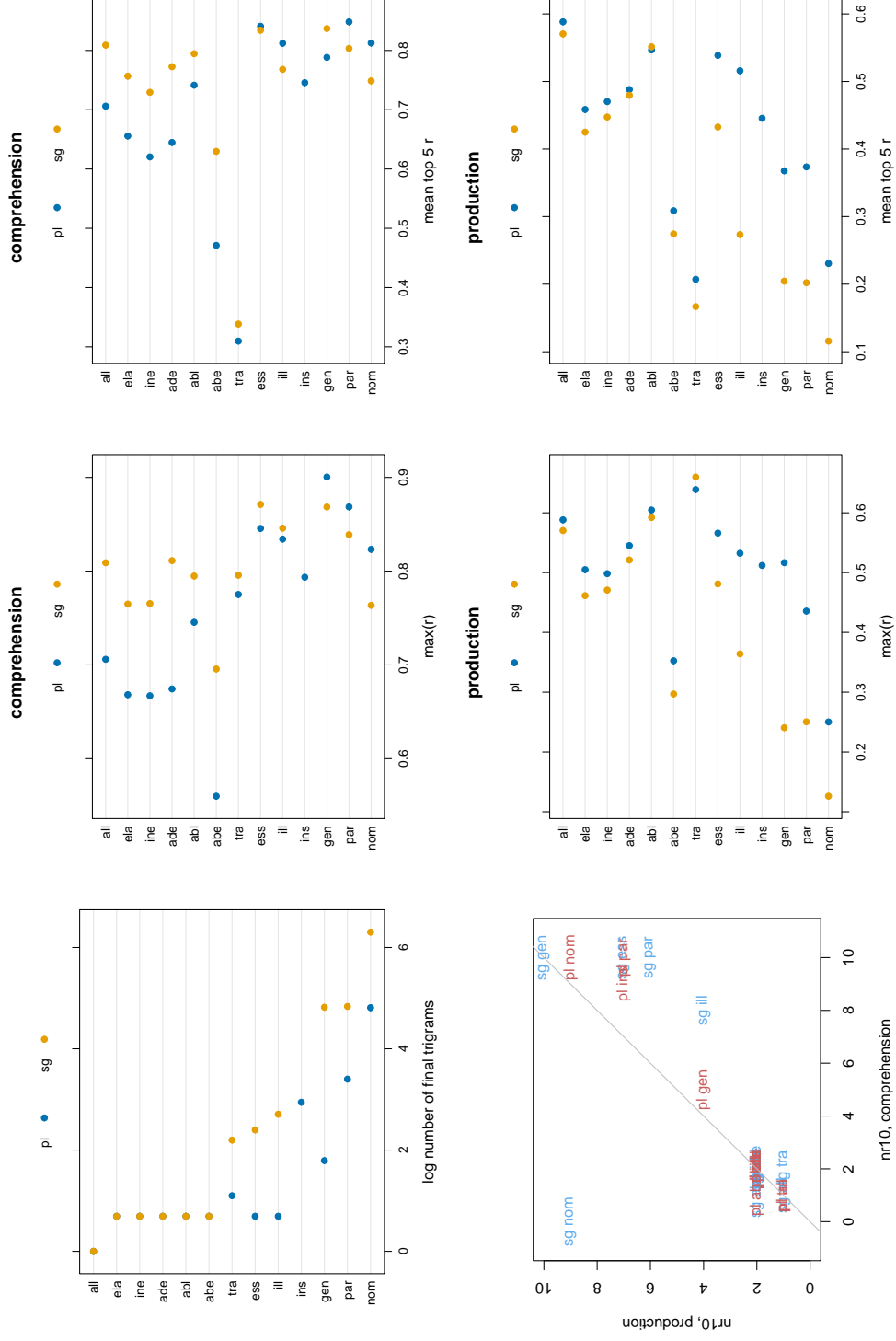


Figure 2: Row vectors of F and column vectors of G in relation to the centroids of case-number combinations in Finnish. Number of final trigrams: word-final three-letter sequences that realize a given case-number combination; max(r): the maximum correlation with the centroid; mean top 5 r: mean of the top-five ranked correlations; nr10: the number of final trigrams among the top 10 highest correlations.

The number of word-final 4-grams for the case+number combinations ranges from 1 (for allative singular and allative plural) to 547 (for the nominative singular). The upper left panel of Figure 2 presents these counts (log-transformed) for the different cases, broken down by number. The nominative clearly pays a high price for being unmarked. The other two grammatical cases, partitive and genitive, also have large numbers of final endings, reflecting that these cases often have somewhat unpredictable stems.

For each of the word-final 4-grams, there is a row in the \mathbf{F} mapping that represents the semantic contribution of that n-gram to its carrier word. Our hypothesis is that this semantic contribution must be similar to the centroid of the case-number combination that the 4-gram is tied to. Figure 3 illustrates the centroids for essive singulars (in red) and illative plurals (in blue), using the t-SNE unsupervised clustering algorithm (Maaten and Hinton, 2008) to re-represent the embeddings of the inflected nouns in a 2-dimensional space. The larger clusters represent case, within clusters, plurals and singulars form sub-clusters. The centroids are in the center of the cluster, and represent the most prototypical meaning of a case-number combination. Our intuition is that the best the linear mapping can do is use embeddings for final 4-grams that are similar to the centroid of the case-number cluster that these 4-grams have to realize.

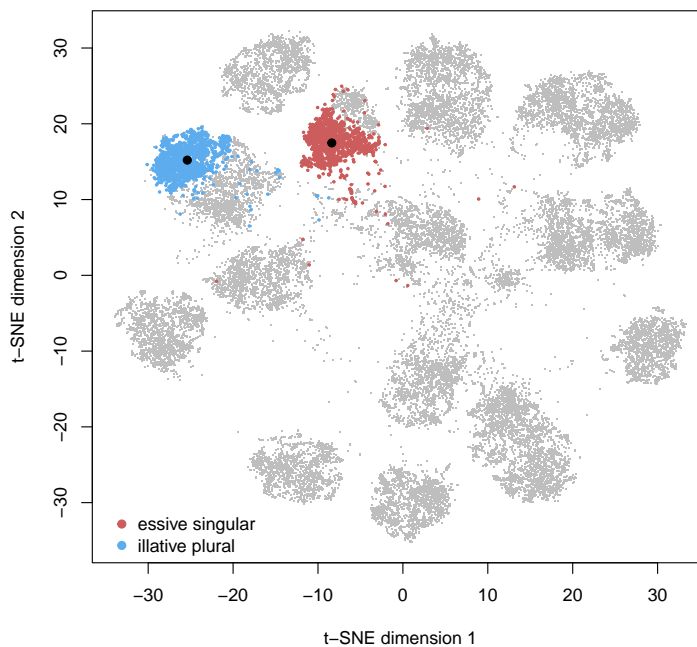


Figure 3: Position of 34,262 Finnish nominal forms in a 2-dimensional t-SNE plane. Clustering is by case. The cluster for essive singular is presented in red, and the cluster for illative plural is given in blue. The black points represent the centroids of these clusters.

Table 1: Correlations of case-number centroid and final 4-gram embedding (row vector of \mathbf{F}) and rank in the ordered sequence of all 4-grams with this centroid.

illative plural		
4-gram	r	rank
iin#	0.83	1
hin#	0.79	2
essive singular		
4-gram	r	rank
una#	0.87	1
ana#	0.86	2
ena#	0.85	3
enä#	0.81	4
ona#	0.77	5
ina#	0.76	6
inä#	0.73	7
önä#	0.72	8
änä#	0.71	9
ynä#	0.62	10
nna#	0.18	1966
nominative singular		
4-gram	r	rank
kka#	0.76	11
eri#	0.76	12
lma#	0.75	19
eli#	0.75	20
uri#	0.72	38
ari#	0.72	47
⋮	⋮	⋮

The reason is straightforward: the centroid averages out the differences in the consequences of the very different affordances of the referents of individual nouns and their consequences for the corresponding embeddings. Because a given word-final 4-gram has to “serve” many different nouns simultaneously, an embedding that is similar to the centroid, the average embedding, is the best solution.

To put this line of reasoning to the test, we calculated the correlation matrix of the centroids with the row vectors of the mapping \mathbf{F} . The following display shows a small part of this $26 \times 13,364$ matrix.

	#vuo	vuon	uonn	onna	nna#	#kel	...
sg ess	0.45	0.12	0.19	-0.20	0.18	0.33	...
sg nom	0.55	-0.12	0.11	-0.00	0.30	0.58	...
sg ill	0.26	-0.00	0.08	0.01	0.15	0.39	...
sg gen	0.42	0.00	0.14	-0.08	0.15	0.42	...
sg tra	0.35	0.06	0.20	-0.12	0.12	0.35	...
sg par	0.46	-0.04	0.07	-0.06	0.13	0.46	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

If our intuition is correct, we expect that the 4-grams with the highest correlations with a given centroid will be the word-final 4-grams that occur in words that realize the relevant combination of number and case. Table 1 presents results for three case-number combinations: illative plural, essive singular, and nominative singular. For the illative plural, there are only two word-final 4-grams, and their embeddings are well-correlated with the centroid of the illative plural cluster. The correlations, 0.83 and 0.79, are the most

correlated of all 13,364 4-grams (ranks 1 and 2). The essive singular is realized by a wider range of 4-grams. For 10 4-grams, the correlations are high and occupy ranks 1-10. It is only the 4-gram *нна#* (as in *vuonna*) that is not well correlated with its centroid: many irrelevant 4-grams (1955) have higher correlations. For the unmarked nominative singular, which is associated with 547 final 4-grams, correlations are lower, and ranks are higher.

The upper central panel of Figure 2 presents the maximal correlation of a word-final 4-gram with its case-number centroid. Except for the nominative singular and the abessive plural, this maximal correlation is always associated with an appropriate word-final 4-gram. In the same figure, the upper right panel presents the mean of the correlations of the 5 most highly ranked correlations. The two panels provide solid support for the idea that the semantics contributed by the word-final 4-grams, the fuzzy ‘exponents’ of the DLM model, realize semantics that are highly similar to the centroid of their corresponding case-number embeddings.

For the mapping required for production, \mathbf{G} , which starts out with the semantic vectors and uses these to predict words’ 4-grams, a similar pattern is observed. It is now the the column vectors of \mathbf{G} , which are used to support 4-grams, that are strongly correlated with the centroids of the case-number clusters. As can be seen by comparing the center and right lower panels in Figure 2 with their upper panel counterparts, the correlations are somewhat lower for production as compared to comprehension. There are four plural cases for which the best correlation is not ranked first. This reduced correlations with the centroids are likely due to the mapping from meaning to form (300 dimensional meaning vectors to 13,364 dimensional form vectors) cannot be very precise as it is not possible to map from a lower-dimensional space into a higher-dimensional space with precision. The lower left panel of Figure 2 presents the count of word-final 4-grams that have ranks of at most 10. The nominative singular is the outlier in this plot: it has good ranks for production, but not for comprehension.

In summary, we have argued that the accuracy of the DLM model on held-out data may be useful as a measure of productivity. This measure lines up well with other already well-established measures of productivity, and elegantly navigates the complexities of the inflectional classes of the Finnish noun system. Furthermore, we have shown that the mappings of the DLM associate word-final 4-grams with the prototypical embeddings of case-number combinations. These centroids form the basis of generalization. Because the centroids are means, they remain fairly stable even when word forms are withheld from the training data, and hence enable generalization to novel forms on which the model was not trained. Although the DLM does not attempt to isolate stems and exponents, and works with multiple word-final 4-grams for a given case-number combination, these 4-grams come very close to ‘realizing’ the prototypical semantics of case and number.

2.3 Compounding

How productive is compounding? A perusal of the CELEX database (Baayen et al., 1995) shows that of all entries in the lemma list for English that are analysed as having two constituents, 7.1% are prefixed words, 42.7 are suffixed words, and 36.9 are compounds. Clearly, compounding is a productive word formation pattern. There are languages such as Vietnamese and Chinese that heavily rely on compounding, with no inflection and hardly any derivation. And yet, in structuralist analyses, compounds are not morphological categories, i.e., sets of words that share aspects of form and aspects of meaning (see, e.g., Schultink, 1961). If compounds have no clear systematic relations between form and meaning to lay the basis for generalization, how then can they be productive?

To address this question, we carried out a computational experiment with 11,170 English words, split into training data (10959 words, 7301 compounds (written as one word), 3658 constituents) and test data (811 compounds with constituents that occur in the training data). Comprehension accuracies for endstate learning (EOL) and frequency informed learning (FIL) are listed in Table 2 for training data and test data, for accuracy (@1) and loosely evaluated accuracy (being among the top 10, @10). For the training data, accuracies are high for the training data (and for FIL, as expected, accuracy is only high using token-wise evaluation). For the held-out test data, all compounds with known constituents, accuracy @1 plummets to 12.6% for EOL, and to zero for FIL. The higher score for EOL (but not FIL) for lenient evaluation @10

for the novel compounds, 69.1%, suggest that the model may get the gist of the semantics of the novel compounds, but lacks precision.

Table 2: Comprehension accuracies for training and test data. For the training data, token-wise accuracy @1 for FIL is 0.858.

	training		testing	
	learning	@1 @10	@1 @10	
EOL	0.866	0.989	0.1258	0.6905
FIL	0.079	0.691	0	0.1134

Table 3: The trigrams of *airfield* and the correlation of their row vectors in \mathbf{F} with the embeddings of selected words. Trigrams spanning the constituent boundary are highlighted in red.

	about	abouts	absorption	absorption-line	abundance	access	air	airfield	craft	field
air	-0.099	-0.067	-0.070	-0.040	-0.016	-0.034	0.063	0.029	-0.032	-0.147
ld#	0.200	0.232	0.152	0.083	0.190	0.171	0.318	0.128	0.377	0.246
#ai	0.130	0.126	0.217	0.200	0.091	0.139	0.331	0.314	0.204	0.232
irf	-0.082	0.020	-0.142	-0.085	-0.025	0.004	-0.145	0.123	-0.120	-0.084
rfi	0.059	0.151	-0.107	-0.077	-0.112	-0.063	0.030	0.229	-0.021	-0.067
fie	-0.137	0.005	-0.141	0.047	0.035	-0.095	-0.054	0.103	-0.084	0.131
iel	0.096	0.001	0.114	0.049	-0.093	0.059	-0.069	0.012	-0.057	0.046
eld	-0.029	-0.066	-0.016	-0.111	0.059	0.018	0.087	-0.042	0.090	-0.119

In order to understand how the model learns to understand the compounds that it has been trained on, it is helpful to consider how the trigrams of a compound contribute to its meaning, using their trigram-specific embeddings in the \mathbf{F} matrix. In what follows, we set aside the effects of usage, and inspect the \mathbf{F} mapping obtained with endstate learning.

Table 3 shows a small part of the correlation matrix for the row vectors of \mathbf{F} and the embeddings in \mathbf{S} . The trigrams shown are those of *airfield*. The trigrams that straddle the constituent boundary between *air* and *field*, *irf* and *rfi*, are highlighted in red. In what follows, for ease of exposition, when we mention the ‘correlation of *irf* and *airfield*’, abbreviated as $r(\textit{irf}, \textit{airfield})$, this is to be understood as the correlation of the row vector of *irf* in \mathbf{F} and the row vector of *airfield* in \mathbf{S} . Table 3 shows that $r(\textit{irf}, \textit{airfield})$ is greater than both $r(\textit{irf}, \textit{air})$ and $r(\textit{irf}, \textit{field})$. The same holds for the corresponding correlations of *rfi*. Furthermore, the trigrams that *airfield* shares with its left constituent (*#ai*, *air*) are more strongly correlated with *air* than with *airfield*. Likewise, three of the trigrams that *airfield* shares with its right constituent (*fie*, *iel*, *ld#*) are more strongly correlated with *field* than with *airfield*. The exception is *eld*, which is negatively correlated with *airfield*, and even more negatively correlated with *field*. To sum up, this example shows that the boundary trigrams support the meaning of the compound much better than the meanings of the constituents. Non-boundary trigrams tend to support the meanings of the constituents much better than they support the meaning of the compound itself. This is unsurprising, as for instance *air* has to be shared with many other compounds, in our dataset *air-pump*, *air-raid*, *airbrake*, *aircraft*, *aircrew*, *airfield*, *airflow*, *airforce*, *airframe*, *airgun*, *airlift*, *airline*, *airlock*, *airmail*, *airman*, *airplane*, *airport*, *airscrew*, *airship*, *airsick*, *airspeed*, *airstrip*, and *airway*.

Figure 4 illustrates that this pattern of results is not unique to *airfield*. For the 2288 two-constituent compounds in our dataset for which a parse is available in the CELEX database (Baayen et al., 1995), this boxplot visualizes the distribution of proportions in a word of trigrams for which the correlation with the compound is greater than the correlation with the left or right constituent. Trigrams are grouped into boundary trigrams, trigrams preceding the boundary (left tri), and trigrams following the boundary (right tri). For some 83 to 84% of the compounds, both trigrams have higher correlations with the compound embedding than with the left or right constituent embeddings. The proportions of compounds for which the trigrams are more correlated with the compound embedding than with the left or right constituent embedding

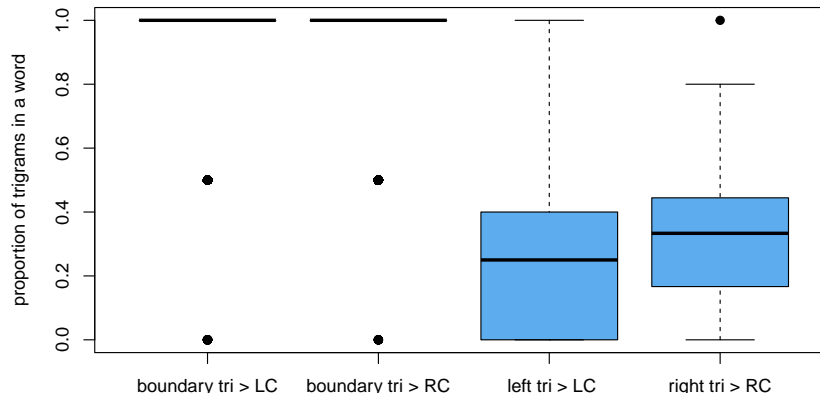


Figure 4: Proportion of trigrams in a word for which the correlation with the compound embedding exceeds the correlation with the embedding of a left (LC) or right (RC) constituent, broken down by boundary trigrams (boundary tri), trigrams preceding the boundary (left tri) and trigrams following the boundary (right tri).

are much lower. In other words, it is the boundary trigrams that have the highest relative functional load of guiding interpretation to the meaning of the compound itself.

Now that we understand how the model manages to learn to understand compounds, it is not difficult to see why it cannot generalize very well. Boundary trigrams have to be shared with other compounds. *rfi* occurs not only in *airfield*, but also in *butterfingers*, *quarterfinal*, *silverfish*, and *starfish*. Likewise, *irf* is also a boundary trigram for *airflow*, *airforce* and *airframe*. As *butterfingers*, *silverfish*, *airflow*, and *airforce* are semantically highly diverse, *irf* is not very useful for semantic generalization.

Limited generalization for compound semantics is possible, as shown by the CAOSS model of [Marelli et al. \(2017\)](#) applied to 8147 English compounds. This model takes the embeddings of the constituents as input, and uses a linear transformation to approximate the embedding of the compound (see the appendix for technical details). Accuracy of this model for all compounds jointly is at 54.3%, but in a training-testing set-up with random selection of 815 held-out compounds, accuracy for the held-out data is down to 23.8%. Interestingly, just adding up the embeddings of the constituents for the held-out data creates predicted compound vectors that already have an accuracy of 18.0%. The fine-tuning that the linear transformation of CAOSS offers over and above vector addition is an increase in accuracy of only 5.8%.

The low generalization accuracy of CAOSS fits well with the findings of [Schäfer and Bell \(2020\)](#), who observed that native speakers of English have great difficulties guessing the meaning of novel compounds such as *acid cap*, even when some of the original context of use is provided. In the light of these considerations, it makes sense that the comprehension mappings of the DLM cannot predict the meanings of held-out compounds with precision.

When modeling the productivity of compounds in production, we again find that the training data can be learned well, but that generalization is highly inaccurate. This is illustrated by a mapping from 300-dimensional embeddings to the 4799-dimensional trigram vectors of the dataset of 11,170 English words. As mapping with precision from a low-dimensional space to a high dimensional space is usually impossible with a linear transformation, we made use of a deep network.¹ Using the same training-test split we used above, we found that for training data, this model predicted the form vector correctly for 92% of the

¹This network had 1000 RELU units, using the binary cross-entropy loss. For details of this kind of network, see [Heitmeier et al. \(2025\)](#).

words. However, for the held-out compounds, accuracy plummeted to 24.7%. The idiosyncratic meanings of the held-out compounds stand in the way of predicting their pronunciations.² This brings us back to our original question: compounds are evidently productive type-wise, but how can this be given that there is no overarching systematicity enabling generalization? We think that compounding is a highly useful for onomasiological tinkering, for creating names by reusing old words in creative ways. This ‘bricolage’ (Lévi-Strauss, 1962) tends to be locally motivated at time of creation (cf. the etymologies of opaque compounds such as *dumbbell*, *moonshine* and *hogwash*), and can lead to small islands of reliability in semantic space. However, the productivity of such islands of reliability may be severely constrained, as shown by Shen and Baayen (2022).

Shen and Baayen (2022) studied compounding in Mandarin Chinese, focusing on how often a given word appears as a constituent of a compound in a given position. Following their terminology, we refer to a position-specific constituent word as a ‘pivot’. Examples of compounds with the pivot *da4*, ‘big’ are *da4jia1* ‘everyone’; *da4xue2* ‘university’ and *da4liang4* ‘generous’. This study observed that as the number of compound types with a given pivot V increases, the Good-Turing probability of unseen types \mathcal{P} (the category-conditioned productivity measure proposed in Baayen, 1993) decreases. Such a negative correlation of V and \mathcal{P} is absent for English derivational affixes. Shen and Baayen (2022) also observed that as \mathcal{P} increases, the embedding of the pivot is more similar to the embedding of its compound, as well as to the embedding of the non-pivot in its compound. Considered jointly, these results suggest that semantic transparency is possible for pivots with small numbers of compounds, but that if a pivot is re-used in more and more compounds, semantic transparency suffers.

Interestingly, the productivity of a pivot decreases if the region in semantic space defined by the compounds of the pivot overlaps too much with the region in semantic space defined by the compounds of the non-pivot. To define a semantic island of reliability for the compounds with a given pivot, they took the embeddings of these compounds, and calculated the centroid, the most prototypical meaning of the pivot in a compound. Next, they calculated, for each compound with the pivot, its correlation with the centroid. For the resulting distribution of correlations, they calculated the 95% confidence interval. All compounds having a correlation within this confidence interval around the centroid are considered to be in the pivot’s island of reliability. Finally, Shen and Baayen (2022) calculated for all compounds with the non-pivot as constituent how many had a correlation with the pivot’s centroid 95% confidence interval. They designated such non-pivot compounds as ‘intruders’. The count of intruders was shown to be negatively correlated with \mathcal{P} . Furthermore, the average of the correlations of the embedding of the pivot with the embeddings of its compounds, a measure of the transparency of the pivot, is negatively correlated with the number of intruders.

In our study of Finnish nominal inflection, the centroids of case-number combinations emerged as a good approximation of the contribution that in the DLM word final trigrams (which capture a substantial part of the exponents for case and number) make to the predicted embedding of the inflected noun. For compounding, our conjecture is that the centroids of pivot constituents are also important, and the more semantically related words are within the area around the centroid, the more productive the pivot can be. However, a pivot that is originally semantically motivated can be put to other uses that move away from the centroid. This onomasiological tinkering rides piggy back on the popularity of a pivot, but because human-perceived order in the natural world is highly constrained by the diversity of the natural world and the limitations of human perception (Kant et al., 1999; Husserl, 1913; Merleau-Ponty et al., 2013; Hoffman, 2019), for productivity, the more tinkering goes on at time t , the less tinkering is attractive at time $t + 1$. An additional complication is that compounds with the non-pivot can enter the pivot’s island of reliability and mess up its semantic transparency.

In summary, morphological tinkering is great. Tinkering poses no problem for learning of training data. L1 learners likely don’t face much of a problem, in part because subliminal statistical learning is easiest when young, and in part because vocabulary development goes hand in hand with general broadening of

²As the present dataset does not include compounds written with internal spaces, these potentially more semantically transparent compounds (Kuperman and Bertram, 2013) (if these are not to be understood as phrases) are not taken into consideration in the present study.

cognitive skills. However, we should not expect compound tinkering to afford much generalization for test data. L1 learners can make use of contextual inferencing skills, but specifically L2 learners face a formidable memorization task, as they are more likely to attempt compositional strategies which run aground as trying to predict what novel compounds mean is largely self-defeating. At the micro-level of individual pivots some generalization is possible, perhaps, but not at the macro-level. [Schultink \(1962\)](#) argued that compounds do not constitute a morphological category. We agree: compounds are *sui generis*, and very useful and productive as onomasiological device, albeit without relying (much) on form-meaning generalization. As a consequence, systemic productivity is low: our estimate for English production is 24.7%, and our estimates for comprehension are 12.6% for endstate learning, and 0 for frequency informed learning.

In the next section, having completed our examination of the extremes of inflection and compounding, we turn to derivation.

2.4 Derivation

A language that is particularly rich in morphological derivation is Malay, an Austronesian language spoken by more than 200 million peoples in various countries of Southeast Asia. Malay has minimal inflection. Affixation, by means of derivational prefixes, is at the heart of what makes Malay productive. In Malay, allomorphy is pervasive such as in prefix families of *beR-* (i.e., be-, bel-, and ber-), *meN-* (i.e., me-, men-, mem-, meny-, meng-, and menge-), *peN-* (i.e., pe-, pen-, pem- peny-, peng-, and penge-), *peR-* (i.e., pel-, and per-), and *teR-* (i.e., te-, and teR-), some of which typically change the meaning or class of a word. In some cases, the initial letter of the stem is omitted to facilitate pronunciation (e.g., *fikir*/think; *pemikir*/a thinker). Although this allomorphy complicates systematicities between form and meaning, [Maziyah Mohamed and Baayen \(2025\)](#) have shown that, nevertheless, derived words cluster by prefix in semantic space. In that study, a *t*-SNE analysis ([Maaten and Hinton, 2008](#)) was conducted on a large set of complex Malay words using high-dimensional word embeddings. Following a key principle of Distributional Semantics, the authors observed a strong correspondence between form and meaning such that words that share a prefix appear closer in semantic space than those that contain a different prefix.

From a computational perspective, generalization is as an index of productivity. A question of primary interest concerns how well generalization occurs in a language in which morphological derivation is extensive. To address this question, the approach taken is two-fold. First, as demonstrated in our working examples for inflection and compounding, we trained the DLM on a set of Malay words, and evaluated its comprehension accuracy on words in the training set and test accuracy on the held-out data. The accuracy with which the DLM predicts the meaning of the unseen forms in the held-out data is contingent on the statistical co-occurrences between form and meaning of the words in the training set. From the series of analyses above, regularities in form and meaning appear rather transparent for inflection and very opaque for compounds. If our intuition holds, the DLM trained on Malay derived words should generalize to held-out data much better than English compounds, but pale in comparison to Finnish inflected forms. Because the DLM is also a cognitive model, a secondary goal is to assess its performance against behavioural data. Implications for the ease with which we read are discussed as a function of form-meaning systematicities.

Our dataset comprises of 8,843 words from the Malay Lexicon Project ([Yap et al., 2010](#)) for which FastText embeddings ([Bojanowski et al., 2017](#)) were extracted. Of these words, 7,959 words make up the training set and the remaining 884 words were reserved as held-out data for testing, such that all cues and derivational features present in the test set appeared in the training set (for details on careful splitting, see [Heitmeier et al. \(2024\)](#)). Words’ forms were represented using 4-gram vectors and words’ meanings were represented using FastText embeddings. For both end-state learning and frequency-informed learning, see Table 4 for the overall comprehension accuracies for the training and test data. Accuracies were evaluated at the top one and top 10 percent, that is, whether the predicted vector was the closest or among the top 10 closest neighbors of the targeted gold-standard embedding.

Indeed, the learning and test performance of the DLM trained on Malay derived words was somewhere in between the accuracies of the models trained on Finnish inflected words and English compounds. Most obvious is the large discrepancy in test accuracy for Malay derived words compared to English compounds, in that generalization to novel forms for derived words far exceeded that of compounds. In contrast, the

DLM outperformed in both learning and test performance when the model was trained on Finnish inflected words than on Malay derived words. It is clear that the extent of form-meaning regularities of the data on which the model was trained on closely corresponds to the rate at which successful generalization occurs. In sum, these comparisons further iterate the need for a strong correspondence between form and meaning for generalizations to be successful.

Table 4: Comprehension accuracies for training and test data (Types).

	training		testing	
learning	@1	@10	@1	@10
EOL	0.870	0.907	0.289	0.594
FIL	0.341	0.427	0.146	0.370

Note. For the training data, token-wise accuracy @1 for FIL is 0.90.

To help us better understand how the model learns derived words, as illustrated using compounds, we extracted the comprehension F mapping, obtained with end-state learning. Here, we correlated a small subset of the row vectors of F with the embeddings of the prefix centroids, that is, the mean embeddings of words containing a particular prefix. [Maziyah Mohamed and Baayen \(2025\)](#) provided empirical evidence for Malay that the correlation between the embedding of a derived word and the centroid was a reliable measure of semantic transparency (i.e., form-meaning systematicity) and the best predictor of lexical decision latencies, among several simple measures of transparency. A stronger correlation with the centroid indicates a more consistent mapping between form and meaning. Because the F mapping is a very large matrix, the row vectors we first analyzed were constrained to word-initial 4-grams that share a spelling with at least one of the prefixes (e.g., *#ber-*; prefix *beR-*). We tested whether the semantic contribution of the row vectors of prefix-like initial n-grams in the F matrix were similar to the centroids of the corresponding prefixes. Figure 5 depicts that that is the case. Not shown here, using 3-grams instead of 4-grams, the same analyses on the F matrix obtained from the DLM trained on the same set of words yield similar results. Then, we further examined the top five 4-grams in the F mapping for which embeddings are most correlated with each centroid. For most prefixes, the prefixal 4-grams contribute most strongly to the meaning of each centroid ($r > .80$; see Table 5). Prefix-like 4-grams are only moderately correlated with the centroids of *peri-* and *pra-*. *Peri-* occurs in very few words in our dataset and appears somewhat more opaque in its mapping between form and meaning (e.g., *peribahasa*/figure of speech, *perihal*/the state of something; *perilaku*/one’s behaviour). On the other hand, *pra-* appears rather transparent in its form-meaning correspondence, and is semantically similar to the English *pre-*, such as in *prauniversiti*/preuniversity. *Pra-*, however, is loaned from Sanskrit and also does not occur frequently in Malay. It is also noteworthy to point out that the row vectors of stem and word-final 4-grams are moderately correlated with each of the centroids as well. In particular, 4-grams that include the first segment of the stem are correlated with prefixes that have fewer than four letters (e.g., *pras*). Several word-final 4-grams that are correlated with the prefix centroids likely include the suffix *-an*, such as in *aan#*, *san#*, and *uan#*. *-an* is highly productive and occurs in more than 60 percent of words containing a suffix in our dataset. Such inspections on the intercorrelations among word-initial, stem, and word-final 4-grams with the centroids provide clarity as to why generalization occurs more seamlessly for derived words relative to compounds, but less extensively compared to inflected forms.

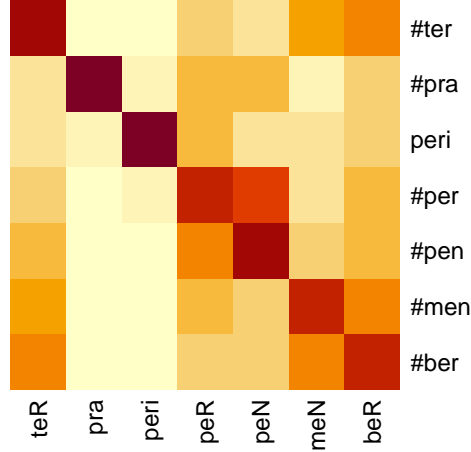


Figure 5: Correlation heatmap of prefix centroid embeddings and a subset of the row vectors of the F matrix that maps the form vectors (4-gram) to its meaning (FastText embedding) in the DLM. The set of 4-grams presented correspond to at least one of the prefixes. Darker hues, compared to lighter hues, indicate a stronger correlation between the embeddings of the centroid and row vectors of the F matrix. The strongest correlations are present for the 4-grams that overlap most with the prefix, indicating that it is the prefixal 4-grams that contribute most to realizing the meaning of the centroid.

Table 5: Top five strongest correlations between the row vectors in F and the centroid embeddings of selected prefixes. Coefficients highlighted in red represent the strongest correlation between the row vectors of the 4-gram and the embeddings of the centroid. Each prefix, except *peri*-, are most correlated with at least one of their allomorphs, providing evidence of differences in degrees of productivity for allomorphs.

	beR	meN	peN	peR	peri	pra	teR
#ber	.847	.717	-	-	-	-	.701
itu#	.739	.720	.677	.682	-	-	.752
aan#	.686	-	.755	.802	.572	.588	.669
#per	.672	-	.768	.809	.580	.549	-
bat#	.668	-	-	-	-	-	-
#men	-	.839	-	-	-	-	-
#mem	-	.753	-	-	-	-	-
#mel	-	.740	-	-	-	-	-
#pen	-	-	.803	-	-	-	-
#pem	-	-	.670	-	-	-	-
san#	-	-	-	.657	-	-	-
uan#	-	-	-	.653	-	-	-
#kes	-	-	-	-	.498	-	-
#kei	-	-	-	-	.497	-	-
#kea	-	-	-	-	.478	-	-
#pra	-	-	-	-	-	.684	-
pras	-	-	-	-	-	.574	-
gka#	-	-	-	-	-	.516	-
#ter	-	-	-	-	-	-	.822
lah#	-	-	-	-	-	-	.654

From a psycholinguistic perspective, a related question concerns how regularities in form and meaning (and, by extension, productivity) influence reading comprehension. To address this matter, we used two measures extracted from the DLM to predict response latencies. These measures are *Target Correlation* and *r target*. In principle, *Target Correlation* and *r target* are conceptually similar. Both measures refer to the correlation between the semantic vectors of the predicted word and the target, except that the estimates

from *Target Correlation* corresponds to words in the training set, and estimates from *r target* corresponds to words in the held-out data. In the subsequent text, we refer to them as **Target Correlation_{train}** and **Target Correlation_{test}** respectively. Lexical decision latencies from a series of previous studies of the Malay Lexicon Project (Maziyah Mohamed et al., 2023; Maziyah Mohamed and Jared, 2023, 2025), were extracted for 1,624 words in the training set and for 200 words in the held-out data. In what follows, we report our results on two sets of GAMs with careful consideration of concavity statistics.

First, we fitted a GAM to the training data. Since **Target Correlation_{train}** and **Target Correlation_{test}** were extracted from the DLM trained with end-state learning, we take into account word frequency and word length. Word frequency and word length are well-established predictors of word recognition. Word frequency was logarithmically transformed and RTs were inverse transformed. Of interest, **Target Correlation_{train}** and the correlation between the derived word and the centroid were entered as predictors of RT, with prefix as a random effect. All predictors significantly predicted RT (see Table 6). The top left panel of Figure 6 shows a facilitative effect of word frequency on RT, that is, faster responses were elicited for higher frequency words than for lower frequency words. The observed effect of frequency on RT is consistent with prior literature. In the bottom left panel of Figure 6, we observe an inhibitory effect of length, and a somewhat surprising facilitative effect for very long words. Note that very few words contained 13 or more letters (about 4 percent of the data). The effect of the correlation between the derived word and the centroid on RT mirrored recent findings in Malay (Maziyah Mohamed and Baayen, 2025). In that study, the effect of the correlation between derived words and their centroid embeddings on RT appeared non-linear and U-shaped. More specifically, faster responses were first observed the stronger the correlation between the embeddings of derived words and those of their centroids, followed by slower responses for very strong correlations ($r > .7$). Similarly, here, in the top right panel of Figure 6, responses were faster the closer a word is to its centroid, with the fastest responses predicted for words that share a moderate correlation with their centroid. However, responses were much slower for words that very closely resembled their centroid.

A separate GAM was fitted only to the held-out data. As in the GAM analysis above, the same predictors were entered in the model, except that **Target Correlation_{test}** was entered as a predictor in place of **Target Correlation_{train}**. Again, all predictors significantly predicted RT (see Table 7). Each panel of Figure 7 shows that the effect of each predictor in the analysis of held-out data exhibited similar trends on RT as illustrated above with the training data.

Table 6: GAMM fitted to the training data. Word frequency was log-transformed. TargetCorTrain = **Target Correlation_{train}**. CorDev-Centroid = Correlation between derived words and their centroid. The model syntax is $\text{inverse RT} \sim \text{s}(\log \text{ frequency}) + \text{te}(\text{TargetCorTrain} * \text{Corr.Dev-Centroid} + \text{s}(\text{word length}) + \text{s}(\text{prefix}, \text{bs} = 're'))$. Inverse RT = $-1000/\text{RT}$; a negative sign is used to ensure that the transformed RT and the observed RTs are positively correlated, facilitating interpretation.

Parametric coefficients				
Variable	Estimate	Std. Error	<i>t</i>	<i>p</i>
Intercept	-1.10	.04	-29.00	<.0001
Smooth terms				
Variable	edf	Ref.df	F	<i>p</i>
Frequency	1.00	1.00	225.94	<.0001
TargetCorTrain*CorDev-Centroid	7.78	9.87	1.88	.0451
Word length	5.34	6.40	57.95	<.0001
Prefix	7.64	9.00	37.78	<.0001
$R^2 = .464$				

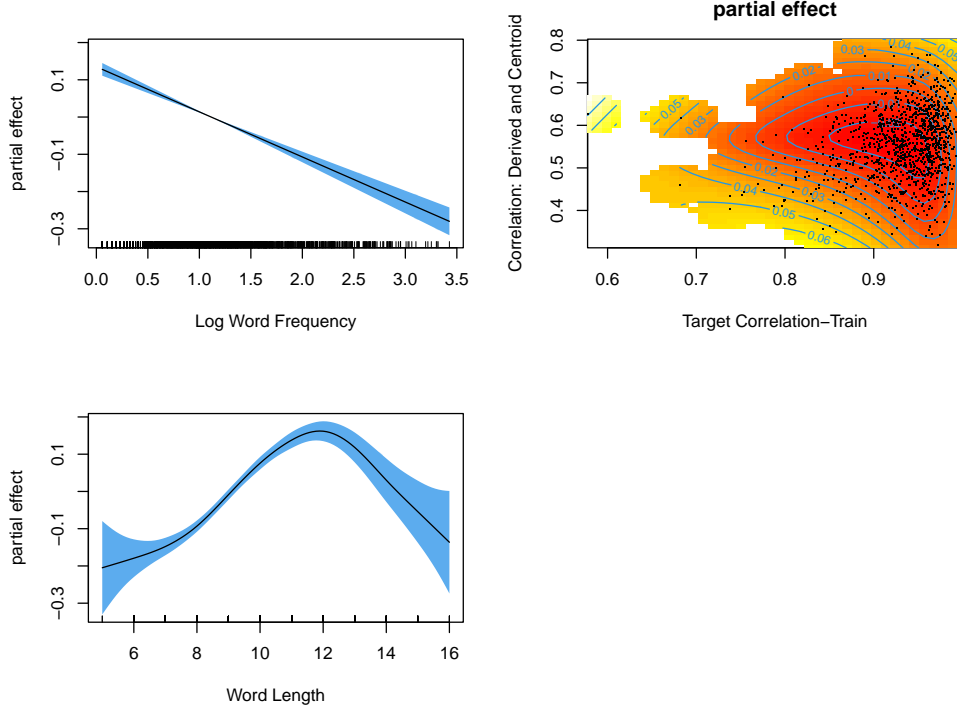


Figure 6: Partial effects of the GAMM fitted to the training data. Left column: Partial effect of word frequency on RT (top). Partial effect of word length on RT (bottom). Rugged lines on the x-axes represent the data. Right column: Partial effect of the interaction between **Target Correlation_{train}** of the DLM and the correlation between each derived word and their centroid. Red hues denote shorter RTs, orange/yellow hues indicate longer RTs. Black dots represent the data.

Table 7: GAMM fitted to the held-out test data. Word frequency was log-transformed. $\text{TargetCorTest} = \text{Target Correlation}_{\text{test}}$. CorDev-Centroid = Correlation between derived words and their centroid. The model syntax is $\text{inverse RT} \sim \text{s}(\log \text{ frequency}) + \text{te}(\text{TargetCorTest} * \text{Corr.Dev-Centroid} + \text{s}(\text{word length}) + \text{s}(\text{prefix}, \text{bs} = \text{'re'})$. $\text{Inverse RT} = -1000/\text{RT}$.

Parametric coefficients				
Variable	Estimate	Std. Error	<i>t</i>	<i>p</i>
Intercept	-1.14	.04	-25.84	<.0001
Smooth terms				
Variable	edf	Ref.df	F	<i>p</i>
Frequency	2.26	2.83	4.44	.00527
TargetCorTest*CorDev-Centroid	3.00	3.00	4.29	.00590
Word length	3.21	4.02	8.27	<.0001
Prefix	5.04	8.00	4.40	<.0001
$R^2 = .401$				

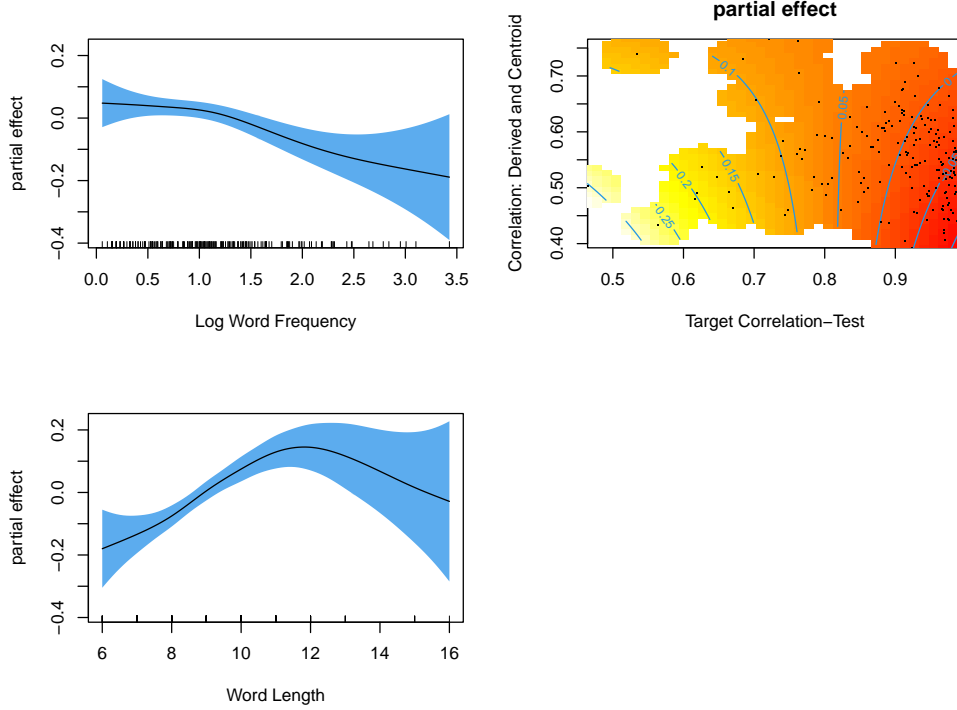


Figure 7: Left column: Partial effect of word frequency on RT (top). Partial effect of word length on RT (bottom). Rugged lines on the x-axes represent the data. Right column: Partial effect of the interaction between $\text{Target Correlation}_{\text{test}}$ and the correlation between each derived word and their centroid. Red hues denote shorter RTs, orange/yellow hues indicate longer RTs. Black dots represent the data.

Ideally, the effect of a productive system on cognition is facilitative for reading comprehension, allowing more automatic processing. Our findings on Malay provide evidence that this is largely the case, albeit not entirely without trade-offs. A very productive system also implies a densely populated morphological space. A word that highly resembles the prototypical meaning of many morphologically related words imposes additional demands. Not only must the reader navigate through large and tightly clustered word representations, but also expend additional cognitive resources in discriminating between related words.

2.5 Discussion

We have shown that generalization to unseen data, the hallmark of systemic morphological productivity, depends on systematic correspondences between form and meaning. Even in a complex semi-agglutinative system as that of Finnish, the DLM model capitalizes on the systematic correspondences between form and meaning for combinations of case and number. For prefixation in Malay, form-meaning correspondences are also detectable, although here the functional load of n-grams (as the fuzzy counterpart of morphemes and allomorphs) is shared with both n-grams of the stem and n-grams of suffixal exponents. For compounding, we have argued that tinkering is the modus operandi, with micro-systematicities at the level of individual pivots playing a minor role. Apparently, we have to distinguish between two fundamentally different kinds of productivity: systemic productivity on the one hand, and the productivity of bricolage on the other hand.

And yet, all the simulations above are flawed. Perhaps not deeply flawed, but flawed they are. The reason is that the DLM is a cognitive model of an individual speaker, but we have been using corpus data, i.e., aggregate data from large communities of speakers. Within such communities, language users have different areas of expertise, each of which come with their own specialized vocabularies. No single speaker is an expert

on the vocabularies of all socio-cultural expertise (lexicographers being a potential counterexample). Testing on held-out data is likely unfair for any individual speaker. In the next section, we present an attempt to address this discrepancy between community behavior (i.e., corpus data) and a computational model for an individual brain (the DLM).

3 Productivity of a concrete individual in their socio-historical context

So far, all measures and models were on the level of the language community. But all language users have their unique window onto language use, their own tailored input, which in turn shapes the way they use language. No two people receive exactly the same utterances. This basic insight of usage-based models makes it necessary to also take the level of the individual language user into account. Ideally, we need a compilation of everything one specific person has ever heard and read, and a compilation of everything they have ever said or written. For a variety of reasons, this is utopian (or rather, dystopian).

We can curate, however, something similar for prolific authors who also kept journals and/or regularly kept correspondences. For these persons, it is in principle possible to reconstruct their reading diet from their journals and letters, and to also collect their output from published texts, manuscripts, journals, letters, etc. This is what we have started to do with German writer Thomas Mann (1875-1955), the 1929 Nobel Prize in Literature laureate. One may argue that our focus on written input introduces a bias into our data. That is certainly true. However, Mann’s spoken input is beyond reach, and most of the new words and syntactic constructions he encountered he probably encountered through reading.

3.1 Data

We reconstructed Mann’s reading diet on the basis of his diaries (1918-21, 1933-1955), his letters (Bürgin and Mayer, 1987) and his personal library in Zurich. Each text was then collected; for practical reasons, we limited our search to texts that are digitally available on the internet. There are four classes of sources that vary in their quality:

- edited digitized versions from literature hubs (“Projekt Gutenberg”, “zeno.org”, “Wikisource”) — excellent quality;
- scanned versions with text extracted with optical character recognition (OCR) from university websites and Google Books — acceptable quality;
- digitized versions from other sites (e.g., Internet Archive), OCR’ed with Tesseract — mostly inadequate quality;
- scans from Mann’s personal library in Zurich’s website (<https://nb-web.tma.ethz.ch/>) — single page scans of texts which were only marginally used due to excessive effort.

So far, we have collected information on Mann’s readings up until the year 1925. Of these, texts until and including 1915 have been collected. Our input corpus consists of 437 texts that Mann states to have read until then, containing 41.4 million tokens. The output corpus consists of 32 texts that Mann wrote during that time, and it contains 500,000 tokens. These numbers show a striking imbalance between Mann’s readings and his writings: Mann read at least 80 times more than he wrote.

The Mann corpus is work in progress and will be expanded continually. The results presented here should be seen as a proof of concept that such a text collection is indeed useful and enables new insights. With that said, there are several shortcomings that have to be noted:

- The first documented texts in the input corpus are from 1889, when Mann was 14 years old; there are hardly any records of his readings in earlier years. His adolescent reading diet is mostly terra incognita.

- Newspapers have not been digitized yet. While we know which papers Mann regularly read, it is impossible to know which articles he actually read in which issue.
- Only a fraction of Mann’s letters has been added to the corpora so far. While the letters he wrote are conveniently collected in edited volumes, the letters he received are often scattered, which makes addition a time-consuming task.
- Some OCR text files are considerably worse than the rest (e.g., ”Geschichte der Lustseuche, die zu Ende des 15. Jahrhunderts in Europa ausbrach” by Philipp Gabriel Hensler from 1805). These texts are excluded for the construction of word embeddings (see below).

However, even if the corpus is not a complete collection of everything Mann ever read (and never will be, for principled reasons), it is possible to use it to investigate his word formations. The corpus may be skewed, but if it is, it is skewed towards the output: We systematically overestimate the number of newly coined formations.

Figure 8 shows the distribution of input texts and output texts (note the different scales for the input texts, left side, and the output text, right side).

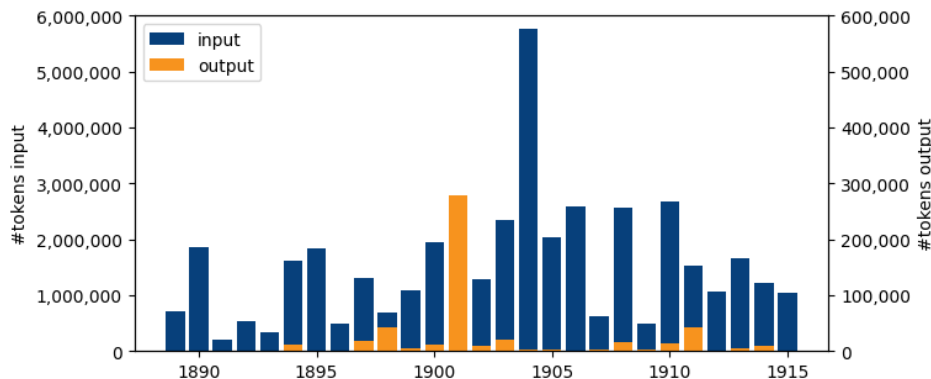


Figure 8: Distribution of running words in the input and the output part of the Mann corpus. The number of words that Mann read exceeds the number of words he wrote by a factor of around 80.

We investigate seven derivational patterns in the Mann corpus, four adjectival and three nominal ones. In previous research (Schneider-Wiejowski, 2011), *-los*, *-bar*, and *-tum* have been judged to be productive derivational patterns at the beginning of the 20th century (using Baayen (1992)’ potential productivity measure), while *-nis* is unproductive (Schneider-Wiejowski, 2011, 121f.). The pattern *-sal*, which Schneider-Wiejowski (2011) did not investigate, was probably unproductive as well (Fleischer and Barz, 2012). *-lich* and *-sam* are both more or less unproductive hapax-wise, but *-lich* exhibits around seven times as many types as *-sam* in a large diachronic corpus of German (Schneider-Wiejowski, 2011, 141f.). *-bar* and particularly *-los* are productive (Schneider-Wiejowski, 2011, 141f.). Table 8 gathers the relevant information on the different patterns.

Note that these are measures collected at the level of the language community. In the present paper, we will investigate whether these productivity values are mirrored at the level of an individual language user. It will also be curious to check the case of *-lich*, which is an outlier: an unproductive pattern with a large number of attested types.

To study the semantic side of word-formation products, we constructed word embeddings. All raw text files were lemmatized and PoS-tagged using the ParZu dependency parser for German (Sennrich et al., 2009), one of the most accurate parsers available (Ortmann et al., 2019). All lemmas were checked against a large lemma database of contemporary German which contains 326,946 entries (DeReWo, 2013) to automatically identify texts with a large degree of misspellings and errors due to faulty OCR. These misspellings and errors introduce unwanted variation when constructing word embeddings. We excluded the bottom quartile

Pattern	POS	example	potential productivity	#V
-bar	A	trinkbar 'drinkable'	0.07	274
-los	A	herzlos 'heartless'	0.11	190
-lich	A	freiheitlich	0.02	354
-sam	A	arbeitsam	0.02	54
-tum	N	bürgertum 'bourgeoisie'	0.06	39
-nis	N	ergebnis 'result'	0.004	45
-sal	N	trübsal 'misery'	-.?	-.?

Table 8: Patterns of interest and their supposed potential productivity values P (from Schneider-Wiejowski, 2011: 108, 114, 129, 132).

of texts (> 60% of types not recognized from the DeReWo list, 107 texts) from further processing; this leaves us with 330 texts with 16.5 million tokens. From these, we excluded punctuation marks and words tagged as either proper nouns or foreign words. Following a suggestion in Baayen et al. (2019), we split sentences longer than ten words into shorter clauses wherever possible, using the next conjunction, relative pronoun, interrogative pronoun, or punctuation mark as a break. The resulting sentences and clauses were then used to construct 300-dimensional Word2Vec embeddings (Mikolov et al., 2013) from the texts’ lemmas that span the whole time period of interest (parameters: window size = 5, iterations = 10, minimal count = 5, epochs = 15). The resulting semantic model consists of Mann-specific embeddings for 42,966 lexemes.

3.2 Descriptive results

Table 9 shows, for each pattern, the number of types and tokens in Mann’s readings and writings, together with the number of hapax legomena in the input and, most importantly for our investigation, the number of new types in Mann’s writings, words he has not read or written before.

Pattern	input			output		
	#types	#tokens	#hapaxes	#types	#tokens	#new types
<i>-lich</i>	991	300,476	2,661	351	4,922	17
<i>-los</i>	676	22,215	2,082	140	511	11
<i>-tum</i>	335	11,169	510	31	95	9
<i>-bar</i>	404	22,884	874	45	364	5
<i>-sam</i>	88	26,430	307	28	641	1
<i>-nis</i>	79	35,010	172	41	452	0
<i>-sal</i>	12	5,274	43	5	49	0

Table 9: Patterns of interest and their types, tokens, hapaxes in the input and output section of the Mann corpus

The number of new types is surprisingly low. In total, the patterns in question have 43 new formations among them, over the course of 26 years. On average, that amounts to less than two new types per year. At least for these derivational patterns, Mann rarely goes beyond what he has read.

The number of types in the input and the number of new types in the output are highly correlated (Spearman’s $\rho = 0.95, p = 0.0008$). The number of different types in the input is a good predictor for new types in the output. Note that the total number of types in the input (2,585) exceeds the number of new types in the output (43) by a factor of 60:1. In a way, then, these word formation patterns exhibit a rather leaky transmission.

The summed presentation in Table 9 gives a first overview, but it omits the temporal dimension that is also contained in the data. For each pattern, we can plot vocabulary growth curves (Baayen, 2001) for words in the input and in the output. For the output, we can further distinguish between words that Mann

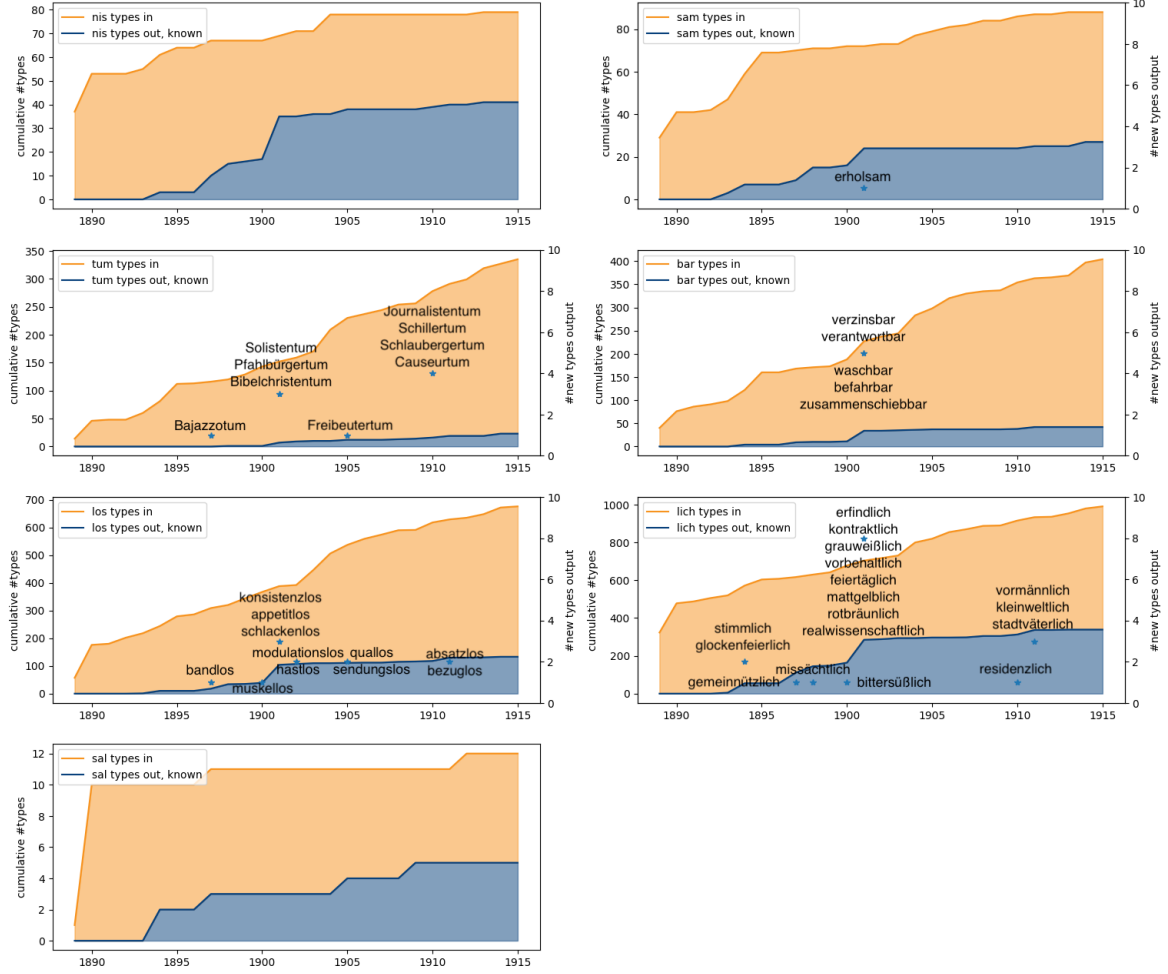


Figure 9: Vocabulary growth curves for the input (orange) and the output for the seven patterns. Output growth curves are plotted for words known from the input; new types that exceed the input are plotted as asterisks (axis on the right hand side).

had encountered in his reading, and which left traces in his memory and may have been repeated, and those that are not contained in the input, and which he probably newly formed.

The result is presented in Figure 9: The orange line is the growth curve for the types of each pattern. For example, in 1905, Mann had come across 78 different *-nis* types. The blue line indicates the growth curves for the known types, which is a fraction of the growth curve for the input types (the scale for the input is on the left hand side of a panel, the scale for the output is on the right hand side, where applicable). In 1905, Mann had used 38 of the 78 types he had read, or 49%. New words that Mann has probably not encountered in his readings are marked by an asterisk. For example, Mann formed one new *-tum* type in 1897, *Bajazzotum*.

The three unproductive patterns *-nis*, *-sam* and *-sal* are clearly distinguishable from the other patterns: Their vocabulary growth curves have reached a limit, and additional input tokens should not change the final number much. The other patterns exhibit a linear growth. The patterns are still expanding, and texts that Mann read will regularly have contained new words of these patterns, never seen before. This is one of the hallmarks of productive patterns.

The degree of re-use of known words differs considerably among the patterns. On the one end, there is

-nis, with a recycle rate of 52% (41 out of 79 types in the input are reproduced). On the opposing end, there is *-tum*, with a 7% rate (23 out of 335 types in the input are reproduced). The other patterns occupy middle ground, with a tendency of more productive patterns towards lower recycle rates.

In general, many hypotheses about the productivity of *-tum*, *-bar*, *-los*, and *-lich* and the unproductivity of *-nis*, *-sam*, and *-sal* are conceivable: Apart from the recycle rate, it could be the number of cumulative types in the input, or the rate of hapaxes among the tokens (productivity in the narrow sense), or the rate of new types in the input among all input types (P_{neo} in Berg, 2020). These and other hypotheses will be tested statistically in the next section.

3.3 Statistical Modeling

We model the number of new words in Mann’s own writings per year, with the following potential predictors for each pattern:

- The number of types in the input in that year, log-transformed;
- the cumulative number of types in the input, log-transformed;
- the number of tokens in the input in that year, log-transformed;
- the number of hapaxes in the input in that year;
- the number of new types in the input in that year, log-transformed;
- the number of known types that Mann uses in his writings, log-transformed;
- the average distance between the embeddings of all types in the input that year and the centroid of all embeddings of all types encountered in the input so far;
- the ratio of hapaxes that year and the sum of all tokens that year (‘productivity in the narrow sense’);
- the ratio of new words that year and all types that year (P_{neo});
- the ratio between the cumulative re-used types and the cumulative set of all types (‘recycle rate’);
- the year of reading/writing.

Each year in which Mann published a text represents one case for each pattern; this leads to a table of 18 years \times 7 patterns, each with one response variable and 10 predictors. We analyzed these data using Generalized Additive Models (GAMs, Wood, 2017), which are able to capture non-linear relationships between predictors and response variables.

We started with a full model with smooth terms for all variables plus an interaction term between centroid distance and the number of different known types Mann uses, and subsequently reduced the number of predictors if they were not significant at least at the 0.05 level. We use the log-transformed number of types that Mann uses as an offset in the model; after all, we are not interested in the absolute number of Mann’s new formations, but as a fraction of the number of types of that pattern — in years with long texts and many relevant types, we expect to see more new types compared to years with a lower output. We also tested whether predictors interact with each other; our criterion for model selection was based on Akaike’s information criterion (AIC). The model with the lowest AIC has one smooth term (P_{neo}) and one tensor product smooth (the interaction between the average distance of a pattern’s embeddings from the pattern’s centroid on the one hand, and the log-transformed number of reproduced types in Mann’s output on the other hand). Both the smooth term and the interaction significantly predict the number of Mann’s new formations (see Table 10).

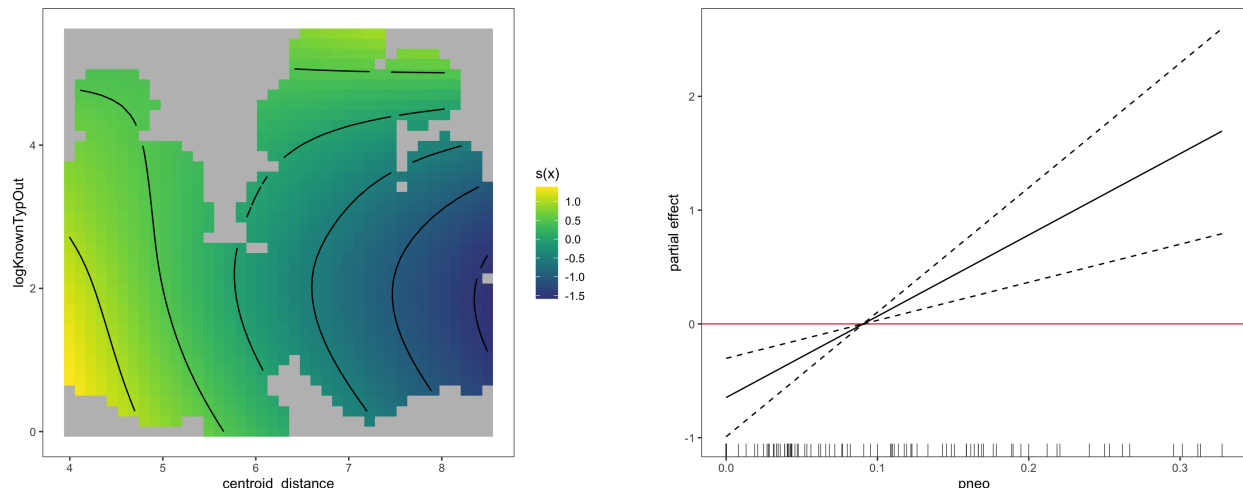


Figure 10: Visualization of the model terms in the best GAM model discussed above: average centroid distance and the number of known types interact (left plot). Smooth of P_{neo} as partial effect (right plot)

Table 10: GAMM fitted to the Mann data. CentroidDistance = Average distance between a word’s embedding and the centroid of the pattern. logKnownTypOut = log-transformed count of different lexemes in Mann’s writings that were part of his readings up until that point.

Parametric coefficients				
Variable	Estimate	Std. Error	<i>t</i>	<i>p</i>
Intercept	-4.3696	0.3126	-13.98	<.0001
Smooth terms				
Variable	edf	Ref.df	F	<i>p</i>
P_{neo}	0.931	9	13.57	<.0001
CentroidDistance*logKnownTypOut	3.104	24	10.67	0.0055
$R^2 = .741$				

The left panel of Figure 10 shows the interaction between the average distance from the centroid (horizontal axis) and the number of known types that Mann produces. Yellow tones indicate a higher probability for a new word. Decreasing semantic homogeneity (i.e., an increasing distance from the centroid) inhibits productivity. This seems plausible: The less homogeneous a pattern is, the harder it is to generalize it to new formations. This effect can (at least partly) be remedied by an increase in the number of known types in the output (top right part of the contour plot). In the right panel of Figure 10, we observe an effect of P_{neo} : The probability for a new formation increases with the ratio of new types among all types.

In other words, the more known words Mann re-uses, and the more semantically homogeneous a pattern is, the more likely he is to go beyond what he has read. The same holds for the ratio of new words among all words: the more new words there are for any pattern relative to known words, the more likely Mann is to produce a new word himself.

From a cognitive perspective, this makes sense. The more types of a pattern somebody reuses, and the more similar they are, the easier it is to go beyond what has been read and heard so far. So easy, in fact, that Mann arguably did not perceive these formations as new, bold and/or odd. This leads to an interesting question: Given the words Mann has read, and the order he has read them in, and also given the meanings we extracted from the respective texts (using word2vec) — can we construct a model of a mental lexicon to assess the ‘ease’ of Mann’s word formations? Of course not; but we can approximate it, and we set out to do so using the discriminative lexicon model introduced above in the remainder of this section.

3.4 Modeling: DLM

We trained various DLM models on pairings of lexemes and word embeddings, preserving the frequency with which the words appear in Mann’s readings. Lexemes are represented as 4-grams, while the 300-dimensional Word2Vec vectors introduced above serve as word embeddings. All calculations were performed using the Julia package JudiLing (Heitmeier et al., 2025)³, using both endstate learning and frequency-informed learning (Heitmeier et al., 2023).

The Word2Vec algorithm contains a frequency threshold (“min_count” in the gensim implementation) which was set to 5. A lower threshold leads to more unreliable embeddings, but is more inclusive of low-frequency words; a higher threshold leads to more robust embeddings, but contains fewer rare words. This is a dilemma for anyone who wishes to investigate morphological productivity (often manifest in rare words) with the means of context-independent word embeddings that are not given access to substrings within words; it means that the embeddings are the bottleneck for our model construction, limiting the number of different lexemes to the 42,966 different lexemes mentioned above. These lexemes occur 13 million times in Mann’s readings between 1888 and 1915.

We trained two models every five years, one using endstate-of-learning (EOL) and one taking frequency of occurrence into account (frequency-informed learning, FIL). Each model was trained with all lexemes that had been read up until and including the year in question: The 1893 model, for example, contains all lexemes in Mann’s reading up until 1893 (for which embeddings exist). Table 11 reports the accuracies of the resulting models, both as a conservative measure (@1, the fraction of all form-meaning correspondences correctly reproduced by the model) and more lenient measure (@10, the fraction of all form-meaning correspondences where the model’s estimate is among the top 10 items). It is obvious that the models get worse over time: The more words there are, the harder it gets for a linear model to distinguish them; after all, between 1898 and 1903, the number of different lexemes more than doubles. It is also obvious that frequency-informed learning fares much worse than endstate-of-learning, when evaluated on accuracy @1 or accuracy @10 on a type basis.

The endstate-of-learning models show what is theoretically possible with the data, given unlimited training time on the input dataset. The frequency-informed models, on the other hand, are more realistic for the situation at hand: They will learn frequent words well, and inevitably struggle with low-frequency words.

Although at first sight the low type accuracies of FIL are surprisingly disappointing, these low accuracies are the inevitable consequence of taking a usage-based approach. To see this, consider the consequence of the fact that in word frequency distributions, the hapax legomena constitute roughly 50% of the types. Words that the model encounters only once cannot be learned given the barrage of tokens of high frequency words that the model is continuously confronted with. It therefore makes sense to evaluate FIL-based models on what proportion of the tokens are properly recognized. Table 11 therefore lists accuracy @1 using token-based evaluation. These accuracies are substantially higher and, across all 5 measurement years, outperform type-based accuracy using endstate learning. In other words, the words that are useful to know because they occur frequently are learned much better by FIL-trained models.

A further consideration to keep in mind is that it is unrealistic to expect a model of isolated word comprehension to be able to predict words perfectly: words typically occur in utterances, and the utterance will often be an excellent guide to a more precise understanding that cannot be achieved on the basis of the bottom-up input of just the word on its own. A FIL model is just a small part of a much more comprehensive set of language skills.

Finally, the performance of FIL raises interesting questions about how well one can generalize from the high-frequency words belonging to a word formation pattern to the low-frequency words in that pattern. Higher frequency words, and especially higher-frequency derived words, tend to have more senses and tend to have more semantic idiosyncrasies. Thus, the more a word formation pattern is dominated by high-frequency words, the less one may expect generalization to low-frequency words in that pattern to be successful. The category-conditioned productivity measure (the Good-Turing ratio of hapax legomena to tokens) captures this by penalizing, however crudely, for token frequencies.

³The JudiLing package is in the julia repository, the latest development version can be downloaded from <https://github.com/quantling/JudiLing>.

Table 11: Comprehension accuracies for training data. Accuracies @1 pertain to perfect recognition, accuracies @10 to accuracy evaluated as belonging to the top 10 nearest neighbors of the intended meaning. Tokens and types refer to token-based and type-based evaluation.

learning	@1 types	@1 tokens	@10 types	#lexemes
1893, EOL	0.809	-	0.929	15,219
1893, FIL	0.121	0.819	0.238	
1898, EOL	0.649	-	0.815	33,777
1898, FIL	0.062	0.826	0.125	
1903, EOL	0.625	-	0.792	38,606
1903, FIL	0.056	0.822	0.112	
1908, EOL	0.615	-	0.78	41,711
1908, FIL	0.052	0.817	0.104	
1913, EOL	0.614	-	0.777	42,747
1913, FIL	0.051	0.815	0.102	

In what follows, we make use of FIL-based models, firstly because they are more in line with usage-based linguistics, and second, because we expect to find structure in the model’s association values, even though model predictions for individual lower-frequency types will lack precision. Specifically, in the following, we use the first FIL model in Table 11 (1893) and ask, for each pattern of interest, how well these patterns have been learned: What is the accuracy for all *-nis* words in the training set, for example?

Table 12 shows that the accuracy varies greatly between the patterns. The level of accuracy for individual patterns is higher than the accuracy of the whole model, save for *-los*. Infrequent patterns (such as *-sal*) and frequent patterns (such as *-lich*) have surprisingly high accuracies which are unexpected given their characterizations as unproductive or less productive. What we are seeing is an effect of the frequency distribution within each pattern. Generally, the more tokens per type there are, the better the model learns the pattern’s words (Spearman’s ρ between acc@1 and tokens per type: 0.829, $p = 0.021$). That means the FIL model favors patterns that are high in tokens but low in types — in other words, unproductive patterns, exactly as expected. Accordingly, token-based accuracy (@1 tokens) is highest for the unproductive patterns. This then would suggest that comprehension accuracy for training data in an FIL model cannot be predictive for a pattern’s productivity, conceptualized as the capacity to generalize of a word formation pattern to unseen types.

Table 12: Comprehension accuracies for training data, model 1893, using FIL, for the individual word formation patterns.

learning	@1 types	@10 types	@1 tokens	# types	# tokens	tokens per type
-sam	0.272	0.546	0.909	22	286	13
-nis	0.27	0.324	0.806	37	397	10.7
-sal	0.25	0.25	0.903	4	72	18
-bar	0.207	0.31	0.769	29	242	8.3
-lich	0.158	0.288	0.752	287	3,047	10.6
-tum	0.158	0.316	0.782	19	142	7.5
-los	0.058	0.161	0.415	87	234	2.6

This conclusion is supported by an examination of how well the FIL model predicts held-out data. Usually, models are validated by carefully splitting the data into training and held-out data (as we did above in section 2). In the case of the Mann corpus, however, we know exactly which new words he encountered after 1893, and we will use these words (grouped into their respective patterns) as validation data. For example, there are 35 *-lich* words that Mann encountered after 1893, among them formations

like *erzbischöflich* ('archbishop-ly') or *weltanschaulich* ('ideologically'). For each of these words, we let the model predict an embedding, and we then compare this predicted embedding with the empirical Word2Vec embedding. Table 13 shows that the FIL-based model trained on data until 1893 does not generalize beyond the training data. For none of the held-out words, the empirical embedding comes in among the ten most similar to the predicted one. Even if we keep in mind that each predicted embedding is evaluated against the full set of 15,210 empirical embeddings, prediction accuracy for FIL is underwhelming.

However, we can lower the accuracy threshold to the set of the 100 or 1000 most similar embeddings, and then some structure becomes visible. The patterns that have the lowest number of tokens per type in Table 12 have the highest accuracy values; both are in fact significantly negatively correlated (Spearman's $\rho = -0.943, p = 0.005$). So while the FIL model has difficulties learning words from productive patterns, it is able to learn something about the pattern itself, which in turn enables it to understand new words slightly better than words from more unproductive patterns. But precise prediction for held-out data is clearly out of reach.

Table 13: Comprehension accuracies for held-out data (Types), model 1893, FIL.

learning	@10	@100	@1000	#new words ≤ 1893
-los	0.0	0.024	0.342	41
-lich	0.0	0.143	0.257	35
-bar	0.0	0.0	0.273	11
-tum	0.0	0.1	0.3	10
-sam	0.0	0.0	0.25	4
-nis	0.0	0.0	0.0	3

Above, we pointed out that training a model with FIL implies giving high-frequency words the opportunity to dominate learning. We anticipated adverse effects for generalization to held-out data. This expectation was borne out, but there is further snag that is worth discussing.

Recall that the Word2Vec algorithm needs a certain minimum amount of occurrences to generalize over a word's co-occurrences and construct high-quality embeddings. In our case, we set the minimum value to 5. This in turn means that we exclude nonce formations, the hallmark of productive patterns. These nonce formations are in general more transparent (in the sense of compositional) than frequent formations, which tend to be charged with any kind of additional semantic, syntactic, or pragmatic information. We expect the DLM model to predict infrequent — and semantically regular — formations particularly well, but sadly, we have no embeddings for these words. If we had, the accuracy values for the FIL model (Table 13) would arguably increase markedly (after all, we are missing out on 255 *-lich* types that occur less than 5 times in the data set, such as *österlich* 'Easterly').

In summary, not only is the model trained without low-frequency words, it is also tested on words that are not the lowest frequency words. Given the Zipfian nature of word frequency distributions, most of the 'low-frequency' words that occur more often than 5 times likely occur much more often than 5 times.

Due to the currently unavoidable limitations of our data, the FIL model is confronted with the task of generalizing from a dataset with systematicities that are much more limited than we had originally hoped. How severe these limitations are becomes apparent when we replace learning with FIL with learning with a deep neural network with a hidden layer of 1000 RELU units, with early stopping to avoid overfitting. This model achieves 99.5% accuracy on the training data, but is not much better at predicting the held-out data. Accuracy @10, for example, is zero for all patterns except for *bar*, which clocks at 9%. As can be seen in Table 14, the deep model gets most of the handful of novel types for *-sam* and *-nis* correct, but for the other word formation patterns, accuracy @1000 is only slightly higher than for FIL. As this deep learning network is not frequency aware, and as a more powerful version of endstate learning informs us about what is learnable in the limit of experience, we can conclude that the present dataset is too unsystematic to afford precise prediction.

One possible way to address this problem would be to measure the model's accuracy not against the

Table 14: Comprehension accuracies for held-out data (types), model 1893, deep mapping.

learning	@10	@100	@1000	# new words > 1893
-los	0.0	0.073	0.463	41
-lich	0.0	0.057	0.371	35
-bar	0.09	0.09	0.363	11
-tum	0.0	0.2	0.3	10
-sam	0.0	0.0	1	4
-nis	0.0	0.0	0.667	3

empirical embeddings, but against a constructed embedding obtained by adding the centroid embedding to the embedding of the base word, see, e.g., [Baayen et al. \(2019\)](#) and [Heitmeier et al. \(2025\)](#). All words that are currently excluded because their frequency is too low to obtain a reliable embedding can then be included with a constructed embedding. These constructed embeddings are likely too regular, and therefore err on the side of semantic transparency. Nevertheless, their inclusion would allow us to test to what extent inclusion of the low-frequency words will equip FIL with improved prediction accuracy for held-out words. Given the overwhelming effect of high-frequency words in FIL learning, it is far from self-evident that prediction accuracy will indeed improve. If the answer is negative, this suggests that research on productivity in a usage-based perspective will have to focus far more on how usage in utterances supports the interpretation of novel words, using contextualized embeddings. We leave this question for further research.

One final question we would like to address in this section is, what is the FIL model learning actually? Even if the accuracy values are low, some of the pattern’s aspects are learned, as witnessed by the close correlation between tokens per type and accuracy. Where is the information about, for example, *-nis* words stored in a model that makes do without words, patterns, rules and constraints? The answer, as in section 2 above, is: at least partly in the pattern’s 4-grams.

Table 15 shows the top five strongest correlations between the row vectors in \mathbf{F} and the centroid embeddings of the seven derivational patterns discussed in this section. Each centroid, save for the most type-frequent pattern *-lich*, is most strongly associated with its respective 4-gram in word-final position.

Table 15: Top five strongest correlations between the row vectors in \mathbf{F} and the centroid embeddings of the seven derivational patterns in an FIL model of Mann’s readings until 1893. Coefficients highlighted in red represent the strongest correlation between the row vectors of the 4-gram and the embeddings of the centroid.

	-BAR	-LICH	-LOS	-NIS	-SAL	-SAM	-TUM
bar#	0.82	0.704	-	-	-	0.572	-
tig#	0.671	0.772	-	-	-	0.575	-
dig#	0.628	0.709	-	-	-	-	-
ial#	0.57	.0634	-	-	-	-	-
siv#	0.547	0.605	-	-	-	-	-
los#	-	-	0.685	-	-	-	-
nlos	-	-	0.642	-	-	-	-
tlos	-	-	0.62	-	-	-	-
#unb	-	-	0.602	-	-	-	-
utal	-	-	0.6	-	-	-	-
nis#	-	-	-	0.949	-	-	-
ung#	-	-	-	0.841	-	-	-
gabe	-	-	-	0.713	-	-	-
ion#	-	-	-	0.703	-	-	0.78
tnis	-	-	-	0.703	-	-	-
sal#	-	-	-	-	0.894	-	-
ksal	-	-	-	-	0.878	-	-
icks	-	-	-	-	0.878	-	-
cksa	-	-	-	-	0.876	-	-
hick	-	-	-	-	0.751	-	-
sam#	-	-	-	-	-	0.891	-
tsam	-	-	-	-	-	0.554	-
egt#	-	-	-	-	-	0.536	-
tum#	-	-	-	-	-	-	0.859
mus	-	-	-	-	-	-	0.742
tät#	-	-	-	-	-	-	0.742
ität	-	-	-	-	-	-	0.715

This demonstrates that the model is on the right track: It has learned that the part of the word that carries the pattern’s letter chunk is connected to the pattern’s semantic contribution. Closer inspection reveals a more intricate knowledge of German derivational morphology:

- Among the top correlations of adjectival -BAR and -LICH are 4-grams containing other adjectival patterns (native *-ig* and foreign *-al* and *-iv*).
- For the privative adjectival pattern -LOS, one of the most strongly correlated forms (apart from *-los* in three variations) is the negative prefix *un-*.
- -NIS forms action nouns or nouns that denote the result of a dynamic action. Among the strongest correlations for the centroid are two other nominal patterns with a similar function, native *-ung* and foreign *-ion*.
- Nominal -TUM is most similar to foreign nominal patterns *-ität*, *-ion*, and *-ismus*, all of which are partially overlapping semantically.
- -SAL is particularly interesting because the highest correlations involve bits of *Schicksal* ‘fate’, by far the most frequent of the few *-sal* words. The pattern is not related to other patterns, possibly because it is not recognized as a pattern by the model.

In all these cases, it is the semantic and/or syntactic overlap between the centroids of different patterns which leads to the observed correlations. It is worth stressing, though, that these correlations emerge without explicitly providing any morphological and/or lexical information. Considered jointly, we conclude that the FIL model, in spite of the limitations of our dataset, has learned a lot about the relations between form and meaning for the German word formation patterns in our Mann corpus.

4 General discussion

In this study, we have used the Discriminative Lexicon Model (DLM) as a tool for probing morphological productivity. Our main findings are the following. First, starting out from the fact that without systematicity, machine learning and human learning cannot generalize to previously unseen, ‘held-out’, data, we formalized the degree of productivity of a word formation pattern as the prediction accuracy of the DLM for held-out data. For Finnish nominal inflection for case and number, the predictions of the DLM align well with the different degrees of productivity of the many inflectional classes of Finnish, as gauged with classical well-established measures. For prefixal derivation in Malay, accuracies for held out data decreased, unsurprisingly as the semantics of derived words typically are less transparent than the semantics of inflected words. For English compounds, accuracy for held out data was the lowest, which is also unsurprising as the meanings of novel compounds can be notoriously difficult to predict (see, e.g., Schäfer and Bell, 2020), and given that there are no systematic form-meaning parallels (which led Schultink (1962) to argue that compounds do not form a morphological category). Since in terms of type counts, compounding is extremely productive, we conclude that systemic productivity can be assessed well with the DLM, but that the productivity of compounding is outside the scope of the model, as this type of word formation is made possible by onomasiological tinkering (‘bricolage’).

Second, we have shown by means of an examination of what Thomas Mann is likely to have read, and what he wrote, that the rate at which Mann produces novel derived words is extremely low. There are far more novel words in his input than in his output. We document a strong correlation between ($\rho = 0.95$) the rate of novel words in Mann’s input and the rate of novel words in his output. Mann’s comprehension productivity (assuming that he understood what he read) was clearly much higher than his production productivity. This asymmetry is also well documented in the acquisition literature (see, e.g., Gershkoff-Stowe and Hahn, 2013, and references cited there) and emerges time and again in the DLM (Chuang et al., 2020; Heitmeier et al., 2025). However, the magnitude of the difference in comprehension and production productivity for a prolific and gifted writer such as Thomas Mann surprised us. But it fits well with existing studies investigating the spread of novel words (De Smedt, 2012) suggesting that innovations spread through social networks, starting with a single innovator, being adopted by followers, and with bursts of popularity possibly standing in the way of entrenchment in a community’s lexicon (Chesley and Baayen, 2010).

Third, across all case studies, the centroid in semantic space of all words sharing a morphological pattern turned out to play an important role. When a DLM model is set up with linear mappings between form and meaning, the row vectors of the comprehension mapping and the column vectors of the production mapping specify semantic vectors. It turns out that the more an n-gram overlaps with an exponent, the stronger its vector in the mappings correlates with the centroid of the embeddings of the words sharing that exponent. The centroids are the ‘average’ meanings of the exponents, and the linear mappings converge on these centroids as the best estimates for the semantics of sublexical n-grams. We document this across all our datasets: Finnish nominal inflection, Malay prefixation, and the derivational suffixes in the writings of Thomas Mann. For Malay, the correlation of the embedding of a derived word with the centroid of its morphological pattern emerges as both a measure of morphological transparency and a robust predictor of lexical decision times. And Thomas Mann is less likely to produce a novel derived word with a given suffix the greater the average distance is of the embeddings of all derived words to the corresponding centroid. In other words, the less transparent the word formation pattern, the less its production productivity. One of the advantages of computational modeling is that the consequences of theoretical decisions are clearly visible. In the present study, the simulation experiments with endstate learning (EOL) and deep learning provide insight into what can be learned with infinite cycles through the dataset of word types. Frequency-informed learning (FIL) implements a usage-based alternative. Training mappings with FIL shows that words that occur once only cannot be learned — a result that will not surprise anyone with experience in machine learning. Since in word frequency distributions, the hapax legomena often make up around 50% of the types, type accuracies for FIL are unavoidably low. On the other hand, since high-frequency words are learned well, the proportion of tokens that is learned correctly is much higher. Nevertheless, a FIL model for Mann trained on the materials from before 1893 struggles with understanding the novel forms encountered after 1893, although it may get the gist correct. Deep learning does not fare much better. The reason for this is

the limitations of our dataset that come with the word2vec embeddings that we used. These embeddings were calculated from Mann’s reconstructed own experience using word2vec. In order to obtain reasonably reliable embeddings, embeddings were calculated only for words with a frequency equal to or greater than 5. As a consequence, both the training data and the test data are biased towards more frequent, and hence less semantically predictable, words. Unsurprisingly, precise prediction suffers dearly. Nevertheless, it is clear from an inspection of how pattern centroids are represented in the comprehension mapping that, first, the FIL model has already learned a lot about the general semantics but also that, second, the centroids of semantically similar affixes, as estimated from our dataset, are similar and are confusable. An open question is whether including constructed ‘regularized’ embeddings for the lowest-frequency words in the Mann corpus will enable FIL to generate more precise predictions for held-out data.

This kind of problem is not specific to our case study of Thomas Mann. The studies of Finnish inflection and Malay derivation make use of corpus frequencies, which are estimates of community usage. As a consequence, for individual language users, the experience with the lowest frequency words is severely overestimated, and experience with higher-frequency words somewhat underestimated. As a consequence, the models are unavoidably biased.

An advantage of computational modeling with the DLM is that this theoretical framework does not require the analyst to make discrete decisions about whether words are exceptional or regular, as in the theory of Yang (2016). According to his tolerance principle, a rule is completely unproductive when the number of types V divided by the natural logarithm of V is greater than the number of types that are exceptional in some way. Otherwise, it is completely productive. But what counts as an exception? In Finnish, the semantics of plurality vary systematically across case (Nikolaev et al., 2023), and within case-number combinations, individual words again show considerable variation. Does this semantic diversification make the whole inflectional system unproductive? Furthermore, at the form side, how to properly set up the inflectional classes of Finnish nouns is far from clear, the current standard being a compromise between lumping and splitting (Nikolaev et al., 2025). The DLM does not force the analyst to make any such ad hoc decisions. The model is not informed about the inflectional classes, nor about case and number. It is provided with rich (but not perfect) information about words’ meanings, and it sets up form embeddings using simple n-grams. The mappings between form and meaning optimize for prediction, but at the same time learn to remember the idiosyncracies of individual words’ forms and meanings. This approach offers novel ways of conceptualizing productivity (generalization as opposed to bricolage), and of assessing degrees of productivity for word formation patterns and associated allomorphs, for inflection and inflectional classes, and for compounding. As we have shown, this approach also helps clarify the challenges for the analyst wishing to understand productivity at the level of the individual. Our ideal for productivity research is a much more ambitious simulation model in which virtual agents with DLM-like mental lexicons are interacting over time in socially stratified communities. Such a model would make it possible to simulate and hopefully predict how innovations might spread in communities, and how productivity changes over the lifetime of both individuals and speaker communities. While such a full model of speakers within their groups within a language community is currently out of reach, studies that focus on simplified aspects are conceivable and may yield interesting and novel results. For example, if we equip a small number of agents with a toy DLM lexicon and have them randomly produce „texts“ based on their mappings, and every agent „reads“ every other agent’s texts, we may observe changes to the individual mappings over time that result in change on the level of the community, as measured by output texts. If these changes turn out to be similar to actual language change, this would both explain the mechanics of language change in a new way, and indicate that an agent-based DLM approach is on the right track.

We are indebted to Alexandre Nikolaev for his feedback on this study.

Appendix

A Specifics of the CAOSS model

Let \mathbf{L} , \mathbf{R} and \mathbf{C} denote the embeddings of the left constituents, the right constituents, and the compounds respectively. The CAOSS model sets up a linear mapping \mathbf{M} by solving

$$[\mathbf{LR}]\mathbf{M} = \mathbf{C}.$$

The mapping \mathbf{M} consists of two blocks, one transforming \mathbf{L} , and one transforming \mathbf{R} . The output of these transformations are added to obtain \mathbf{C} :

$$[\mathbf{LR}] = \begin{bmatrix} \mathbf{M}_L \\ \mathbf{M}_R \end{bmatrix} \mathbf{C}.$$

Figure 11 clarifies that the highest values of \mathbf{M}_L and \mathbf{M}_R are found on the diagonal. If all off-diagonal elements were exactly zero, these matrices would be identity matrices, and their joint effect would be simply addition of the constituent embeddings:

$$\hat{\mathbf{C}} = \mathbf{L} + \mathbf{R}.$$

Figure 11 shows that CAOSS is finding a solution that is close to, but not identical to, simple vector addition.

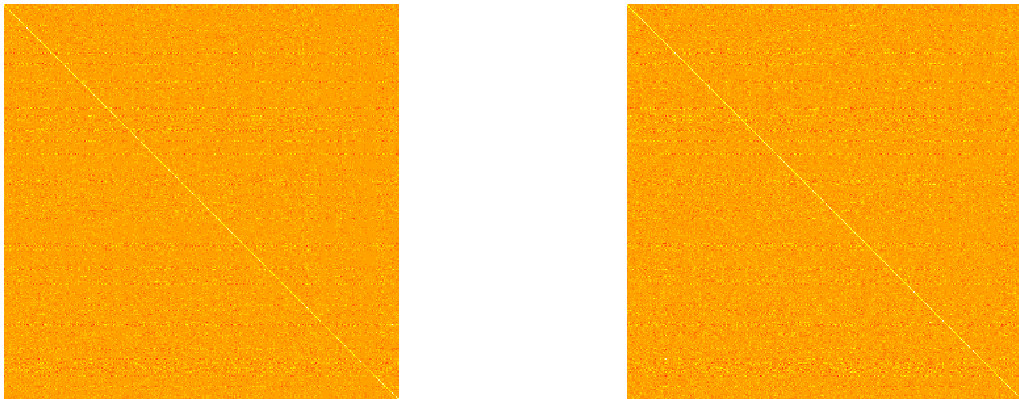


Figure 11: CAOSS mapping matrices, estimated for the dataset with 8147 English compounds.

References

- Baayen, R. H. (1992). A quantitative approach to morphological productivity. In Booij, G. E. and Marle, J. v., editors, *Yearbook of Morphology 1991*, pages 109–149. Kluwer, Dordrecht.
- Baayen, R. H. (1993). On frequency, transparency, and productivity. In Booij, G. E. and van Marle, J., editors, *Yearbook of Morphology 1992*, pages 181–208. Kluwer Academic Publishers, Dordrecht.
- Baayen, R. H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.
- 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*.
- Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.
- Berg, K. (2020). Changes in the productivity of word-formation patterns: Some methodological remarks. *Linguistics*, 58(4):1117–1150.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Booij, G. E. (1977). *Dutch Morphology. A Study of Word Formation in Generative Grammar*. Foris, Dordrecht.
- Bürgin, H. and Mayer, H.-O. (1977-1987). *Die Briefe Thomas Manns: Regesten und Register*. Fischer Frakkfurt.
- Chesley, P. and Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48:1343–1374.
- Chuang, Y.-Y., Bell, M., Banke, I., and Baayen, R. H. (2020). Bilingual and multilingual mental lexicon: a modeling study with Linear Discriminative Learning. *Language Learning*, pages 219–292.
- De Saussure, F. (1966). *Course in General Linguistics*. McGraw, New York.
- De Smedt, K. (2012). Ash compound frenzy: A case study in the norwegian newspaper corpus. In *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, pages 241–256. John Benjamins Publishing Company.
- DeReWo, K. W. (2013). vww-bll-320000g-2012-12-31-1.0.
- Fleischer, W. and Barz, I. (2012). *Wortbildung der deutschen Gegenwartssprache*. Walter de Gruyter.
- Gershkoff-Stowe, L. and Hahn, E. R. (2013). Word comprehension and production asymmetries in children and adults. *Journal of experimental child psychology*, 114(4):489–509.
- Heitmeier, M., Chuang, Y., Axen, S., and Baayen, R. H. (2023). Frequency-informed linear discriminative learning. *Front. Hum. Neurosci., Sec. Speech and Language*, 17.
- Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2024). The discriminative lexicon: Theory and implementation in the julia package judiling.
- Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2025). *The Discriminative Lexicon: Theory and implementation in the Julia package JudiLing*. Cambridge University Press, Cambridge. in press.
- Hoffman, D. (2019). *The case against reality: Why evolution hid the truth from our eyes*. WW Norton & Company.

- Husserl, E. (1913). *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*. Felix Meiner Verlag (2009).
- Kant, I., Guyer, P., and Wood, A. W. (1781/1999). *Critique of pure reason*. Cambridge University Press.
- Kuperman, V. and Bertram, R. (2013). Moving spaces: Spelling alternation in English noun-noun compounds. *Language and Cognitive Processes*, 28(7):939–966.
- Lévi-Strauss, C. (1962). *Savage mind*. University of Chicago.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Marelli, M., Gagné, C. L., and Spalding, T. L. (2017). Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166:207–224.
- Maziyah Mohamed, M. and Baayen, R. H. (2025). An exploratory analysis on the explanatory potential of embedding-based measures of semantic transparency for Malay word recognition. *arXiv preprint arXiv:2505.05973*.
- Maziyah Mohamed, M. and Jared, D. (2023). The distributional properties of prefixes influence lexical decision latencies: Evidence from Malay. *The Mental Lexicon*, 18(2):218–264.
- Maziyah Mohamed, M. and Jared, D. (2025). Malay lexicon project 3: The impact of orthographic–semantic consistency on lexical decision latencies. *Quarterly Journal of Experimental Psychology*, 78(1):22–47.
- Maziyah Mohamed, M., Yap, M. J., Chee, Q. W., and Jared, D. (2023). Malay lexicon project 2: Morphology in Malay word recognition. *Memory & Cognition*, 51(3):647–665.
- Merleau-Ponty, M., Landes, D., Carman, T., and Lefort, C. (2013). *Phenomenology of perception*. Routledge.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Nikolaev, A., Chuang, Y.-Y., and Baayen, R. H. (2023). A generating model for Finnish nominal inflection using distributional semantics. *The Mental Lexicon*.
- Nikolaev, A., Chuang, Y.-Y., and Baayen, R. H. (2025). Analyzing Finnish inflectional classes through discriminative lexicon and deep learning models. *arXiv preprint*. Manuscript submitted for publication.
- Ortmann, K., Roussel, A., and Dipper, S. (2019). Evaluating off-the-shelf nlp tools for German. In *KONVENS*.
- Schäfer, M. and Bell, M. J. (2020). Constituent polysemy and interpretational diversity in attested English novel compounds. *The Mental Lexicon*, 15(1):42–61.
- Schneider-Wiejowski, K. (2011). *Produktivität in der deutschen Derivationsmorphologie*. PhD. dissertation, University of Bielefeld.
- Schultink, H. (1961). Produktiviteit als morfologisch fenomeen. *Forum der Letteren*, 2:110–125.
- Schultink, H. (1962). *De Morfologische Valentie van het Ongelede Adjectief in Modern Nederlands*. van Goor & Zonen, Den Haag.
- Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A new hybrid dependency parser for german. *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

- Shen, T. and Baayen, H. R. (2022). Productivity and semantic transparency: An exploration of word formation in Mandarin Chinese. *The Mental Lexicon*.
- Yang, C. (2016). *The Price of Linguistic Productivity*. The MIT Press, Cambridge, MA.
- Yap, M. J., Liow, S. J. R., Jalil, S. B., and Faizal, S. S. B. (2010). The Malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior research methods*, 42:992–1003.