

# Predicting the Growth of Morphological Families from Social and Linguistic Factors

Valentin Hofmann<sup>\*‡</sup>, Janet B. Pierrehumbert<sup>†\*</sup>, Hinrich Schütze<sup>‡</sup>

<sup>\*</sup>Faculty of Linguistics, University of Oxford

<sup>†</sup>Department of Engineering Science, University of Oxford

<sup>‡</sup>Center for Information and Language Processing, LMU Munich

valentin.hofmann@ling-phil.ox.ac.uk

## Abstract

We present the first study that examines the evolution of morphological families, i.e., sets of morphologically related words such as “trump”, “antitrumpism”, and “detrumpify”, in social media. We introduce the novel task of Morphological Family Expansion Prediction (MFEP) as predicting the increase in the size of a morphological family. We create a ten-year Reddit corpus as a benchmark for MFEP and evaluate a number of baselines on this benchmark. Our experiments demonstrate very good performance on MFEP.

## 1 Introduction

Lexical change is a prime indicator of topical dynamics in social media. When people or events attract the attention of a user community, this is reflected by the *token frequency evolution of individual words*. The burst in token frequency of the word “trump” in social media before the 2016 presidential election (see Figure 1), e.g., mirrors the increasing presence of Donald Trump in public discourse during that time.

However, token frequency is only one way of measuring changes in topical prominence. Accompanying the increase in token frequency of “trump”, there was a parallel increase in the number of words morphologically related to “trump”, i.e., words like “trumpification”, “antitrumpism”, and “detrumpify” (see Figure 1, Table 1). Most of these words have a very low token frequency and are removed in the first steps of a typical NLP pipeline.

Here, we present the first study of lexical change in social media that takes as its main unit of analysis the *type frequency evolution of morphological families*, i.e., changes in the number of morphologically related words such as “trump”, “trumpification”, “antitrumpism”, and “detrumpify”. We show that morphological families allow for a fresh view

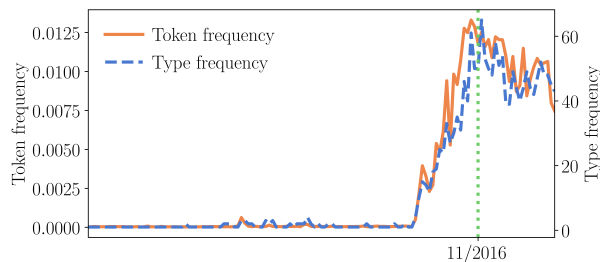


Figure 1: Token frequency of “trump” and type frequency of derivations of “trump” in the `r/politics` Subreddit between 08/2007 and 07/2018. Vertical line: date of the 2016 presidential election.

Month	Words
04/2015	“trump”
05/2015	“trump”, “trumpish”, “trumpster”, “trumpy”
06/2015	“trump”, “trumpening”, “trumper”, “trumpish”, “trumpness”, “trumpology”, “trumpster”, “trumpy”
07/2015	“trump”, “trumpening”, “trumper”, “trumpic”, “trumpification”, “trumpiness”, “trumpish”, “trumpism”, “trumpistan”, “trumpness”, “trumpster”, “trumpy”

Table 1: Derivations of “trump” in four subsequent months of the `r/politics` Subreddit.

of lexical change in social media, making them a promising tool for studies in the social sciences that draw on NLP techniques. At the same time, our work adds to the growing body of computational research on derivational morphology (Cotterell et al., 2017; Vylomova et al., 2017; Cotterell and Schütze, 2018; Deutsch et al., 2018; Pierrehumbert and Granell, 2018; Hofmann et al., 2020) by introducing a temporal perspective.

**Contributions.** We introduce the novel task of Morphological Family Expansion Prediction (MFEP), which aims at predicting whether a morphological family will increase in size or not. We publish a benchmark for MFEP and show that the growth of morphological families can be successfully modeled using social and linguistic factors

relating to the morphological parent. Furthermore, our results add a new perspective to the growing body of research on the link between cultural and linguistic change in social media.<sup>1</sup>

## 2 Morphological Families

We define a morphological *family* as a set  $F$  of words  $w$  with a shared free morpheme. Thus, “trump”, “trumpification”, “antitrumpism”, and “detrumpify” are in the same morphological family because they share the free morpheme “trump”. By contrast, “antitrumpism” and “antiprogressivism” are in different morphological families: even though both words have two morphemes in common (“anti” and “ism”), they do not belong to the same morphological family according to our definition since “anti” and “ism” are bound morphemes, not free morphemes like in the case of “trump”, “trumpification”, “antitrumpism”, and “detrumpify”. In this study, we only consider derivational morphology.<sup>2</sup> Compounds such as “trump-wall” are not split into their parts, but they can become a parent (see below). Thus, each word belongs to exactly one morphological family. The cardinality  $|F|$  of a family will be referred to as the morphological family *size*, a term also used in other studies (Schreuder and Baayen, 1997; de Jong et al., 2000; del Prado Martín et al., 2004).

The morphological *parent*  $w^*$  is the morphologically most basic word of a family  $F$ . The word “trump” is the parent of “antitrumpism”, “trumpification”, and “detrumpify”. We denote the morphological family of  $w^*$  as  $F(w^*)$ .

Except for the parent, all members of a family are morphological *children* and form a subset  $\tilde{F}$  of the entire family  $F$ . The words “antitrumpism”, “trumpification”, and “detrumpify” are all morphological children. We further distinguish between *old* children  $\tilde{F}_o$ , established words in the lexicon, and *new* children  $\tilde{F}_n$ , innovative forms. While “trumpify” can still be considered a new child of “trump”, “trumpster” is on its way to becoming an old child in the family.<sup>3</sup>

<sup>1</sup>We make all our data and code publicly available at <https://github.com/valentinhofmann/mfep>.

<sup>2</sup>The distinction between inflection and derivation is gradual, not binary (Haspelmath and Sims, 2010). The suffix “ly”, e.g., is variously defined as inflectional or derivational (Bauer, 2019). We try to exclude inflectional morphology as far as possible (e.g., by lemmatizing), but we are aware that a clear separation does not exist in linguistic reality.

<sup>3</sup>“Trumpster” is already listed in the English Wiktionary at <https://en.wiktionary.org/wiki/Trumpster>.

As the example of “trumpster” shows, morphological families are in constant flux. Specifically, there are three types of change in morphological families: word *birth*  $\emptyset \rightsquigarrow \tilde{F}_n$ , word *entrenchment*  $\tilde{F}_n \rightsquigarrow \tilde{F}_o$ , and word *death*  $\tilde{F}_n, \tilde{F}_o \rightsquigarrow \emptyset$ . The topic of this paper is word birth, i.e., we ask: given a set  $\mathcal{F}$  of morphological families, which will increase in size during a specified time interval? This differentiates our work from previous research on lexical change in social media which has focused on word entrenchment (see Section 7). One question we are particularly interested in is whether endogenous (language-internal) or exogenous (language-external) factors are better predictors of morphological family growth; these factors have been previously compared for changes in word token frequency (Altmann et al., 2011), but not for changes in morphological type frequency.

## 3 Experimental Data

We develop MFEP using Reddit, a social media platform hosting discussions about a variety of topics. Reddit is divided into smaller communities centered around a shared interest, so-called *subreddits* (SRs), which are highly conducive to linguistic innovation (del Tredici and Fernández, 2018). Concretely, we draw upon the Baumgartner Reddit Corpus, a collection of (almost) all publicly available comments posted on Reddit since 2005.<sup>4</sup> A three-year slice of this corpus was used in a study on lexical change by Stewart and Eisenstein (2018). Gaffney and Matias (2018) show that the corpus’s coverage of Reddit is not complete, but we do not expect this to affect our analysis.

Our study examines data from 2007 to 2018 in the four SRs `r/gaming`, `r/movies`, `r/nba`, and `r/politics`. These SRs were chosen because they are of comparable size, belong to the largest SRs of Reddit, and at the same time all reflect distinct areas of interest (Table 2). For each month, we also draw a random sample of comments from all SRs that will be used for computing word topicality (Section 4). The size of the sample equals the average size of the four selected SRs within the respective month.

**Preprocessing.** As in previous work (Tan and Lee, 2015), we filter posts for known bots and spammers. We remove abbreviations (“viz.”), strings containing numbers (“b4”), references to

<sup>4</sup>The corpus is available at <https://files.pushshift.io/reddit/comments/>.

SR	Start	End	$N_I$	$N_W$	$N_U$	$N_T$	$ \mathcal{F} $	$\mu_{ \mathcal{F} }$	$\sigma_{ \mathcal{F} }$
r/politics	08/2007	07/2018	19	2,970,509,554	1,519,039	2,020,843	19,000	4.64	4.05
r/gaming	09/2007	08/2018	19	1,119,096,999	2,875,931	1,859,228	19,000	4.00	3.48
r/movies	06/2008	05/2018	17	738,365,964	1,772,085	613,158	17,000	3.44	3.09
r/nba	10/2010	09/2018	13	898,483,442	486,746	721,641	13,000	3.20	2.93

Table 2: SR summary statistics. Start: first month included; end: last month included;  $N_I$ : number of intervals;  $N_W$ : number of word tokens;  $N_U$ : number of users;  $N_T$ : number of threads;  $|\mathcal{F}|$ : number of family examples ( $1,000 \times N_I$ );  $\mu_{|\mathcal{F}|}$ : mean family size per month;  $\sigma_{|\mathcal{F}|}$ : standard deviation of family size per month.

users (“u/user”) and SRs (“r/subreddit”), and both full and shortened hyperlinks. We convert British English spelling variants to American English and lemmatize all words to remove inflectional morphology. We follow Han and Baldwin (2011) in reducing repetitions of more than three letters (“ni-iiiiice”) to three letters. Except for stopwords, we do not employ a frequency threshold; in particular, we include words that occur only once.

**Computing morphological families.** Given a collection of texts  $S$ , we define the morphological families as follows. Let  $V_S$  be the vocabulary of  $S$ , i.e., all words occurring in it. We define the set of parents  $O_S \subset V_S$  as the 1,000 most frequent words in  $S$ , regardless of whether the word is decomposable or not. This means that parents are not necessarily morphological roots (Haspelmath and Sims, 2010).<sup>5</sup> We attempt to segment all other words  $w$  using affixes from a representative list of productive prefixes and suffixes in English (Crystal, 1997). We define the set  $C$  of candidate parents of  $w$  as follows. If  $w \in O_S$ , then  $C(w) = \{w\}$ . Otherwise,  $C(w) = \bigcup_{b \in B(w)} C(b)$ , where  $B(w)$  is the set of bases that remain when one of  $w$ ’s affixes is removed. For  $w^* \in O_S$ , we then define its morphological family as

$$F(w^*) = \{w \in V_S | w^* \in C(w) \wedge |C(w)| = 1\}.$$

Procedurally, families can be identified by a recursive bottom-up algorithm. The algorithm is sensitive to morpho-orthographic rules of English (Plag, 2003); e.g., when “ness” is removed from “trumpiness”, the result is “trumpy”, not “trumpi”.

<sup>5</sup>In a situation where both “sense” and “sensation”, e.g., fall above the frequency threshold, we get two separate morphological families of the parent “sense” (a root) with “non-sense”, “sensitive”, etc. and the parent “sensation” (not a root) with “sensational”, “sensationalism”, etc. (the children of “sensation” are not added to the family of “sense”). However, most morphological parents are in fact roots.

## 4 Predictors for MFEP

In a setup similar to Altmann et al. (2011), we formalize MFEP as a binary classification task. Given a *context interval*  $i^{(c)}$ , a following temporally adjacent *probe interval*  $i^{(p)}$ , and a morphological parent  $w^*$ , we ask: what properties of  $w^*$  can predict whether  $|F(w^*)|$  increases in  $i^{(p)}$  or not?

Here, we set the length of  $i^{(c)}$  to 18 months and the length of  $i^{(p)}$  to 6 months. Morphological families are computed separately for each pair of  $i^{(c)}$  and  $i^{(p)}$ . The lowest frequency count of a parent in  $i^{(c)}$  is 244. Table 2 summarizes statistics of the morphological families for each SR.

We define a number of predictors for family expansion that are measurements of properties of  $w^*$ . All predictors are motivated by work in psycholinguistics and NLP. They fall into three natural classes: (i) a *type frequency-based* predictor ( $|\overline{F}|$ ), (ii) *token frequency-based* predictors ( $f_r$ ,  $\bar{z}^{(p)}$ ), and (iii) *dissemination-based* predictors ( $D_w^U$ ,  $D_w^T$ ,  $Q_w$ ). All predictors except for  $\bar{z}^{(p)}$  (which is measured in  $i^{(c)}$  and  $i^{(p)}$ ) are measured in  $i^{(c)}$ .

**Family size  $|\overline{F}|$ .** The family size is a prime example of an endogenous (language-internal) factor, i.e., one that depends on the linguistic system. A morpheme with a large family might combine more readily with new affixes than a morpheme that occurs only with a small number of affixes. This idea bears a theoretical connection to smoothing techniques such as Witten-Bell and Kneser-Ney smoothing, which model the probability of previously unseen  $n$ -grams containing a given word ( $\approx |\tilde{F}_n|$ ) by assuming a rich-get-richer process (Manning and Schütze, 1999; Teh, 2006). It is also in line with lexical growth models based on preferential attachment (Steyvers and Tenenbaum, 2005). Intuitively, the fact that morphological children themselves can become the basis for new derivations also suggests a rich-get-richer process.

Notice that  $|\overline{F}|$  is equivalent to the type fre-

quency of  $w^*$ . In linguistics, type frequency is known to be a good predictor of the productivity of inflectional patterns (Bybee, 1995). Furthermore, it has been shown that the morphological family size facilitates lexical processing (Schreuder and Baayen, 1997). To probe whether type frequency also influences the likelihood of a family to grow, we include the predictor  $\overline{|F|}$ , morphological family size averaged over the 18 months of  $i^{(c)}$ .

**Relative token frequency of parent  $f_r$ .** Frequent words are known to be more accessible in lexical processing than rare words (Jescheniak and Levelt, 1994). Therefore, they might be more available for use in novel derivations, causing an increase in morphological family size.

**Trending behavior  $\bar{z}^{(p)}$ .** Changes in the relative frequency of a morphological parent might be indicative of concomitant changes in the morphological family size. If a word gains in popularity and becomes more frequent, this could increase the chances of new morphologically related words being created. The trending behavior is a prime example of an exogenous (language-external) factor, i.e., one that depends on non-linguistic events (e.g., a presidential election). Therefore, we measure whether the parent increases in frequency.

This is done in the following way: we calculate the z-score of the frequency distribution of the parent in the probe interval  $i^{(p)}$  relative to the frequency distribution in the context interval  $i^{(c)}$ . The mean of these z-scores is then used as a continuous variable in the model,

$$\bar{z}^{(p)} = \frac{1}{|i^{(p)}|} \sum_{j=1}^{|i^{(p)}|} z_j^{(p)} = \frac{1}{|i^{(p)}|} \sum_{j=1}^{|i^{(p)}|} \frac{x_j^{(p)} - \mu^{(c)}}{\sigma^{(c)}},$$

where  $|i^{(p)}| = 6$  is the length in months of the probe interval and  $x_j^{(p)}$  is the relative frequency of the parent in month  $j$  ( $1 \leq j \leq |i^{(p)}|$ ) of  $i^{(p)}$ ;  $\mu^{(c)}$  and  $\sigma^{(c)}$  are mean and standard deviation of relative frequency of the parent in the 18 months of the context interval  $i^{(c)}$ . The measure detects increases in frequency relative to the intrinsic variation in usage frequency of a particular word. This is necessary since some words naturally exhibit stronger short-term fluctuations, which we do not want to count as frequency bursts. Similar methods for peak detection in time series are frequently used, e.g., in Baskozos et al. (2019). We use both  $i^{(c)}$  and  $i^{(p)}$  for calculating  $\bar{z}^{(p)}$  because this captures the idea of exogenous forcing without any additional

assumptions; notice that the metric is calculated on the parent only and does not include any information about what is being predicted in MFEP, namely changes in the morphological family size.

**User dissemination  $D_w^U$ .** Following findings by Church and Gale (1995) and Altmann et al. (2011), we define user dissemination  $D_w^U$  as the extent to which the number of users of a specific word  $w$  deviates from a Poisson process,

$$D_w^U = \frac{U_w}{\tilde{U}(f_w)},$$

where  $U_w$  is the number of users who posted at least one comment including  $w$  in  $i^{(c)}$ ,  $f_w$  is the relative frequency of  $w$  in  $i^{(c)}$ , and  $\tilde{U}(f_w)$  is the expected number of users under a Poisson model given the relative frequency  $f_w$ .  $\tilde{U}(f_w)$  can be calculated as

$$\tilde{U}(f_w) = \sum_{j=1}^{N_U} \tilde{U}_j \approx \sum_{j=1}^{N_U} \left(1 - e^{-f_w m_j^U}\right),$$

where  $N_U$  is the number of users,  $\tilde{U}_j$  is the probability that the posts of user  $j$  contain  $w$  at least once, and  $m_j^U$  is the total number of words posted by user  $j$  in  $i^{(c)}$ . The approximation is valid for  $f_w \ll 1$  and  $m_j^U / \sum_{j=1}^{N_U} m_j^U \ll 1$ . Our data satisfy both requirements.

User dissemination and the following dissemination measures have a cognitive justification: it has been shown that items that are used in more diverse situations and contexts are stored in human memory in a way that makes them more retrievable (Anderson and Milson, 1989; Brysbaert et al., 2016). Thus, words with a higher dissemination are more accessible to speakers and could figure more prominently among bases for new formations. The dissemination measures fall into a gray area between exogenous and endogenous factors since they reflect the cognitive representation of language-external properties (Altmann et al., 2011).

**Thread dissemination  $D_w^T$ .** Similar to user dissemination, thread dissemination  $D_w^T$  is defined as the extent to which the number of threads containing a specific word  $w$  deviates from a Poisson process (Altmann et al., 2011),

$$D_w^T = \frac{T_w}{\tilde{T}(f_w)},$$

where  $T_w$  is the number of threads that include at least one instance of  $w$ , and  $\tilde{T}(f_w)$  is the expected



number of threads under a Poisson model.  $\tilde{T}(f_w)$  can again be calculated as

$$\tilde{T}(f_w) = \sum_{j=1}^{N_T} \tilde{T}_j \approx \sum_{j=1}^{N_T} \left(1 - e^{-f_w m_j^T}\right),$$

where  $N_T$ ,  $\tilde{T}_j$ , and  $m_j^T$  are defined analogously to  $N_U$ ,  $\tilde{U}_j$ , and  $m_j^U$ . The approximation is again valid since the data satisfy  $m_j^T / \sum_j^{N_T} m_j^T \ll 1$ .

**Topicality  $Q_w$ .** Because SRs are communities centered around interests, words that are characteristic of a SR’s topic are more frequent in that SR than in the others. Topicality has been shown to have an impact on lexical dynamics at long time scales (Church, 2000; Montemurro and Zanette, 2016). It could also influence the productivity of morphological families: higher topical dissemination, i.e., lower topicality, could facilitate growth. To capture this effect, we introduce a metric of topical distinctiveness,  $Q_w$ , which we define as

$$Q_w = \frac{f_w}{\tilde{f}_w},$$

where  $f_w$  is the relative frequency of the word  $w$  in a SR in  $i^{(c)}$ , and  $\tilde{f}_w$  is the expected relative frequency of  $w$  based on a random sample of posts from all SRs in  $i^{(c)}$ .

The polarity of  $Q_w$  is reversed to  $D_w^U$  and  $D_w^T$ , i.e., a word that is very clumped in SR space will have a *high* value of  $Q_w$ , but a word that is very clumped in user or thread space will have a *low* value of  $D_w^U$  or  $D_w^T$ , respectively.

## 5 Experimental Setup

**Finding growing families.** We use two different notions of growth for MFEP: *absolute growth* and *relative growth*. We define absolute growth as

$$\delta_a(F) = \mu_{|F|}^{(p)} - \mu_{|F|}^{(c)},$$

where  $\mu_{|F|}^{(p)}$  and  $\mu_{|F|}^{(c)}$  are the mean morphological family size in  $i^{(p)}$  and  $i^{(c)}$ , respectively. Relative growth is defined similarly as

$$\delta_r(F) = \frac{\mu_{|F|}^{(p)}}{\mu_{|F|}^{(c)}}.$$

For both  $\delta_a$  and  $\delta_r$ , we define binary features based on thresholds  $l_a$  and  $l_r$ , i.e., we define a morphological family  $F$  to be a positive example if

$\delta_a(F) > l_a$  for a pair of  $i^{(p)}$  and  $i^{(c)}$  (in the case of absolute growth). We thus train two models: one for predicting whether  $\delta_a(F) > l_a$ , one for predicting whether  $\delta_r(F) > l_r$ .

**Model.** We use Random Forests (RF) to perform the classification (Breiman, 2001). RF offers two main advantages in comparison with other models. Firstly, as opposed to other tree-based models, RFs decorrelate trees, which is important if the features are correlated (as is the case here). Secondly, the feature importance scores of a RF provide a transparent way to compare the predictive power of features. We do not use more complex albeit potentially better performing methods such as deep architectures since our primary goal is to compare various features and show that MFEP is a feasible computational task.

Since the data contain considerably more negative than positive examples, we randomly sample one negative example for every positive example for the final data. The interval pairs from all SRs were merged into one dataset, which was then split into 0.8 and 0.2 for train/dev and test sets. The train/dev set was split again into 0.8 and 0.2 for train and dev sets. Thus, all sets contain a balanced sample of interval pairs from all SRs.<sup>6</sup>

We use a total of 68,000 pairs of intervals ( $i^{(c)}$ ,  $i^{(p)}$ ) where  $i^{(c)}$  is the context interval and  $i^{(p)}$  the probe interval (see also Table 2). Recall that  $i^{(c)}$  has length 18 months and  $i^{(p)}$  6 months. Temporally adjacent interval pairs are overlapping by  $|i^{(p)}|$  months, i.e., every month in the original data is used exactly once in a probe interval and three times in a context interval.

We do not perform hyperparameter tuning and instead choose typical values for the hyperparameters of RF: 80 for the number of trees, and 20 for tree depth. For our initial MFEP models, we set the thresholds  $l_a = 2.4$  and  $l_r = 1.6$ , two values in the mid-range of existing values for  $\delta_a$  and  $\delta_r$ . We will later analyze the sensitivity to these hyperparameters in greater detail.

## 6 Results

**Overall performance.** As shown in Table 3, the RF models exhibit a good performance with an overall prediction accuracy of 80.9% for  $l_a = 2.4$  and 70.8% for  $l_r = 1.6$  (random baselines: 50.0%).

<sup>6</sup>Since the chosen SRs cover diverse topics, the model should have a high transferability to other SRs, but we have not tested this.

	$l_a (\delta_a)$							$\mu \pm \sigma$	$l_r (\delta_r)$							$\mu \pm \sigma$
	0.0	0.8	1.6	2.4	3.2	4.0	4.8		1.0	1.2	1.4	1.6	1.8	2.0	2.2	
All	.618	.696	.745	.809	.769	.844	.846	.761 $\pm$ .077	.618	.631	.651	.708	.701	.724	.744	.683 $\pm$ .045
$\overline{ F }$	.602	.674	.724	.742	.705	.754	.831	.719 $\pm$ .066	.602	.607	.629	.645	.628	.681	.744	.648 $\pm$ .046
$f_r$	.522	.542	.524	.551	.534	.574	.492	.534 $\pm$ .024	.522	.507	.509	.465	.558	.483	.605	.521 $\pm$ .044
$\bar{z}^{(p)}$	.520	.540	.523	.528	.506	.475	.631	.532 $\pm$ .045	.520	.521	.521	.555	.544	.595	.488	.535 $\pm$ .031
$D_w^U$	.506	.513	.506	.509	.466	.484	.569	.508 $\pm$ .030	.506	.499	.494	.505	.485	.569	.558	.517 $\pm$ .031
$D_w^T$	.521	.546	.508	.518	.542	.590	.431	.522 $\pm$ .045	.521	.525	.492	.490	.478	.578	.535	.517 $\pm$ .032
$Q_w$	.512	.504	.506	.516	.522	.484	.415	.494 $\pm$ .034	.512	.507	.505	.516	.496	.543	.558	.520 $\pm$ .021
$\bar{F}_n$	.557	.730	.883	.846	.909	—	—	.785 $\pm$ .129	.557	.567	.621	.618	.657	.701	.703	.632 $\pm$ .054
12 m.	.608	.668	.749	.811	.840	.838	.946	.780 $\pm$ .106	.608	.629	.684	.673	.771	.733	.912	.716 $\pm$ .095
24 m.	.630	.688	.732	.780	.797	.818	.847	.756 $\pm$ .071	.630	.645	.679	.710	.690	.748	.786	.698 $\pm$ .051

Table 3: Prediction accuracies for varying experimental setups and thresholds  $l_a, l_r$ . All: results for model trained on all predictors;  $\overline{|F|}$ : family size only;  $f_r$ : relative token frequency of parent only;  $\bar{z}^{(p)}$ : trending behavior of parent only;  $D_w^U$ : user dissemination only;  $D_w^T$ : thread dissemination only;  $Q_w$ : topicality only.  $\bar{F}_n$ : results for model trained on new children only. 12 m.: results for model trained on context intervals of 12 months; 24 m.: 24 months. Best accuracies per column are highlighted in gray, second-best accuracies in light gray.

	$l_a (\delta_a)$							$\mu \pm \sigma$	$l_r (\delta_r)$							$\mu \pm \sigma$
	0.0	0.8	1.6	2.4	3.2	4.0	4.8		1.0	1.2	1.4	1.6	1.8	2.0	2.2	
$\overline{ F }$	.202	.296	.344	.393	.458	.495	.489	.382 $\pm$ .101	.202	.208	.223	.253	.286	.261	.204	.234 $\pm$ .031
$f_r$	.153	.145	.140	.152	.127	.128	.164	.144 $\pm$ .013	.153	.153	.148	.144	.129	.122	.125	.139 $\pm$ .012
$\bar{z}^{(p)}$	.168	.150	.149	.145	.154	.136	.138	.149 $\pm$ .010	.168	.172	.178	.179	.195	.218	.264	.196 $\pm$ .032
$D_w^U$	.156	.134	.119	.098	.084	.070	.058	.103 $\pm$ .033	.156	.156	.149	.136	.130	.117	.096	.134 $\pm$ .021
$D_w^T$	.162	.139	.124	.105	.081	.090	.067	.110 $\pm$ .031	.162	.158	.150	.145	.132	.147	.145	.148 $\pm$ .009
$Q_w$	.158	.137	.125	.108	.096	.081	.083	.113 $\pm$ .027	.158	.155	.152	.143	.128	.135	.167	.148 $\pm$ .013

Table 4: Importance loadings for individual features. Features as in Table 3. Importance loadings are calculated for different values of the thresholds  $l_a$  and  $l_r$ . Highest importance loadings for features are highlighted in gray, second-highest in light gray.

The strongest predictor for both models is type frequency with a feature importance of 39.3% and 25.3%, respectively (Table 4). Models trained only on this feature already achieve accuracies of 74.2% and 64.5%, respectively (Table 3). However, the effect of  $\overline{|F|}$  is reversed: while larger morphological families have higher absolute growth values, which is in accordance with theories of lexical growth based on preferential attachment (Steyvers and Tenenbaum, 2005), smaller morphological families have higher relative growth rates. This can be explained by the observation that a large family needs much higher increases in family size to have the same relative growth rate as a small family. The fact that larger families are generally more likely to grow does not seem to counteract this imbalance.

We now analyze how different design choices impact the performance of our model.

**Thresholds  $l_a$  and  $l_r$ .** We systematically vary  $l_a$  and  $l_r$  in the range  $0.0 \leq l_a \leq 4.8$  and  $1.0 \leq l_r \leq 2.2$  to examine their influence on per-

formance.<sup>7</sup> We find that accuracies for predicting larger  $\delta_a$  and  $\delta_r$  are considerably higher than for smaller increases (Table 3). For  $l_a = 4.8$  (i.e., the family size increases by more than 4.8 members on average), the model has an error rate of 15.4%, which is less than half compared to the error rate of 38.2% for  $l_a = 0.0$ . This striking result is in line with studies on the predictability of extreme events in social media (Miotto and Altmann, 2014) and statistical physics (Hallerberg and Kantz, 2008) showing that extreme events are generally better predictable than non-extreme events.

We then train single-feature models for varying  $l_a$  and  $l_r$ . The best predictor for all values of  $l_a$  and  $l_r$  is  $\overline{|F|}$  (Table 3). The overall second-best predictor is  $\bar{z}^{(p)}$ , even though it is (sometimes only marginally) outperformed by  $f_r$  and  $D_w^T$  on several values of  $l_a$  and  $l_r$ .

To further analyze the relative importance of individual features, we examine the RF feature

<sup>7</sup>The number of positive examples with  $\delta_a > 4.8$  and  $\delta_r > 2.2$  is too small for achieving reliable results.

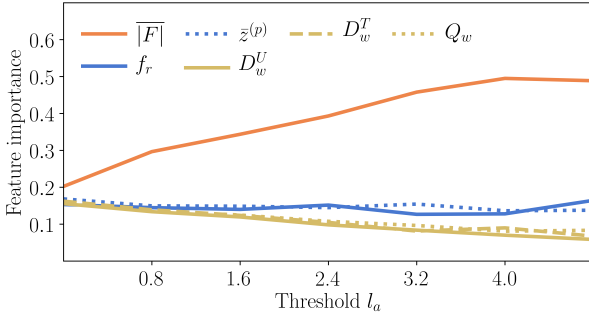


Figure 2: Feature importances of MFEP RF models ( $\delta_a$ ) for varying the absolute growth threshold  $l_a$ . The predictors are grouped into the classes *type frequency-based* ( $|F|$ ), *token frequency-based* ( $f_r$ ,  $\bar{z}^{(p)}$ ), and *dissemination-based* ( $D_w^U$ ,  $D_w^T$ ,  $Q_w$ ).

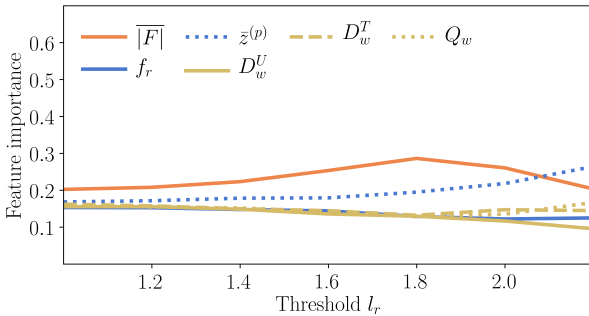


Figure 3: Feature importances of MFEP RF models ( $\delta_r$ ) for varying the relative growth threshold  $l_r$ . The predictors are grouped into the classes *type frequency-based* ( $|F|$ ), *token frequency-based* ( $f_r$ ,  $\bar{z}^{(p)}$ ), and *dissemination-based* ( $D_w^U$ ,  $D_w^T$ ,  $Q_w$ ).

importance loadings for varying  $l_a$  and  $l_r$  (Table 4, Figure 2, Figure 3). While  $|F|$  is again the best predictor overall,  $\bar{z}^{(p)}$  much more clearly comes out as the second-best predictor: especially for  $\delta_r$ , it steadily increases with higher values for  $l_r$  and even surpasses  $|F|$  for  $l_r = 2.2$ .

These results indicate that while the family size is most predictive of morphological family growth in general, high growth rates are particularly likely for families of trending parents (most of which have initially small family sizes). An example for the second case is the burst in the “trump” family illustrated in Section 1. This would explain how small morphological families can grow in the first place given the overall dominating importance of a large family size: small families need exogenous forcing (Altmann et al., 2011), i.e., external events leading to a burst in token frequency and a subsequent increase in type frequency. In order to test this hypothesis, we retrain the model for  $\delta_a$  on small families ( $1.5 \leq |F| \leq 1.6$ ), varying

	$l_a$						$\mu \pm \sigma$
	0.0	0.2	0.4	0.6	0.8	1.0	
$f_r$	.175	.187	.196	.193	.165	.190	.184±.011
$\bar{z}^{(p)}$	.222	.216	.223	.253	.269	.234	.236±.019
$D_w^U$	.183	.198	.190	.181	.179	.190	.187±.006
$D_w^T$	.212	.196	.201	.192	.195	.233	.205±.014
$Q_w$	.208	.202	.190	.181	.192	.153	.188±.018

Table 5: Importance loadings for individual features with small families ( $1.5 \leq |F| \leq 1.6$ ). Features as in Table 3. Importance loadings are calculated for different values of the thresholds  $l_a$ . Highest importance loadings for features are highlighted in gray, second-highest in light gray.

$0.0 \leq l_a \leq 1$  (which is the range of  $\delta_a$  for families of that size);  $\bar{z}^{(p)}$  has the highest feature importance for all values of  $l_a$  (Table 5).<sup>8</sup>

Furthermore, it is interesting to note that the frequency-based as well as the dissemination-based measures are considerably clumped together in feature importance space for absolute growth  $\delta_a$ , with the frequency-based predictors topping the dissemination-based ones. This is in line with recent work on the relative importance of frequency and social dissemination in lexical change (Stewart and Eisenstein, 2018). Higher values in these features correlate with a higher likelihood of growth, except for topicality: here, growth is more likely with lower topicality, which as discussed above indicates higher topical dissemination.

**Length of  $i^{(c)}$  and  $i^{(p)}$ .** In previous experiments, the length of  $i^{(c)}$  and  $i^{(p)}$  was set to 18 and 6 months, respectively. We now analyze how the choice of the interval length, specifically of the length of  $i^{(c)}$ , influences the performance of our MFEP model. We retrain the model for  $0.0 \leq l_a \leq 4.8$  and  $1.0 \leq l_r \leq 2.2$  with  $|i^{(c)}| = 12$  and  $|i^{(c)}| = 24$ , i.e., the context intervals are six months shorter and longer than previously. The length of the probe interval is kept unchanged,  $|i^{(p)}| = 6$ .

The performances of the two MFEP models is comparable with  $|i^{(c)}| = 18$  (Table 3). Both show top performance at  $l_a = 4.8$  and  $l_r = 2.2$ . However, it is interesting to note that the performance with  $|i^{(c)}| = 12$  tends to be better than  $|i^{(c)}| = 24$  with large values for  $l_a$  and  $l_r$ , but worse with smaller values, suggesting that shorter context intervals have an advantage in predicting large increases in family size while longer context intervals have an advantage in predicting smaller increases.

<sup>8</sup>Even smaller families do not allow to vary  $l_a$  on such a large range due to data sparsity, but show similar results.

**New children.** In our main study design, growth in the families is not necessarily due to *new* children being added to the family, i.e., due to an increase of  $|\tilde{F}_n|$ . A rare but established English word  $w \in \tilde{F}_o$  that only occurs a couple of times in the data counts as much to the growth as an innovative form. Here, we try to exclude fluctuations due to  $\tilde{F}_o$  by excluding all words in the data that are listed on a comprehensive list of English words encompassing over 400,000 word types, an independent estimate of established words.<sup>9</sup> Training the model on the resulting data, we find that accuracies tend to be higher for  $\delta_a$ <sup>10</sup> but lower for  $\delta_r$  than corresponding accuracies on the full dataset (Table 3). This result indicates that our model is not only capable of forecasting the evolution of the entire family but also of predicting the birth of new morphological children.

**Error analysis.** The segmentation algorithm is doomed to produce a certain number of false positives. To get a clearer picture of its accuracy, we manually examine 500 randomly selected families from one month in the data. Macro-averaged over families, 8.8% of the words are errors, i.e., they do not belong to the morphological family assigned by the algorithm. However, the error rate is not distributed evenly: only 10 of the 500 families are responsible for more than 60% of the errors.

One frequent source of erroneous segmentations is incorrect orthography. The word “representatives”, e.g., is frequently written as “reprentatives” due to its being pronounced without the consonant “t”. The algorithm then segments “reprentatives” into “re+pre+senate+ive+s” and adds it to  $\tilde{F}$  (“senate”). Another frequent case is the erroneous segmentation of emphatical repetitions of vowels, e.g., “hey” is segmented as “hey+y” and added to  $\tilde{F}$  (“hey”). Such false positives are a major source of distortion in the data.

## 7 Related work

**Morphological families and productivity.** The concept of morphological families was introduced in psycholinguistic work on lexical processing (Schreuder and Baayen, 1997; de Jong et al., 2000; del Prado Martín et al., 2004). These studies show that response latencies in lexical decision are not only influenced by *token frequency* but also by *type*

*frequency*, i.e., the size of their morphological family. Morphological families have also been used for analyzing lexical change on historical time scales (Keller and Schultz, 2013, 2014). However, this work is not comparable to our study since it relies on dictionaries, which typically exclude the transparently formed, non-entrenched words in  $\tilde{F}_n$  we are interested in (Bauer, 2001).

The main question of our study (how can we predict the growth of morphological families?) is related to a long-standing problem in traditional linguistic scholarship, i.e., what factors influence morphological productivity. The productivity of a morpheme is defined as its propensity to be used in novel combinations and traditionally understood to refer to bound morphemes (Haspelmath and Sims, 2010). Pierrehumbert and Granel (2018) highlight the fact that morphological productivity, just as morphology (Rácz et al., 2015) and other components of language (Labov, 1963) in general, is heavily influenced by social variation. Social groups differ in the morphological patterns they use and in the extent to which they extend these patterns to new words. This makes morphological productivity an exciting new area for future research in computational social science, and it further underscores the relevance of MFEP for that field.

**Derivational morphology in NLP.** Derivational morphology has received increasing attention in NLP recently. Key challenges include segmenting (Cotterell et al., 2016; Luo et al., 2017; Cotterell and Schütze, 2018), modeling the meaning (Lazaridou et al., 2013; Kisselew et al., 2015; Padó et al., 2016; Cotterell and Schütze, 2018), and predicting the form (Vylomova et al., 2017; Cotterell et al., 2017; Deutsch et al., 2018) as well as morphological well-formedness (Hofmann et al., 2020) of derivatives. Whereas all these studies approach derivational morphology from a *synchronic* standpoint, MFEP is to the best of our knowledge the first computational task that addresses *diachronic* aspects of derivation.

**Lexical change in social media.** Language change (Croft, 2000; Bybee, 2015) is most visible on the lexical level. New words like “detrumpify” attract attention, often becoming the subject of public discourse (Metcalf, 2002). Since innovations are taking place at a much faster rate on internet media (Crystal, 2004), social media have become a central resource for studies on lexical change over the last decade (Altmann et al., 2011; Garley

<sup>9</sup>The list is available on <https://github.com/dwyl/english-words>.

<sup>10</sup>We only trained the model for  $0 \leq l_a \leq 3.2$  since there was not enough data for larger threshold values.



and Hockenmaier, 2012; Danescu-Niculescu-Mizil et al., 2013; Grieve et al., 2016; Kershaw et al., 2016; Sang, 2016; Stewart and Eisenstein, 2018; del Tredici and Fernández, 2018).

One central question in this field is: what factors determine whether a word will survive in the lexicon of an online community? Usage frequency is a well-known factor that influences the evolution of a word at historical time scales (Pagel et al., 2007). Studies on lexical change in online groups have shown that this is also true for shorter time scales (Altmann et al., 2011; Stewart and Eisenstein, 2018). Another main factor is the dissemination of a word, i.e., how widely a word is spread across different social and linguistic contexts. Generally, the more disseminated a word is, the more likely it is to grow. This holds for social dissemination across users and threads (Altmann et al., 2011) as well as linguistic dissemination across different lexical collocations (Stewart and Eisenstein, 2018).

The studies mentioned so far focus on token frequency. An exciting new approach looks instead at the meaning of words using diachronic word embeddings (Hamilton et al., 2016). del Tredici et al. (2019), e.g., explore short-term meaning shifts on Reddit and identify considerable changes even within a period of eight years.

A main goal of this study is to add a third approach to studies on lexical change in social media besides word frequency and word embeddings: word families. From a linguistic point of view, these three approaches can be viewed to be complementary: whereas word frequency is *context-independent*, both word embeddings and word families reflect *context-sensitive* measures. However, while word embeddings reflect proximity on the *utterance* level (which words are close to each other in spoken sentences?), word families reflect proximity on the *system* level (which words are close to each other in the mental lexicon?).

## 8 Conclusion

In this paper, we have proposed MFEP (Morphological Family Expansion Prediction), a new task that aims at predicting how morphological families evolve over time. We have shown that changes in morphological family size provide a fresh look at topical dynamics in social media, thus complementing token frequency as a metric.

Furthermore, we have presented a random forest model for MFEP that achieves very good accuracy,

particularly in predicting extreme growth in morphological family size. The strongest predictor of growth is the morphological family size itself, an endogenous factor. However, the initial growth of small families is mainly driven by the trending behavior of the parent, an exogenous factor. This reflection of external events makes morphological families a promising tools for various fields drawing upon NLP techniques for tracing temporal dynamics in text (e.g., virality detection).

Overall, we see our study as an exciting step in the direction of bringing together computational social science and derivational morphology. In future work, we intend to further fine-tune our methodological apparatus for tackling MFEP.

## Acknowledgements

Valentin Hofmann was funded by the Arts and Humanities Research Council and the German Academic Scholarship Foundation. This research was also supported by the European Research Council (Grant No. 740516). We thank the reviewers for their detailed and helpful comments.

## References

- Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a determinant of word fate in online groups. *PLoS ONE*, 6(5).
- John R. Anderson and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719.
- Georgios Baskozos, John M. Dawes, Jean S. Austin, Ana Antunes-Martins, Lucy McDermott, Alex J. Clark, Teodora Trendafilova, Jon G. Lees, Stephen B. McMahon, Jeffrey S. Mogil, Christine Orengo, and David L. Bennett. 2019. Comprehensive analysis of long noncoding rna expression in dorsal root ganglion reveals cell-type specificity and dysregulation after nerve injury. *Pain*, 160(2):463–485.
- Laurie Bauer. 2001. *Morphological productivity*. Cambridge University Press, Cambridge, UK.
- Laurie Bauer. 2019. *Rethinking morphology*. Edinburgh University Press, Edinburgh, UK.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.
- Marc Brysbaert, Michaël Stevens, Paweł Mander, and Emmanuel Keuleers. 2016. The impact of word prevalence on lexical decision times: Evidence from the dutch lexicon project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42(3):441–458.

- Joan Bybee. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(425-455).
- Joan Bybee. 2015. *Language change*. Cambridge University Press, Cambridge, UK.
- Kenneth Church. 2000. Empirical estimates of adaptation: The chance of two noriegas is closer to  $p/2$  than  $p^2$ . In *Conference on Computational Linguistics (COLING) 18*, pages 180–186.
- Kenneth W. Church and William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.
- Ryan Cotterell and Hinrich Schütze. 2018. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 6:33–48.
- Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A joint model of orthography and morphological segmentation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 15*.
- Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm completion for derivational morphology. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2017*.
- William Croft. 2000. *Explaining language change: An evolutionary approach*. Pearson, Harlow, UK.
- David Crystal. 1997. *The Cambridge encyclopedia of the English language*. Cambridge University Press, Cambridge, UK.
- David Crystal. 2004. *Language and the internet*. Cambridge University Press, Cambridge, UK.
- Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *The Web Conference (WWW) 22*.
- Fermín del Prado Martín, Raymond Bertram, Tuomo Häikiö, Robert Schreuder, and R. Harald Baayen. 2004. Morphological family size in a morphologically rich language: The case of finnish compared with dutch and hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1271–1278.
- Marco del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *International Conference on Computational Linguistics (COLING) 27*.
- Marco del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 17*.
- Daniel Deutsch, John Hewitt, and Dan Roth. 2018. A distributional and orthographic aggregation model for english derivational morphology. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*.
- Devin Gaffney and J. Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PLoS ONE*, 13(7):e0200162.
- Matt Garley and Julia Hockenmaier. 2012. Beefmoves: Dissemination, diversity, and dynamics of english borrowings in a german hip hop forum. In *Annual Meeting of the Association for Computational Linguistics (ACL) 50*.
- Jack Grieve, Andrea Nini, and Diansheng Guo. 2016. Analyzing lexical emergence in modern american english online. *English Language and Linguistics*, 21(1):99–127.
- Sarah Hallerberg and Holger Kantz. 2008. Influence of the event magnitude on the predictability of an extreme event. *Physical Review E*, 77(1):011108.
- William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Annual Meeting of the Association for Computational Linguistics (ACL) 54*.
- Bo Han and Timothy Baldwin. 2011. Lexical normalization of short text messages: Makn sens a #twitter. In *Annual Meeting of the Association for Computational Linguistics (ACL) 49*.
- Martin Haspelmath and Andrea D. Sims. 2010. *Understanding morphology*. Routledge, New York, NY.
- Valentin Hofmann, Hinrich Schütze, and Janet B. Pierrehumbert. 2020. A graph auto-encoder model of derivational morphology. In *Annual Meeting of the Association for Computational Linguistics (ACL) 58*.
- Jörg D. Jescheniak and Willem J. Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20(4):824–843.
- Nivja de Jong, Robert Schreuder, and R. Harald Baayen. 2000. The morphological family size effect and morphology. *Language and Cognitive Processes*, 15(4/5):329–365.
- Daniela B. Keller and Jörg Schultz. 2013. Connectivity, not frequency, determines the fate of a morpheme. *PLoS ONE*, 8(7):e69945.

- Daniela B. Keller and Jörg Schultz. 2014. Word formation is aware of morpheme family size. *PLoS ONE*, 9(4):e93978.
- Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *International Conference on Web Search and Data Mining (WSDM)* 9.
- Max Kisselew, Sebastian Padó, Alexis Palmer, and Jan Šnajder. 2015. Obtaining a better understanding of distributional models of german derivational morphology. In *International Conference on Computational Semantics (IWCS)* 11.
- William Labov. 1963. The social motivation of a sound change. *Word*, 19(3):273–309.
- Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Annual Meeting of the Association for Computational Linguistics (ACL)* 51.
- Jiaming Luo, Karthik Narasimhan, and Regina Barzilay. 2017. Unsupervised learning of morphological forests. *Transactions of the Association for Computational Linguistics*, 5:353–364.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA.
- Allan Metcalf. 2002. *Predicting new words: The secret of their success*. Houghton Mifflin Company, Boston, MA.
- José M. Miotto and Eduardo G. Altmann. 2014. Predictability of extreme events in social media. *PLoS ONE*, 9(11):e111506.
- Marcelo A. Montemurro and Damián H. Zanette. 2016. Coherent oscillations in word-use data from 1700 to 2008. *Palgrave Communications*, 2(1):405.
- Sebastian Padó, Aurélie Herbelot, Max Kisselew, and Jan Šnajder. 2016. Predictability of distributional semantics in derivational word formation. In *International Conference on Computational Linguistics (COLING)* 26.
- Mark Pagel, Quentin Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449(7163):717–720.
- Janet Pierrehumbert and Ramon Granel. 2018. On hapax legomena and morphological productivity. In *Workshop on Computational Research in Phonetics, Phonology, and Morphology (SIGMORPHON)* 15.
- Ingo Plag. 2003. *Word-formation in English*. Cambridge University Press, Cambridge, UK.
- Péter Rácz, Janet Pierrehumbert, Jennifer Hay, and Viktória Papp. 2015. Morphological emergence. In Brian MacWhinney and William O’Grady, editors, *The handbook of language emergence*, pages 123–146. Wiley, Hoboken, NJ.
- Erik T. Sang. 2016. Finding rising and falling words. In *Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*.
- Robert Schreuder and R. Harald Baayen. 1997. How complex simplex words can be. *Journal of Memory and Language*, 37:118–139.
- Ian Stewart and Jacob Eisenstein. 2018. Making “fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2018.
- Mark Steyvers and Joshua Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.
- Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *International Conference on World Wide Web (WWW)* 24.
- Yee W. Teh. 2006. A hierarchical bayesian language model based on pitman-yor processes. In *Annual Meeting of the Association for Computational Linguistics (ACL)* 44.
- Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, and Trevor Cohn. 2017. Context-aware prediction of derivational word-forms. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)* 15.