Article

R. Harald Baayen*

The wompom

https://doi.org/10.1515/cllt-2024-0053
Received May 7, 2024; accepted May 13, 2024; published online October 25, 2024

Keywords: corpus linguistics; linguistic theory; large language models; computational modeling

> You can do such a lot with a wompom,
> You can use every part of it too.
> For work or for pleasure, it's a triumph it's a treasure
> Oh there's nothing that a wompom cannot do.
> **Flanders and Swan**
> https://www.youtube.com/watch?v=CsaNgsoQcO4

1 Introduction

In this essay, I reflect on the past, present, and future of corpus linguistics. My own research has focused primarily on morphology, the mental lexicon, and lexical processing. A consequence of this narrow, myopic focus on just a tiny part of the rich phenomenon of language is that many important and inspiring developments in corpus linguistics have remained completely out of my sight. I beg the reader's forgiveness for a narrative that is fragmentary, selective, and anchored in my own research.

In recent years, large language models (LLMs) have sprung into existence. LLMs offer unprecedented opportunities and unprecedented dangers for both the study of language and for society. Paraphrasing a line from the wonderful song by Flanders and Swan about an imaginary creature with infinite uses, the wompom, it seems that "there is nothing that an LLM cannot do". What are the implications of a talking wompom for corpus linguistics?

***Corresponding author: R. Harald Baayen**, Department of Linguistics, University of Tübingen, Tuebingen, Germany, E-mail: harald.baayen@uni-tuebingen.de

Open Access. © 2024 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

In what follows, Section 1 looks back at the developments in corpus linguistics that unfolded over the last decades, and that influenced my thinking and research. Section 2 then present two case studies illustrating the opportunities and challenges of the tools and resources that are now available for understanding the lexicon and lexical processing. Section 3 looks forward, and reflects on the huge challenges for the immediate future that I believe corpus linguistics will have to address.[1]

2 The past

2.1 Resources and computational infrastructure

Looking back over the past decades, there have been enormous developments in computational infrastructure and the written and spoken resources that have become available. In the eighties, a megabyte was an enormous amount of memory. When I was working on my PhD thesis in the eighties, I was advised to travel to Nijmegen, where the CELEX lexical database was being developed, in order to extract morphologically analysed word frequency lists from this database. At that time, I was investigating whether the Good-Turing probability for unseen types (Good 1953) would be a reasonable statistic for evaluating morphological productivity. Since this probability is estimated by the ratio of hapax legomena and the corpus size, I was particularly interested in the words with frequency 1 in CELEX, which typically constitute roughly half of the types. To my disappointment, it turned out that CELEX only included a handful of hapax legomena. One of the reasons turned out to be that adding hapax legomena to the database would have led to a doubling of the size of the database, and to doubling the size of external memory required. As in the early eighties a hard disk storing 1 megabyte had a price tag of around 1 million US$, adding hapax legomena to the database was prohibitively expensive. Fortunately, the costs of storing devices have plummeted while the amounts of data that can be stored have increased dramatically. In parallel, the resources available for corpus linguistic research have also increased exponentially.

Given the abundant resources we now have, the work done by early pioneers of corpus linguistics such as Zipf, Yule, and Ellegåard, which was done by hand (Ellegård 1953; Yule 1944; Zipf 1949), is all the more impressive. My earliest experiences

[1] I am grateful to Maria Heitmeier, Sebastian Otte, Yu-Xiang Tseng, and Stefanie Wulff for their detailed feedback on earlier versions of this study, and to Dagmar Divjak, Petar Milin, Shu-Kai Hsie, Yu-Ying Chuang, Konstantin Sering, Melanie Bell, Karen Beaman, Susanne Gahl, Arvi Tavast, and Kristian Berg for essential and inspiring discussions. The author is indebted to support from the ERC, project SUBLIMINAL-101054902.

with corpus linguistics involved the study of aspect in Biblical Hebrew (Baayen 1997). The instrument that I used to collect the data for this study was a keyword in context concordance in book format (Mandelkern 1896), a monument of Jewish corpus linguistics long before the advent of computers. (Now, I would use a different tool, the computerized and meticulously linguistically annotated electronic version of the Hebrew Bible https://shebanq.ancient-data.org/hebrew/text.) When I started working on my PhD thesis in the mid-eighties, my initial insights were gleaned from a perusal of a book with word frequencies compiled by computer (probably using punch cards) from a Dutch corpus of some 600,000 words (Uit den Boogaart 1975). This corpus was modeled after the Brown corpus, which had 1 million words, and for which a book with word frequencies had been published almost a decade earlier (Kučera and Francis 1967). Fortunately, the frequency tables of Uit den Boogaart (1975) included information on all the hapax legomena. Soon after, I found myself learning UNIX, programming in C, and scripting in awk, on a small VAX computer that was in constant use by some 15 people simultaneously. But this computer gave me access to the online version of the Dutch corpus.

Since then, developments have been rapid. In the nineties, the British National Corpus (Burnard 1995) was created, comprising 100 million words. A decade later, the WaCky corpora (Baroni et al. 2009) extended the size of written corpora to more than 1 billion words. The Common Crawl February/March 2024 archive brings together no less than 3.16 billion web pages.[2]

Over the years, resources for spoken language have also been increasing rapidly. The switchboard telephone conversation corpus comprises some 260 h of recorded and transcribed speech (Godfrey et al. 1992). The Buckeye corpus (Pitt et al. 2005) comprises 40 one-hour face-to-face interviews with a stratified sample of speakers from Columbus, Ohio, with audio, orthographic transcriptions, and broad phonetic transcriptions. Inspired by the British National Corpus, a corpus of spoken Dutch was constructed and completed in 2004, with some 900 h of speech, transcribed at various levels http://hdl.handle.net/10032/tm-a2-k6 (Oostdijk et al. 2002). A much smaller corpus of spontaneous speech in German was created by Arnold and Tomaschek (2016), but this corpus provides, for 6.5 h of speech from 13 speakers, not only the audio and transcriptions, but also recordings of tongue movements traced with electromagnetic articulography. A truly vast multimedia corpus is being compiled by the Little Red Hen Consortium (https://www.redhenlab.org/home), which provides videos, audio, and aligned transcription for TV news broadcasts, mainly from the USA, but with increasing contributions from other languages as well. The NewsScape

[2] https://www.commoncrawl.org/blog/february-march-2024-crawl-archive-now-available.

infrastructure developed by Uhrig (2018) provides access to some 350,000 h of broadcasts, comprising some 2 billion words. Pipelines are currently in development for adding detailed information of movements of the head, torso, arms, and hands in the videos, enabling large scale study of co-speech gesture, and the consequences of gesture for the realization of speech in natural settings (see, e.g., Alcaraz Carrión et al. 2020).

2.2 Methods

The development of tools for the analysis of corpora has also been impressive. With respect to specifically statistical methods, the logistic regression modeling framework for variable rule analysis introduced in the early seventies (see, e.g., Cedergren and Sankoff 1974) gave sociolinguistics an impressive head start compared to other subdisciplines of linguistics. Extensions within the general approach of linear mixed models (Pinheiro and Bates 2000) and the generalized additive (mixed) model (GAM, Wood 2017) provide the analyst not only with tools for studying count data, but also with tools for evaluating measurements made on corpora, such as the durations in speech of segments or words (see, e.g., Gahl 2008; Plag et al. 2017). GAMs provide an especially attractive analytical tool for the analysis of non-linear trends in time, or non-linear properties of linguistic units in spoken corpora, such as F_0 contours for stress (Koesling et al. 2012) and tone (Chuang et al. 2021, 2024).

Multivariate methods such as principal components analysis, multidimensional scaling, linear discriminant analysis, and more recently t-distributed stochastic neighbor embedding (tSNE, Maaten and Hinton 2008) have become part of the corpus linguist's toolkit. My first encounter with this group of methods goes back to an ACH-ALLC conference at which John Burrows presented his stylometric analyses of authorial hands and language change (Burrows 1986, 1992, 1993). As a young post-doc, I was stunned by the analytical power of the relative frequencies of the most frequent function words (just 40 function words will already get you a long way), when analysed using principal components analysis.

About a decade later, researchers started moving away from monocausal explanations to multifactorial explanations. In cognitive linguistics and corpus linguistics, Stefan Gries and Dagmar Divjak developed behavioral profiling (Arppe 2008; Arppe and Järvikivi 2007; Divjak 2004; Divjak and Gries 2006; Gries 2000; Gries 2006; Gries 2010; Gries and Divjak 2009; Levshina 2022). But also in subsequent more formal approaches to language spearheaded by Joan Bresnan (Bresnan 2006; Bresnan et al. 2007), multifactorial explanations gained traction. Across sociolinguistics, cognitive linguistics, and corpus linguistics, a variety of methods for count data is now available to predict which of a possible set of alternative language variants is most likely to be used in a given context. These methods range from statistical methods such as mixed effects logistic regression and support vector machines to methods from machine learning such as random forests, gradient boosting machines, and memory-based learning (Daelemans and Van den Bosch 2005).

At the time that logistic models and variable rule analysis were making their way into linguistics, psychologists were studying learning rules for very simple neural networks (Rescorla and Wagner 1972). The Parallel Distributed Processing books (McClelland and Rumelhart 1986; Rumelhart and McClelland 1986) were published at the same time that the idea of network models was introduced to morphology by Bybee (1985, 1988). In the eighties, seminal algorithms and network architectures such as back-propagation Rumelhart, Hinton, and Williams (1986) and recurrent networks (Jordan 1986) were already on the table. Recurrent networks were developed further by Elman (1990), and not much later enriched with other kinds of neurons (Hochreiter and Schmidhuber 1997) that made these networks even more powerful. More recent transformer technology (Vaswani et al. 2017) has been applied to huge datasets such as the Common Crawl, resulting in large language models (LLM) such as the current GPT series initiated by Open-AI (Radford et al. 2018). There is hardly a natural language processing task left that an LLM, supplemented with some fine-tuning, cannot do. According to Binz and Schulz (2023a, 2023b), with some fine-tuning pre-trained LLMs can also be turned into generalist cognitive models. It seems that there is nothing that an LLM cannot be made to do.

2.3 Linguistic theory

Looking back at the development of linguistic theory over the last decades, one constant, across widely different linguistic theories, has been that explanations have been relying on symbolic descriptors that partition complex continuous and multifaceted phenomena into discrete subsets with superordinate labels such as phoneme, morpheme, word category, aspect, tense, and voice. According to early proposals, both in post-Bloomfieldian American structuralism and in generative grammar, these symbolic descriptors are part of a calculus that governs how symbolic units are arranged (Blevins 2016). However, over time, it was recognized that the constructions in a language provide islands of systematicity in a variegated whole (Croft 2001; Goldberg 2005). Monofactorial explanations were replaced by multifactorial explanations, and researchers in NLP started replacing deterministic grammars by probabilistic grammars.

Making sense of language as a variegated, probabilistic system that is subject to many simultaneous pressures and constraints is not straightforward. It is all fine to have models predicting the probabilities of alternative realizations of a given feature. Social stratification of language features can be dealt with relatively straightforwardly by positing different probabilistic grammars for dialects, sociolects and idiolects. But variability within one language variety remains enigmatic. If a speaker selects a less-probable variant, is this the inevitable consequence of the randomness in the firing patterns of biological neurons? Or, more along the lines of current next-word prediction based large language models, is every choice made ultimately predictable from a deterministic chain of previous experiences?

Exemplar models and analogical models provide a different kind of explanation. These models propose that generalization and selection in language is driven by analogical reasoning over exemplars stored in memory. Examplars can range from structured chunks of words (Bod 1998, 2006; Borensztajn et al. 2009) to structured sequences of phones (Daelemans et al. 1995; Skousen 1989). Analogical reasoning based on the most similar structured exemplars in memory then accounts for generalization. As more frequent phenomena will have more exemplars in memory, effects of frequency and probability are straightforwardly explained by selecting that class that is supported by most exemplars. The changing rates at which phenomena are encountered in individual language use, and hence the changing composition of analogical sets of exemplars, then explain within-individual fluctuations in choice behavior.

But exemplar models face two challenges. One challenge is that feature engineering is required. Some metric is required on the basis of which it can be decided how similar exemplars are. Often, exemplars are described by means of arrays of discrete feature-value pairs. This may not be optimal, a point to which I return below.[3]

The other challenge is the overwhelming numbers of exemplars. The challenges that come with storing and evaluating huge numbers of exemplars is especially prominent for spoken language, and becomes visible already at the word level. Although an exemplar model for auditory word comprehension was developed in the late nineties (Johnson 1997), network-based models for auditory comprehension were considered already at the same time (Gaskell and Marslen-Wilson 1998), and have been explored subsequently from many different theoretical perspectives (Magnuson et al. 2020; Scharenborg 2008; Shafaei-Bajestan et al. 2023). Simple network models operating on discrete featural representations, using the learning

[3] Radical exemplar models as conceptualized by Ambridge (2020b) argue against stored abstractions. However, without clear proposals as to how exemplar similarity is to be assessed, this seems not very helpful. The idea that exemplars would be ‘re-represented’ at multiple levels of abstractions in deep learning networks (Ambridge 2020a) is for me incompatible with the idea that exemplars would be discrete representations in memory, that they in some way would ‘exist’.

rule of Rescorla and Wagner (1972) are also available for exploring both lexical processing (Baayen et al. 2011; Milin et al. 2016) as well as higher levels of language structure (see, e.g., Romain et al. 2022).

For many years, I believed that network models avoid the problems faced by exemplar models. But with the advent of large language models, I'm not so sure anymore. These models work with billions of parameters (175 billion for GPT-3, more than a trillion for more recent AI systems). Some engineers now describe language modelling as a form of compression (Delétang et al. 2023). From their perspective, LLMs are compressed re-representations of the data they were trained on. Importantly, compression results in a memory that is productive and that makes accurate predictions for unseen data. From this perspective, Chat-GPT3 is a corpus that talks back.

But this suggests to me that there is considerable continuity with exemplar-based models. Daelemans et al. (1999) observed, using memory-based learning, that forgetting is harmful: performance drops when low-frequency words are not taken into consideration. Furthermore, in order to optimize algorithms for efficient exemplar-driven generalization such as TiMBL (Daelemans et al. 2007), they also found that it is essential to make use of data compression.

The analogical sets of nearest neighbors on which the models of Skousen (1989) and Daelemans et al. (1995) base generalization offer the analyst considerable insight. Analogical modeling has successfully and insightfully been applied to a range of phonomena, ranging from the study of stress (Arndt-Lappe 2011; Eddington 2000) and final devoicing (Ernestus and Baayen 2003) to the study of derivational semantics (Plag et al. 2023).

By contrast, a current LLM is more like a black box. In fact, the LLMs put the connectionist debate that raged for two decades in a new light. This debate came to an inconclusive end in 2002 with a series of exchanges in *Trends in Cognitive Sciences* (McClelland and Patterson 2002a, 2002b, 2002c; Pinker and Ullman 2002). This debate left many researchers with the feeling that connectionist networks are just not good enough. However, history has proven the connectionists right. Neural network models can be made to work, and a well-designed connectionist memory is productive. Unfortunately, LLMs are not straightforwardly interpretable. In this respect, history has proven the connectionists wrong. For instance, Seidenberg and Gonnerman (2000) viewed the hidden layer of their networks as a layer of morphological representation, where orthographical, semantic, and phonological systematicities (in their terminology, 'codes') converge. According to this proposal, just as there are distributed representations of form and meaning, there is a distributed representation of morphology which captures regularities attributed to morphemes without having to represent morphemes with discrete units. However, the many layers of units and the internals of attention heads in current transformer-based LLMs are opaque to any such interpretation.

3 The present

For corpus linguistics and corpus-informed linguistic theory, the importance of statistical learning is unsurprising. However, with the advent of LLMs, we now have computational proof that human language skills can be very closely approximated by statistical learning algorithms that work extremely well without depending on any traditional linguistic concepts such as morphemes, tense, aspect, case, number, voice, relative clause, syntactic dependencies, and constructions.

The excellent performance of LLMs suggests to me that abstracting away from micro-variation by means of symbolic features and symbolic rules operating with these features is harmful, at least when huge volumes of data are available for training with immensely large and deep networks. Perhaps this should be unsurprising. Abstraction results in categories that ideally are maximally distinct and as a consequence informative for main trend prediction with some classifier, but within-category differences become invisible and can no longer fine-tune prediction. LLMs therefore challenge current practice in corpus linguistics, cognitive linguistics, and construction grammar, of analyzing and describing tokens of language use by means of featural descriptions distinguishing between discrete values, such as employed in behavioral profiling. To be clear, I believe behavioral profiling is an important and informative tool that has led to many important insights. But LLMs challenge us to do even better.

Working with abstract features may also be sub-optimal for a very different reason. LLMs absorb not only language structure (whatever that is), but also huge chunks of world knowledge. In part, this happens because LLMs are currently developed not only on textual data, but also on digital materials containing programming code, equations from logic, mathematics, physics, and chemistry. In part this is also due to including wikipedia among the training materials.[4]

The implication for linguistic theory is that it is counterproductive to assume that knowledge of language and world knowledge are completely separated and disjunct. When in the early nineties I was working as a post-doc with Rochelle Lieber on lexical semantics, she argued that language provides the ‘skeleton’ for morphology, a skeleton that is featural and hierarchical in nature, and that encyclopedic knowledge is a holistic body that is grafted onto this skeleton (Lieber 2004). At the

[4] It appears there are hardly any benefits of visual grounding for contextualized embeddings calculated from very large text corpora, whereas benefits are visible when training proceeds on substantially smaller corpora (Shahmohammadi et al. 2023).

time, this distinction made a lot of sense to me, not in the least for practical reasons. But with the new technologies that are now available, we are in a much better position to study the interaction of knowledge of language and knowledge of the world.

In what follows, I present two case studies that illustrate the above thoughts. The first addresses the importance for linguistic theory of taking world knowledge into account. In this case study, which focuses on nominal number in English, I make use of standard embeddings (obtained with word2vec), which I take to be word-type specific ‘continuous’ (instead of discrete) behavioral profiles. The second case study uses contextualized embeddings from a large language model (GPT-2) to predict the English determiners (*the, a, an*). It illustrates the problems that come with the use of abstract features, and at the same time shows how an LLM can be useful as a tool for corpus linguistics.

3.1 Oak leaves

In standard approaches in theoretical linguistics to the formation of noun plurals in English, an operation on form (the addition of a dental fricative with phonological adjustments) has a parallel operation on meaning: the change from singular number to plural number. Typically, the semantics of pluralization is described as involving the change in value of an abstract feature for number, either by adding a feature specification for plurality, or replacing a singular feature value by a plural feature value. Although the semantics of pluralization in English is well known to be more interesting than this (see, e.g., Quirk et al. 1985), in distributional semantics, pluralization has been conceptualized as a unitary semantic operation. For instance, taking pluralization to involve an equipollent opposition, we can set up proportional analogies such as

$$\begin{matrix} \textit{apple} \\ [\text{SING}] \end{matrix} : \begin{matrix} \textit{apples} \\ [\text{PLUR}] \end{matrix} = \begin{matrix} \textit{cow} \\ [\text{SING}] \end{matrix} : \begin{matrix} \textit{cows} \\ [\text{PLUR}] \end{matrix}$$

and rewrite this in vector notation as follows:

$$\overrightarrow{\textit{apple}} : \overrightarrow{\textit{apples}} = \overrightarrow{\textit{cow}} : \overrightarrow{\textit{cows}}.$$

We next express the plurals as a function of their singulars,

$$\overrightarrow{\text{APPLES}} = \overrightarrow{\text{APPLE}} + \overrightarrow{\text{PLURAL}}$$

$$\overrightarrow{\text{COWS}} = \overrightarrow{\text{COW}} + \overrightarrow{\text{PLURAL}},$$

and it now follows that

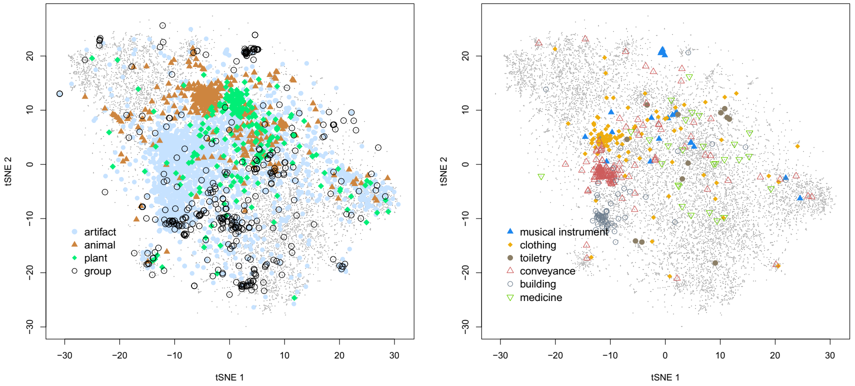

Figure 1: Shift vectors for 11,665 English nouns, using word2vec embeddings. Left panel: selected high-level WordNet categories. Right panel: WordNet-based subcategories of the higher-level category of artifacts. Individual shift vectors are represented by tiny dots, shift vectors of selected categories are represented by larger symbols.

$$\overrightarrow{\text{PLURAL}}_{\text{apple}} = \overrightarrow{\text{APPLES}} - \overrightarrow{\text{APPLE}}$$

$$\overrightarrow{\text{PLURAL}}_{\text{cow}} = \overrightarrow{\text{COWS}} - \overrightarrow{\text{COW}}.$$

Here, I specified separate plural vectors for *apple* and *cow*, one reason being that word embeddings inevitably contain measurement noise. In what follows, I will refer to the vectors $\overrightarrow{\text{PLURAL}}_{\text{apple}}$ and $\overrightarrow{\text{PLURAL}}_{\text{cow}}$ as the shift vectors for *apple* and *cow*. Given the measurement noise in the embeddings, we average over all shift vectors to obtain an average shift vector calculated over all n pairs of singulars and their corresponding plurals in a given dataset (see, e.g., Drozd et al. 2016):

$$\overrightarrow{\text{AVG SHIFT}} = \frac{1}{n}\sum_{i=1}^{n}(\overrightarrow{\text{plural}_i} - \overrightarrow{\text{singular}_i}).$$

However, as pointed out by Shafaei-Bajestan et al. (2022, 2024), upon closer inspection, the empirical shift vector for English is not very informative. To see this, consider the left panel of Figure 1, which plots the shift vectors (calculated from word2vec embeddings, Mikolov et al. 2013) for 11,665 English nouns in a two-dimensional plane, obtained by applying a tSNE dimension reduction (Maaten and Hinton 2008). The average shift vector (not shown) is located near the origin of the scatterplot. Shift vectors show considerable clustering by high-level WordNet-based semantic class, four of which are highlighted (artifacts in gray, animals in brown, plants in green, and group nouns in black). A straightforward linear discriminant analysis with leave-one-out cross-validation is able to predict the high-level semantic class (N = 26) for a shift vector with 63 % accuracy (majority baseline: 23 %).

The right panel of Figure 1 shows that within a large category such as that of artifacts, words belonging to further subclasses show additional clustering. For more fine-grained WordNet-based semantic classes, classification accuracy increases even further (see Shafaei-Bajestan et al. 2024, for detailed discussion). The emergence of clusters in just a two-dimensional t-SNE plane indicates that English shift vectors are well-separated in a relatively low-dimensional subspace of the embedding space (cf. Stupak and Baayen 2022).[5]

What are the theoretical implications of the clustering of English shift vectors by semantic class? It might be argued that we are observing performance noise, and that English pluralization is at its core a symbolic operation – the adding or exchanging of a feature value. However, as pointed out by Shafaei-Bajestan et al. (2024), languages such as Swahili (Polomé 1967) and Kiowa (Harbour 2008) have elaborate gender systems with exponents that vary with a set of high-level semantic classes. Given the present observations for English, it is conceivable that Swahili and Kiowa have grammaticalized parts of the differential semantics of nominal pluralization, whereas English has not.

Of course, one's response to this observation could be: so what? On the other hand, what if the observed clustering of shift vectors is actually part of English speakers' knowledge of their language? If so, there are non-trivial implications for our understanding of inflectional productivity. Rules depend on simplicity. It is this simplicity that enables generalization. If English nouns have plurals that are to some extent class-specific and even item-specific, then the implication is that predicting the precise meaning of a noun plural that has not been encountered before, given a singular that has been encountered before, is not possible.

Let me illustrate this simple point with an example. Consider how listeners might interpret the plural compound *oak leaves*. Listeners who do not know what an oak tree is will think of the leaves of some plant. Listeners who know that an oak is some kind of tree, will understand that the compound references the leaves of some tree that goes by the name of oak. Those listeners who can distinguish oak trees from willow trees and ash trees, will recognize that leaves such as depicted in Figure 2 are the leaves of oak trees. But European listeners will primarily think of the leaves of the European oak (left panel), whereas listeners in North America will first think of the leaves of the American oak (right panel). What the precise meaning of an inflected word is varies across speakers and listeners, and depends on their knowledge of the world.

One might argue that this variability in understanding does nothing more than clarify that all that a grammatical description can reasonably accomplish is to

[5] Plural shift vectors in Russian, and Finnish Chuang et al. (2023), Nikolaev et al. (2023) have also been found to form clusters, but these clusters are based on case, rather than semantic class.

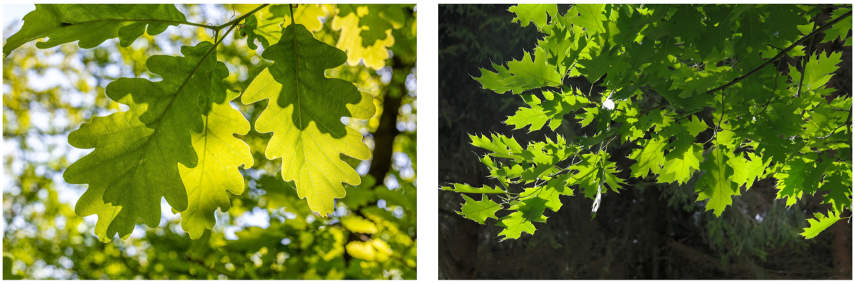

Figure 2: Leaves of the European oak (left) and the American oak (right). Source: Pixabay.com.

associate plurality with a single general operation, such as an average shift vector in distributional semantics, or a change in a symbolic feature an in theoretical morphology. Undoubtedly, this keeps things simple, but it has an important downside. If embeddings capture important aspects of meaning, then linguistic theories that work with a single fixed featural change for nominal plurality give up completely on making precise predictions for the actual meanings of plural nouns.

Alternatively, we can embrace embeddings, and accept the challenge to think of how we might predict words' meanings from their forms as well as possible. Before addressing this question, a cautionary note on embeddings is in order. The knowledge base underlying embeddings is super-human in many ways, both in terms of volumes of text and the wide coverage of topics. Embeddings cannot be expected to be precise for any single language user. Nevertheless, they are useful, especially as similarity scores of embeddings have been found to be precise predictors for aggregated human perceived semantic similarity (Boleda 2020). I look at them as crutches to walk with.

If we accept embeddings as legitimate and useful approximators of words' meanings, then the question arises how close a listener can get to a word's embedding given its form as input. This question can be addressed with a computational model of the mental lexicon that we have been developing in Tübingen, the discriminative lexicon model (DLM, Baayen et al. 2019; Chuang and Baayen 2021; Heitmeier et al. 2024).

The DLM is a simple model for studying the relation between words' forms and their meanings. Both forms and meanings are represented by high-dimensional numeric vectors.[6] The task that the model is designed for is to predict meanings from

[6] However, in the DLM, forms and meanings do not have representations stored in memory. In comprehension, for instance, form vectors represent ephemeral visual or auditory input that is dynamically mapped on the fly onto equally ephemeral semantic representations (see Gahl and Baayen 2024, for detailed discussion).

forms (comprehension), and forms from meanings (production). This is accomplished with the help of networks that map form and meaning onto each other. The model's simpler and computationally light mappings make use of the linear algebra that underlies multivariate multiple regression, its more complex and computationally intensive mappings make use of deep learning networks. In what follows, I illustrate what can be achieved with the computationally light linear mappings.

An implementation of a DLM model provides a high-level computational approximation of the lexicon of a single language user. Although ideally a DLM model would be exposed to the language use and language exposure of a given speaker at a given point in their life, the data required for implementing such a model are not available (and for ethical reasons, I am hoping such data will never become available). As a consequence, a DLM model is usually trained on community-based form representations (in what follows, words' standard written forms) and community-based semantic representations (in what follows, word2vec embeddings).

A linear mapping $\boldsymbol{F}$ that transforms a form vector $\boldsymbol{c}$ into its corresponding embedding $\boldsymbol{s}$,

$$\boldsymbol{s} = \boldsymbol{F}\boldsymbol{c},$$

can be estimated in three ways, given a corpus-based lists of word forms, the corresponding word frequencies, and their embeddings. First, one can estimate an optimal mapping for the set of word types without taking the frequencies with which these types are used into account. Such mappings estimate what we have called the 'endstate of learning' (henceforth EOL), which provides the regression solution for $\boldsymbol{F}$. This regression solution is mathematically equivalent to the asymptotic solution that would be obtained with infinite step-by-step learning experience with the set of word types.

Second, one can let the model incrementally learn a mapping, step by step and token by token. This method comes into its own for training data that are explicitly ordered in time. Heitmeier et al. (2023b) used it to study within-experiment lexical learning, and also applied it to chronologically ordered data sets from CHILDES (MacWhinney 2000). Differences in frequency of use in the training data are taken into account automatically, as more frequent words will be encountered more often during learning. However, this method is computationally intensive and relatively slow.

When no information on the temporal order of learning events is available, or when order information is not essential, then one can implement a form of frequency-weighted regression, henceforth FIL (frequency-informed learning, see Heitmeier et al. 2023a, for detailed discussion). FIL is as computationally light as EOL, and its predictions are similar to those of incremental learning, except for words with tokens that appear primarily either early in training, or late in training.

In what follows, we use the DLM to assess the pros and cons of working with regularized plural vectors, obtained by taking the empirical vectors of singulars and adding the average plural shift vector. Figure 3 presents the accuracies of EOL and FIL mappings constructed for the 11,500 singular and plural noun pairs the embeddings of which were previously visualized in Figure 1. The blue dots in Figure 3 represent the case in which a prediction is counted as correct when the predicted semantic vector has its targeted gold-standard vector as closest neighbor (accuracy@1). Orange dots represent the case in which the target vector is among the closest 10 neighbors of the predicted embedding (accuracy@10). Models using FIL are summarized in the upper panels, and models using EOL are visualized in the lower panels. The left panels present type-wise accuracy, the right panels present accuracy when token frequency is taken into account.

Figure 3 clarifies that when we regularize plurals by deriving them from their singulars with the average shift vector, learning accuracy always increases. The learning task is simpler, the plurals are more predictable. Second, when frequency of use is not taken into consideration (EOL learning), the advantage of working with average plural shift vectors is greatest. Higher-frequency words are more difficult to learn, they have more form neighbors. Regularizing high-frequency plurals alleviates this problem especially for EOL mappings. Third, with FIL, higher-frequency words are learned well, in contrast to lower-frequency words. However, with lenient evaluation (accuracy@10), performance is just about as good for empirical plural vectors as for regularized plural vectors.

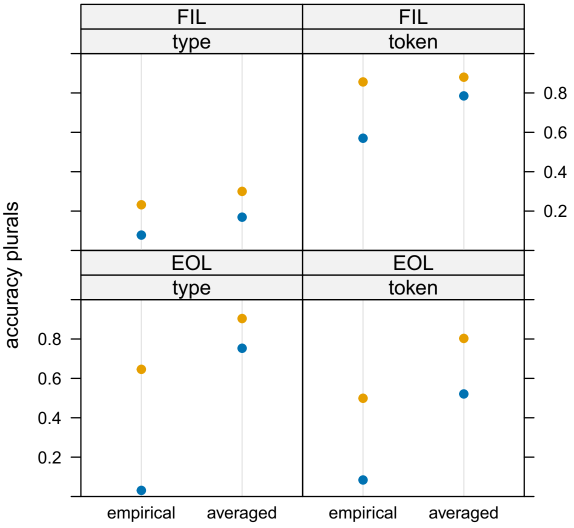

Figure 3: Accuracy@1 (blue) and Accuracy@10 (yellow) for mappings trained with FIL (upper panels) and EOL (lower panels) for plurals with empirical shift vectors and regularized plurals obtained with average shift vectors. Accuracies are evaluated over types (left panels) and over tokens (right panels).

From a usage-based perspective, a model trained with FIL is the most attractive. Heitmeier et al. (2023a) report that FIL outperforms EOL when predicting visual lexical decision latencies. FIL provides a good window on what words are more difficult to learn, and where speakers and listeners are likely to have learned words less well. Evaluation on a token-basis is attractive because it assesses how well learning serves actual use: knowing words that are encountered frequently is more important than knowing words that are hardly ever used. Evaluating with Accuracy@10 is also attractive, for two reasons. First, we are dealing with a classification task with 11,500 possible outcomes. Getting close is already an achievement. Second, from a cognitive perspective, it is highly unlikely that a listener would arrive at exactly the same conceptualization as the speaker, as language users have different experiences with the language and different life histories. Provided interlocutors succeed in communicating meanings that are similar enough, communication can proceed unhindered. The fact that token-wise Accuracy@10 is very high for empirical plural vectors, and almost as good as Accuracy@10 with regularized plural vectors, therefore suggests that from a usage-based perspective, not much is to be gained by regularizing plural vectors.

At the same time, regularizing plural vectors comes with a clear loss. Regularized plurals are a far cry from the true meanings of plural forms as gauged with the empirical embeddings. The averaging process used to generate regularized plurals filters out all world knowledge. The different ways in which multiple objects of the same type configure in the world (see Figure 4 for examples) is not part of what grammars with general shift vectors (or symbolic features) can predict. If the task of a grammar is to model and predict as well as possible what speakers really understand given text or speech input, and what speakers say when they have an idea or a message in mind, then simplified plural vectors simply don't do the job. Prediction of exactly what a speaker had in mind is not possible, and not necessary, but this modeling example shows we can get the gist right. For instance, in the hypothetical

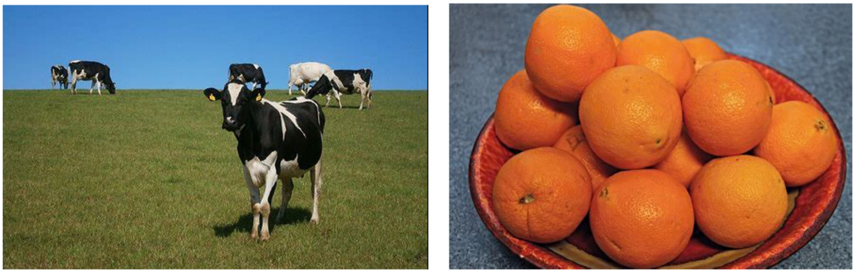

Figure 4: Multiple cows (left) and multiple oranges (right). A decent computational model should be able to get the gist of plural configurations right. Source: Pixabay.com.

case that the plural *apples* has not been encountered, the similarities in shape and function of apples and oranges may well be picked up from the embedding space, allowing the model to predict for apples configurations in the world similar to those of oranges and other large fruits.

3.2 Determining determiners

Behavioral profiling, analogical modeling of language (Skousen 1989) and memory-based learning (Daelemans and Van den Bosch 2005) typically work with categorical predictors that are carefully curated by the analyst. For any algorithm, decisions about representations are unavoidable. For behavioral profiling, compiling well-motivated and replicable feature values for large numbers of sentences, when done by hand, can be prohibitively expensive. More important for the present discussion than this pratical issue is that categorical features may not offer the precision that we really need.

By way of example, consider the problem of predicting the English determiner. Divjak et al. (2023) studied determiner use in the British National Corpus. To my mind, it is an exemplary study of what can be accomplished with behavioral profiling. And yet, have these authors really solved the problem?

Divjak et al. (2023) coded 2,200 tokens of nouns by hand for the number and countability of the noun, for whether the noun is part of a set phrase, for the presence of pre- or post-modifiers, for hearer knowledge, and for specificity. They report very high accuracies (close to perfect) using recursive partitioning trees and random forests. At first sight, this might suggest that with carefully selected and theoretically well-motivated discrete features, the problem of determiner selection in English is now solved.

However, the features provided to the tree and forest classifiers are themselves based on considerable human interpretation. For instance, ‘Hearer Knowledge’ is reported to be “based on whether it appears that the speaker assumes that the hearer knows what they are referring to” (p. 12). The authors admit that applying these kinds of concepts to actual data is not always straightforward. In fact, in the case of ‘Hearer Knowledge’, a problem of form (predicting the determiner) is replaced by a problem of meaning: figuring out when ‘Hearer Knowledge’ is present or absent. Although I can get some intuitive meaning out of this term, my mind shuts down on me when I try to think through what it actually means. I can imagine that for non-native speakers of English, developing a sense of what in the English speaking world counts as ‘Hearer Knowledge’ is a formidable task. These considerations suggest to me that what we need is a model for predicting the value of ‘Hearer Knowledge’ for a given utterance token. This predicted value can inform a feature-driven symbolic classifier. Unfortunately, this would imply a two-step process, and the lesson from end-to-end modeling in NLP and AI is that in-between steps tend to be suboptimal for modeling accuracy.

Large language models also predict English determiners with high accuracy. Figure 5 illustrates this for some 15,000 nouns in the Buckeye corpus (Pitt et al. 2005), using contextualized embeddings obtained with GPT-2. Two context windows were used: a wide window with 100 sublexical units (about 50 words) to the left of the article, and a narrower window with 20 sublexical units (about 10 words).[7] For both windows, prediction accuracy of a linear discriminant analysis (LDA) with leave-one-out crossvalidation is shown for the contextualized embeddings of the articles themselves (distance 0) as well as for the words preceding the article, up to a distance equal to 3. Baseline accuracy (using the majority choice) is at 0.558. Except for words at distance 3, GPT-2 accuracies are significantly higher than the baseline accuracy.[8] The somewhat better performance of contextualized embeddings obtained with a wider context window suggests that information at quite some distance from the target noun is already providing information on what determiner to use. It is also interesting that without any knowledge of the upcoming noun, and hence without any direct information about the upcoming noun's number or countability, the LDA models are already extremely successful. In other words, the contextualized embeddings of the GPT-2 model capture micro-level variation that may motivate some features (e.g., 'Hearer knowledge'). They also may render others unnecessary. In the case of grammatical number, for instance, the GPT-2 model may already be anticipating the upcoming grammatical number, and hence may not need to be provided explicitly with information about grammatical number that is realized on the noun. Contextualized embeddings appear to be looking into the future. The challenge for linguistics is to explain why.

Of course, the question of whether actual human language users can make use of the information that LLMs extract from planet-wide global usage can only be addressed by experiments with real speakers, and with computational models trained on human-scale input using cognitively plausible learning mechanisms. LLMs are models constructed at a superhuman scale, using non-human learning algorithms. The way in which LLMs predict upcoming determiners is a far cry from what is currently known about human language processing. But my hunch is that they are powerful statistical tools for probing dependencies in language use.

[7] I am indebted to Sean Tseng for calculating these contextualized embeddings for the Buckeye corpus.

[8] For related research supporting the idea that contextualized embeddings provide a lens on the future, see Pal et al. (2023), and for mathematical reasons why this is possible in deep networks, see Wu et al. (2024).

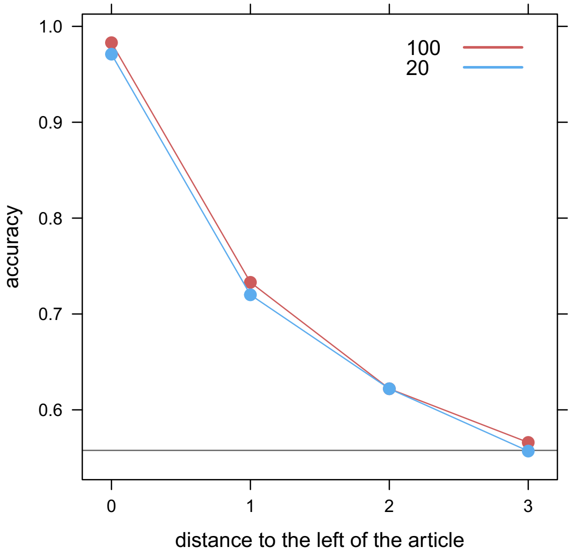

Figure 5: Contextualized embeddings are looking into the future: prediction accuracy for English determiners (*the*, *a*, *an*) in the Buckeye corpus using GPT-2 derived contextualized embeddings of the words at distances 3–0 from the determiner. The gray horizontal lines denotes majority baseline accuracy. Accuracy for a wide preceding context window is shown in red, and accuracy for a narrow preceding context window is shown in blue.

3.3 Models and models

Thus far, I argued that abstract features such as ‘plural’ and ‘Hearer Knowledge’ can be suboptimal because abstraction may render useful fine within-category detail invisible, because abstraction may filter out essential knowledge of the world, and because an abstract feature may itself stand in need of explanation. What then is the role of abstraction, and of symbols, in the scientific method?

I can only offer a tentative line of thought that takes as point of departure the distinction made by Breiman et al. (2001) between machine learning and statistical modeling. Breiman and colleagues describe machine learning as goal oriented, it not being important to understand how a model works as long as it does what it is designed to do. By contrast, in statistics one sets up a mathematical model with predictors and response variables that are represented symbolically, in the hope is that the model can predict and even generate the data, apart from measurement noise, and in this way provide the analyst with an explanation of why we observe what we observe.

LLMs are machine learning models. Just as a smartly compressed version of the collected works of Plato does not provide a model of Plato’s language or Plato’s thought that humanist scholars would find useful, LLMs (conceptualized as smartly compressed corpora that are productive) do not provide a model of human thought or even human language. I think of them as Large Productive Compressed Corpora, rather than as language models.

When we build statistical models, we always work with abstraction and with symbols, but, paraphrasing (Box 1976, p. 176), we should guard against falling in love with our own models. The cautionary note in Box and Draper (1987) that all models are approximations, and thus, essentially wrong (p. 424), also applies to linguistic theories using abstract features for quantitative prediction. The practical question is how wrong a linguistic model has to be to not be useful (cf. Box and Draper 1987, p. 74). Of course, this raises the question of what counts as 'useful'.

Symbolic quantitative models can be useful for language teaching. Here, linguistics has a long history. Assuming that in the near future we will not be wearing automatic translation earplugs, and that there still will be people who would like to learn another language themselves, being able to provide explanations for main trends in language use will continue to be very helpful. Symbolic quantitative models also help us to dig deeper, to engage in the iterative process of refining features so that prediction become increasingly more accurate. In principle, when applied to the same problem, with the same goal, machine learning and iterated linguistic modeling should converge. But I fear that as linguistic models become increasingly refined, the pleasure of understanding will decrease exponentially. Anyone who has tried to make sense of an ANOVA model with a five-way interaction, or a recursive partitioning tree generating hundreds of splits, will know what I mean.

4 The future

What is the future of corpus linguistics and linguistic theory? Within the constraints of the limitations of my own understanding of language and linguistics – the complexities of language and language use continue to both excite and puzzle me – I see three issues that the field will have to address, if it is to survive.

1. We will need to clarify what the benefits for society are for tax money going into funding corpus linguistics and linguistic theory.
2. We will need to make our research carbon neutral.
3. We will need to change the way we work, moving from 'butterfly collection' to incremental model building.

In what follows, I discuss these three issues in turn.

4.1 Why fund research in corpus linguistics and linguistic theory?

In the coming years, I anticipate that funding for the humanities, including linguistics, will decrease as the economic costs of the unfolding climate catastrophe for our societies skyrocket. The current global political situation adds to this with increasing defense spending. For linguistics, and corpus linguistics, to have a chance of survival in the academic marketplace, it is imperative that we can clarify to the tax-payer what the benefits of our discipline for society are.

As a humanities scholar, I strongly believe in the added value of better understanding how the human mind works within societies and their cultures. But as countries have to take measures to deal with, for instance, rapidly rising sea levels, increasing frequency and magnitude of storms, as well as raging forest fires, reduced amounts of funding across academia will flow primarily to applied sciences. Funding agencies in many countries are already requesting grant proposals to clarify their practical value for society, and I expect the pressure on submitting proposals outlining applicable research to increase over the coming years. At the moment, we can get away with paying lip service to applicability. However, as a field, we will have to provide clear answers to the question why linguistics, and corpus linguistics, should be funded. After all, with LLMs, society now has unprecedented highly effective language technology at its fingertips – it seems that there is nothing that this wompom cannot do. Why would citizens be required to fund academic research in corpus linguistics and linguistic theory now that engineers “have figured it all out”? This is a harsh question, and I don’t have any good answers. But here is one thought.

People without linguistic training tend to think that there is correct language use, as opposed to incorrect language use. This rigid perspective on language reinforces ingroup versus outgroup behavior and is easily harnassed to serve the interests of demagogues. However, the accumulated results of corpus linguistics over the past decades indicate that language is not deterministic and fixed, but stochastic, a patchwork of constructions, idioms and collocations, with writing conventions that diverge considerably from actual conversational speech. But this is not common knowledge. Although teachers at secondary schools must have some knowledge of the key findings of corpus linguistics, sociolinguistics, and dialectology, little of this knowledge is passed on to their students. These students, some of whom will later in life be policy makers, need to be much better informed about the variability that is inherent in language.

I’m hoping that a better understanding of the true nature of language will reduce the risk of policy makers to overvalue the standard language or their own sociolect. To achieve this, it will be essential that we better communicate the core insights of our field to the general public. An example from sociolinguistics that I find impressive is the documentary entitled “Talking black in America” (https://www.youtube.com/watch?v=8QFpVgPl9tQ), an excellent response of American sociolinguistics to widespread negative sentiments and misconceptions in the USA about English as spoken by black Americans. For educational purposes, having similar documentaries about key results from corpus linguistics need to be developed. Many outstanding videos explaining central concepts from mathematics and AI are available on the web. Similar materials, perhaps gamified, need to be developed to enable the general public to get a better understanding of language and language use, the variability of language, and the social and cultural importance of language diversity.

4.2 Environmentally sustainable corpus linguistics

In 1950, Sir Ronald Fisher pointed out the following:

> For the future, so far as we can see it, it appears to be unquestionable that the activity of the human race will provide the major factor in the environment of almost every evolving organism. Whether they act consciously or unconsciously human initiative and human choice have become the major channels of creative activity on this planet. Inadequately prepared we unquestionably are for the new responsibilities, which with the rapid extension of human control over the productive resources of the world have been, as it were, suddenly thrust upon us. (Fisher 1950, p. 20)

More than 70 years later, human-induced climate change is unfolding at a catastrophic pace. The planet is seriously ill: it has intelligence. Corpus linguistics and corpus based natural language processing are contributing to climate change, just as all other sectors of modern society. Figure 6 plots cumulative carbon emissions as a function of time from 1930 to 2022 (data from Friedlingstein et al. 2023). This is what

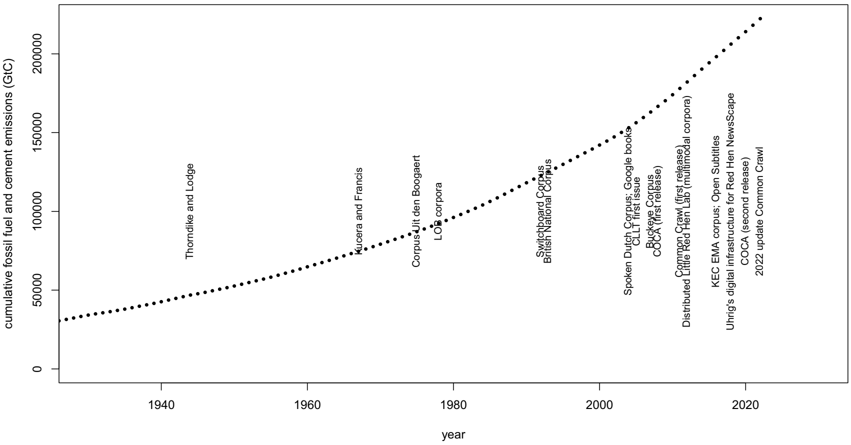

Figure 6: Cumulative fossil fuel and cement emissions (in GtC) for 1930–2022, annotated for the emergence of selected resources for corpus linguistics.

humanity has been pumping into the planet's atmosphere. There is no sign that this behavior will stop anytime soon.

The date of publication of a selection of resources for corpus linguistics has been added to the plot, beginning with *A teacher's word book of 30,000 words* by Thorndike and Lorge (1944) and the word frequency lists published by Kučera and Francis (1967), and ending with the 2022 release of Common Crawl. The magnitude of resources now available to corpus linguistics has increased along with the ever growing global carbon footprint of humanity. The contribution of increasingly large corpora to humanity's global carbon emissions is likely to be small, and I know, correlation is not causation. But it is not unlikely that even though storage of electronic data is becoming cheaper, the rate at which storage devices are used is increasing even faster, leading to a net increase in construction costs and of costs of use, across society,[9] including corpus linguistics.

Environmental costs are not only increasing with storage space, but also with computational resources. Data centers housing LLMs and AI systems currently are responsible for 2 % of global electricity consumption, and this figure will have doubled in the near future. These systems also need enormous amounts of fresh water to cool processors. Within years, large AI systems may be consuming as much energy as entire nations (Crawford 2024). The planet not only has intelligence, but now it has artificial intelligence as well. It is important to realize that the carbon footprint of LLMs is not restricted to development, but also extends to use (think of applications such as Meta-Llama 3). A search using generative AI is estimated to use four to five times the energy required by a conventional web search (Crawford 2024).[10]

Even though computations and storage have become increasingly energy efficient, the rate at which our use of the new tools and resources increases has been increasing even more. AI and machine learning are no exception (see, for AI and NLP Wu et al. 2022). Although efforts are under way to mitigate the environmental costs of AI and LLMs (Zhu et al. 2023), what we really need is green AI, defined by Schwartz et al. (2020) as "AI research that yields novel results without increasing computational cost, and ideally reducing it". As long as a novel AI application is not carbon neutral, for the sake of future generations, it should be banned from large-scale public use.[11]

[9] See, e.g., Berners-Lee (2021), p. 95, and Berner et al. (2022).

[10] The difference in costs of using simpler as opposed to complex LLMs can also be huge. The contextualized embeddings calculated (using 50 preceding words) for the Buckeye corpus, which figured in the analyses reported above in Section 3.2, required 0.011 kWh of (local) electricity. An estimate of the cost of using Llama-3-8b instead of GPT-2 is 0.950 kWh, two orders of magnitude higher.

[11] The way that homo 'sapiens' is managing its own waste suggests to me that the social intelligence of our species is remarkably similar to that of the coronavirus.

The challenge to become carbon neutral is not only urgent for AI and machine learning, it extends also to corpus linguistics (and any other scientific discipline). Simple measures at the personal level are to minimize air travel, to drastically reduce meat consumption, and to switch off your modem when not in use. But collectively, we also need to become much better aware of the carbon costs that come with the use of computer programs and applications. What is the carbon footprint of downloading English fasttext embeddings, or of training a deep learning model for 2 weeks, or fitting a linear mixed model with maximal random effects that takes 3 h in R and then reports it didn't converge?[12]

At present, we act without thinking about environmental costs. I confess guilty. We have ethics boards that assess whether our experiments could be harmful to participants, but the harm of carbon-intensive research for upcoming generations seems to be outside the radar of academic ethical awareness. Funding agencies should be encouraging us to explain how we plan to keep emissions low or how we will offset them. Perhaps journals, including Corpus Linguistics and Linguistic Theory, should ask researchers to publish estimates of the amount of electricity that was required for the research reported. As researchers, we will need to learn how to balance the higher carbon costs of energy-intensive methods, as compared to carbon-leaner alternatives, against the benefits of their use (see also Dhar 2020; Schwartz et al. 2020; Strubell et al. 2019).

Let me provide two examples of the kind of issues that I am running up against in my own research practice. The first example concerns fitting statistical models to empirical data. A complex model may require hours of computation time. A simple model may require only a few minutes. Is 3 h of computation time justified by the additional precision that is supposedly gained? This question is exacerbated by experiences indicating that for reasonably sized datasets, statistical models that take many hours to fit are likely overfitting the data.

My second example concerns computational modeling within the discriminative lexicon framework. We have been developing software facilitating the implementation of DLM models (Heitmeier et al. 2024). This software is written in the Julia language, following recommendations of Douglas Bates and Reinhold Klieg, a language that is highly optimized for numerical computations. Most of our models have been implemented with linear mappings between meaning and form. It is surprising how well these simple mappings perform. Recently, however, we have been exploring deep networks. Deep mappings tend to outperform linear mappings with

[12] The computationally highly effecient re-implementation by Douglas Bates in the MixedModels package for julia is highly recommended: it doesn't have any convergence problems, it is several orders of magnitude faster, and hence has a much smaller carbon footprint.

respect to comprehension and production accuracy.[13] The problem that we are now running into is whether to recommend using the simple linear mappings, which for standard datasets can be estimated on a normal CPU within a few seconds, or should we recommend to use deep networks, which take several hours to train on a GPU. Does the greater precision of deep networks justify a carbon footprint that is a thousand times greater?

Perhaps a reasonable environmentally-aware 'code of conduct', both for statistical modeling and for computational modeling, is to conduct exploration with models that are simple,[14] and only build complex models when doing so is truly crucial for theory development or practical application. If so, a complex model should be run once only.

4.3 From butterfly collection to model building

Unfortunately, to put it bluntly, a lot of research on language looks disquietingly like butterfly collecting. An individual study can be well designed, competently executed, and highly informative. Across many such studies, a general picture is emerging of how language works. This is all great. Unfortunately, a body of loosely connected research questions, analysed with a wide variety of research methods (however well-motivated) now has to compete with LLMs. Ideas about how language might work are no match for a technology that actually works, that works extremely well, and that is going to be integrated into widely used application software.

I believe that as a field we should be more ambitious, and aim for comprehensive computational models that are cognitively motivated, trained on human-scale data, with algorithms that respect important properties of human learning. Without implementation, theories of language are at best food for thought, and in the worst case belief systems similar to the many other belief systems that permeate human culture (see Horton 1967, for discussion of the rationality of pre-scientific thought). Humanistic cognitive computational models will be theories of language that enable us to predict much more effectively the details of language acquisition, language change over the lifespan, the interaction of language and social identity, and the interplay of language and cognition. What (corpus) linguistics needs is something along the lines of what Anderson's ACT-R model (see, e.g., Taatgen et al. 2006) is in psychology: an integrated model that generates precise quantitative predictions for vast arrays of empirical findings.

[13] But when it comes to modeling human trial-by-trial lexical learning (Heitmeier et al. 2023b), linear mappings outperform deep mappings (Heitmeier 2024).

[14] Here, I follow Box (1976): "…it is fruitless to attempt to allow for all contingencies in advance so in practice model building must be accomplished by iteration…" (p. 3 of the 1979 technical report).

I am hoping that humanistic cognitive modeling in an interdisciplinary context will also contribute to greening AI. As pointed out by Schwartz et al. (2020), “Ironically, deep learning was inspired by the human brain, which is remarkably energy efficient.” The long-term goal of humanist cognitive computational modeling of language can actually be energy-lean models that take inspiration from on the one hand the fact that we now know for sure that probability in language can be dealt with effectively (as demonstrated by LLMs), and on the other hand by the fact that we also know a green solution exists as well (as demonstrated by the human brain).

5 Concluding remarks

I am hoping that technological advances in the design of computer chips may bring the environmental costs of LLMs down, while at the same time providing linguists with the means of building models that better approximate the green computations of our brains. I am thinking in particular about modeling language processing with spiking neurons. Eliasmith (2013) built a model of the brain, inspired by ACT-R, using spiking neurons. Unfortunately, simulating spiking neurons with standard hardware is extremely computationally intensive. However, simulation studies show that ensembles of spiking neurons can be highly energy efficient (Higuchi et al. 2024b; Yin et al. 2021).

New perspectives are offered by the neuromorphic chips that are currently being developed by the Intel Neuromorphic Research Community, which is making some significant advances on a variety of computational tasks (Davies et al. 2021). One wishful thought is that linguistics will be actively engaged in exploring the potential of spiking neurons. What if neuronal firing in the brain measured with collected high-density intracranial recordings from the human speech cortex on the superior temporal gyrus (see, e.g., Oganian et al. 2023) can be approximated by spiking neurons that respond selectively to different frequency bands, and can extract frequency patters directly within the time domain, at a fraction of the computational costs of a fast fourier transform (Higuchi et al. 2024a, 2024b)?

I am also hoping that linguistic theory will not be incarcerated in the mental prison of next word prediction, in the way that linguistic theory was enthralled for many years by the properties of formal (programming) languages.[15] Next word prediction technology goes back to Markovian models, but are language users really predicting the next word given the sequence of preceeding words, constantly

[15] The history of how recursion made its way into programming languages, as a means of facilitating conceptualization of an iterative procedure (Daylight 2011), makes one wonder why in certain linguistic circles recursion has been elevated to the defining property of language.

recalibrating predictions for each successive word? What if we are anticipating the message, rather than the next word?[16] The discriminative lexicon model, for better or for worse, works with holistic vectors for form that are mapped onto holistic vectors for meaning. At the level of words, this can be made to work. It is an open question whether the model can be made to work also for utterances. However, even within the constraints of our simple model, effects usually attributed to narrowing down of probabilities (Levy 2008) emerge and arise as a consequence of sequential inputting of information to a network (cf. Baayen et al. 2016; Shafaei-Bajestan et al. 2023).

I bring this essay to a close with a partial quote of an abstract that appeared more than 65 years ago:

> The “understanding” of verbal behavior is something more than the use of a consistent vocabulary with which specific instances may be described. . . .The extent to which we understand verbal behavior in a “causal” analysis is to be assessed from the extent to which we can predict the occurrence of specific instances and, eventually, from the extent to which we can produce or control such behavior by altering the conditions under which it occurs. In representing such a goal it is helpful to keep certain specific engineering tasks in mind. How can the teacher establish the specific verbal repertoires which are the principal end-products of education? How can the therapist uncover latent verbal behavior in a therapeutic interview? How can the writer evoke his own verbal behavior in the act of composition? How can the scientist, mathematician, or logician manipulate his verbal behavior in productive thinking? (Skinner 1957)

B. F. Skinner was far ahead of his time. What he describes gets very close to current-day prompt engineering for and fine-tuning of pre-trained LLMs. History has proven him right. In my opinion, linguistics owes Skinner a huge apology. Now that Skinner’s vision has become reality, at least in a technical sense, we need to think through what this means for our understanding of language. I briefly touch upon three issues.

First, LLMs require training on huge volumes of textual and multimodal data. LLMs are not embodied, and this may explain in part why they need far larger volumes of language data, compared to human learners.[17] However, human language use also depends on extensive and intense processes of multimodal learning that unfold across the lifespan.[18] Rather than faulting LLMs for the large amount of training data they (currently) need, I think we should take stock and consider the possibility that we have been severely underestimating the amount of ‘training’ that

[16] Reflecting on the challenges that face any author writing about a complex topic, next-word prediction is a caricature of the actual creative process.

[17] How much data is actually required for LLMs is hotly debated also in AI, see, e.g., Bender et al. (2021), Hoffmann et al. (2022), and more recently, Eldan and Li (2023) and Gunasekar et al. (2023).

[18] Our vocabularies keep expanding as we grow older (Keuleers et al. 2015), and we now also know that our mental lexicons are constantly updated and fine-tuned to experience as we speak (Heitmeier et al. 2023b); for continuous learning in vision, see, e.g., Marsolek (2008).

speakers actually receive from very early on as they learn to navigate their physical and social worlds at the same time.

Second, we are far more predictable than we are comfortable with. We may not like being predictable, but that doesn't change that we are, and that this predictability is exploited. The abovementioned work by Burrows on authorial hands more than 30 years ago has in recent years been taken to unanticipated heights. Companies can now infer with surprising accuracy one's gender and age, and optimize their advertisements accordingly. Every time Chat-GPT3 provides an answer that you find useful, it is succeeding in predicting what you like. Programmers are now finding that LLMs help them write better code. Soon, authors will be using LLMs to write better novels. Of course, as with any technology, LLMs come with real and potentially large dangers of misuse.[19] But Skinner's vision, and LLMs, are not intrinsically good or bad.

Third, it seems to me there is far more continuity between human language and animal communication than has often been assumed. Prairie dogs, for instance, pack meanings into their calls (Dennis et al. 2020; Slobodchikoff et al. 2009) in ways that are reminiscent of how human languages pack motion, figure, path, and manner into verbs (Talmy 1985). Pepperberg (2020) showed that a grey parrot, with a brain the size of a walnut, can learn the concept of zero. Her work suggests that the parrot not just memorized answers to specific questions, but was also able to generalize and answer non-trivial questions about novel configurations of objects. Thus, the parrot was not just 'parroting', it had a memory that generalized. One might argue that parrots need enormous amounts of training, but so do LLMs, and so do we. To be clear, I am not arguing that parrots and prairie dogs are just as intelligent as we are, or that animal languages are just like ours. What I am arguing is that across several species (prairie dogs, parrots, whales, dogs, elephants, octopuses) brains are making use of similar statistical learning algorithms to compress experience into productive memory systems for communication.

References

Alcaraz Carrión, Daniel, Cristóbal Pagán Cánovas & Javier Valenzuela. 2020. Enaction through co-speech gesture: The rhetorical handing of the mental timeline. *Zeitschrift für Anglistik und Amerikanistik* 68(4). 411–431.

Ambridge, Ben. 2020a. Abstractions made of exemplars or 'you're all right, and I've changed my mind': Response to commentators. *First Language* 40(5–6). 640–659.

[19] It will be interesting to see what the feedback loop effect is once LLMs are trained on materials that in increasing amounts comprise materials produced with the help of LLMs.

Ambridge, Ben. 2020b. Against stored abstractions: A radical exemplar model of language acquisition. *First Language* 40(5–6). 509–559.

Arndt-Lappe, Sabine. 2011. Towards an exemplar-based model of stress in English noun–noun compounds1. *Journal of Linguistics* 47(3). 549–585.

Arnold, Dennis & Fabian Tomaschek. 2016. The Karl Eberhards corpus of spontaneously spoken southern German in dialogues – audio and articulatory recordings. In C. Draxler & F. Kleber (eds.), *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, 10–13. Muenchen: Ludwig-Maximilians-Universitaet.

Arppe, Antti. 2008. *Univariate, bivariate and multivariate methods in corpus-based lexicography. A study of synonymy*. Helsinki: University of Helsinki.

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159.

Baayen, R. Harald. 1997. The pragmatics of the 'tenses' in biblical Hebrew. *Studies in Language. International Journal Sponsored by the Foundation "Foundations of Language"* 21(2). 245–285.

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 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*. 2019 https://doi.org/10.1155/2019/4895891.

Baayen, R. Harald, Petar Milin, Dusica Filipović Durdević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118. 438–482.

Baayen, R. Harald, Cyrus Shaoul, John Willits & Michael Ramscar. 2016. Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition, and Neuroscience* 31(1). 106–128.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Shmar- garet Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Berner, Anne, Stephan Bruns, Alessio Moneta & David I. Stern. 2022. Do energy efficiency improvements reduce energy use? Empirical evidence on the economy-wide rebound effect in Europe and the United States. *Energy Economics* 110. 105939.

Berners-Lee, Mike. 2021. *There is no planet B: A handbook for the make or break years-updated edition*. Cambridge: Cambridge University Press.

Binz, Marcel & Eric Schulz. 2023a. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Binz, Marcel & Eric Schulz. 2023b. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences* 120(6). e2218523120.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Bod, Rens. 1998. *Beyond grammar: An experience-based theory of language*. Stanford, CA: CSLI publications.

Bod, Rens. 2006. Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review* 23(3). 291–320.

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6. 1–22.

Borensztajn, Gideon, Willem Zuidema & Rens Bod. 2009. Children's grammars grow more abstract with age – evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science* 1(1). 175–188.

Box, George E. P. 1976. Science and statistics. *Journal of the American Statistical Association* 71. 791–799.

Box, George E. & Norman R. Draper. 1987. *Empirical model-building and response surfaces*. New York: John Wiley & Sons.

Breiman, Leo. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* 16(3). 199–231.

Bresnan, Joan. 2006. Is knowledge of syntax probabilistic? Experiments with the English dative alternation. In *Pre-proceedings of the international conference on linguistic evidence. Empirical, theoretical and computational perspectives*, 2–4.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Kraemer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Burnard, Lou. 1995. *Users guide for the British national corpus*. Oxford: British National Corpus Consortium, Oxford University Computing Service.

Burrows, John F. 1986. Modal verbs and moral principles: An aspect of Jane Austen’s style. *Literary and Linguistic Computing* 1. 9–23.

Burrows, John F. 1992. Computers and the study of literature. In C. S. Butler (ed.), *Computers and written texts*, 167–204. Oxford: Blackwell.

Burrows, John F. 1993. Noisy signals? Or signals in the noise? In *ACH-ALLC conference abstracts*, 21–23. Georgetown.

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

Bybee, Joan L. 1988. Morphology as lexical organization. In Michael Hammond & Michael Noonan (eds.), *Theoretical morphology: Approaches in modern linguistics*, 119–141. London: Academic Press.

Cedergren, Henrietta & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2). 333–355.

Chuang, Yu-Ying & R. Harald Baayen. 2021. Discriminative learning and the lexicon: NDL and LDL. In *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press.

Chuang, Yu-Ying, Janice Fon, Ioannis Papakyritsis & R. Harald Baayen. 2021. Analyzing phonetic data with generalized additive mixed models. In Martin J. Ball (ed.), *Handbook of clinical phonetics*, 108. London: Routledge.

Chuang, Yu-Ying, Dunstan Brown, Roger Evans & R. Harald Baayen. 2023. Paradigm gaps are associated with weird “distributional semantics” properties: Russian defective nouns and their case and number paradigms. *The Mental Lexicon* 17(3). 395–421.

Chuang, Yu-Ying, Melanie J. Bell, Yu-Hsiang Tseng & R. Harald Baayen. 2024. *Word-specific tonal realizations in Mandarin*. Manuscript, Tubingen: University of Tübingen.

Crawford, Kate. 2024. Generative AI’s environmental costs are soaring – and mostly secret. *Nature* 626. https://doi.org/10.1038/d41586-024-00478-x.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Daelemans, Walter & Antal Van den Bosch. 2005. *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, Walter, Peter Berck & Steven Gillis. 1995. Linguistics as data mining: Dutch diminutives. In T. Andernach, M. Moll & A. Nijholt (eds.), *CLIN V, papers from the 5th CLIN meeting*, 59–71. Enschede: Parlevink.

Daelemans, Walter, Antal Van den Bosch & Jakub Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning, Special Issue on Natural Language Learning* 34. 11–41.

Daelemans, Walter, Jakub Zavrel, Ko Van der Sloot & Antal Van den Bosch. 2007. TiMBL: Tilburg memory based learner reference guide. Version 6.1. Technical Report ILK 07-07. Computational Linguistics Tilburg University.

Davies, Mike, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A. Fonseca Guerra, Prasad Joshi, Philipp Plank & Sumedh R. Risbud. 2021. Advancing neuromorphic computing with loihi: A survey of results and outlook. *Proceedings of the IEEE* 109(5). 911–934.

Daylight, Edgar G. 2011. Dijkstra's rallying cry for generalization: The advent of the recursive procedure, late 1950s–early 1960s. *The Computer Journal* 54(11). 1756–1772.

Delétang, Grégoire, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau. 2023. Language modeling is compression. *arXiv preprint arXiv:2309.10668*.

Dennis, Patricia, Stephen M. Shuster & C. Slobodchikoff. 2020. Dialects in the alarm calls of black-tailed prairie dogs (cynomys ludovicianus): A case of cultural diffusion? *Behavioural Processes* 181. 104243.

Dhar, Payal. 2020. The carbon impact of artificial intelligence. *Nature Machine Intelligence* 2(8). 423–425.

Divjak, Dagmar. 2004. Degrees of verb integration: Conceptualizing and categorizing events in Russian. Belgium: University of Leuven Dissertation.

Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23–60.

Divjak, Dagmar, Laurence Romain & Petar Milin. 2023. From their point of view: The article category as a hierarchically structured referent tracking system. *Linguistics* 61(4). 1027–1068.

Drozd, Aleksandr, Anna Gladkova & Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king − man + woman = queen. In Y. Matsumoto & R. Prasad (eds.), *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 3519–3530. The COLING 2016 Organizing Committee.

Eddington, David. 2000. Spanish stress assignment within the analogical modeling of language. *Language* 76(1). 92–109.

Eldan, Ronan & Yuan-Zhi Li. 2023. Tinystories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*.

Eliasmith, Chris. 2013. *How to build a brain: A neural architecture for biological cognition*. USA: OUP.

Ellegård, Alvar. 1953. *The auxiliary do: The establishment and regulation of its use in English*. Stockholm: Almquist & Wiksell.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14. 179–211.

Ernestus, Mirjam & R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79. 5–38.

Fisher, Ronald A. 1950. *Creative aspects of natural law*. Cambridge: CUP Archive.

Friedlingstein, Pierre, Michael O'sullivan, Matthew W. Jones, Robbie M. Andrew, Dorothee C. Bakker, Judith Hauck, Peter Landschützer, Corinne Le Quéré, Ingrid T. Luijkx, Glen P. Peters. 2023. Global carbon budget 2023. *Earth System Science Data* 15(12). 5301–5369.

Gahl, Susanne. 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84(3). 474–496.

Gahl, Susanne & R. Harald Baayen. 2024. *Time* and *thyme* again: Connecting spoken word duration in English to models of the mental lexicon. *Language*, to appear.

Gaskell, M. Gareth & William Marslen-Wilson. 1998. Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 24(2). 380–396.

Godfrey, John J., Edward C. Holliman & Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, vol. 1, 517–520. IEEE Computer Society.

Goldberg, Adele. 2005. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40. 237–264.

Gries, Stefan Th. 2000. *Towards multifactorial analyses of syntactic variation: The case of particle placement*. Hamburg: University of Hamburg Dissertation.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of to run. *Trends in Linguistics Studies and Monographs* 172. 57.

Gries, Stefan Th. 2010. Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5(3). 323–346.

Gries, Stefan Th. & Dagmar S. Divjak. 2009. Behavioral profiles: A corpus-based approach towards cognitive semantic analysis. In Vyvyan Evans & Stephanie S. Pourcel (eds.), *New directions in cognitive linguistics*, 57–75. Amsterdam & Philadelphia: John Benjamins.

Gunasekar, Suriya, Yi Zhang, Jyoti Aneja, Caio César T. Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Harbour, Daniel. 2008. *Morphosemantic number: From Kiowa noun classes to UG number features*. Studies in Natural Language and Linguistic Theory. Dordrecht: Springer.

Heitmeier, Maria. 2024. Mappings in the discriminative lexicon model. Tubingen: University of Tübingen Dissertation.

Heitmeier, Maria, Yu-Ying Chuang, Seth Axen & R. Harald Baayen. 2023a. Frequency-informed linear discriminative learning. *Frontiers in Human Neuroscience, Section Speech and Language* 17. https://doi.org/10.3389/fnhum.2023.1242720.

Heitmeier, Maria, Yu-Ying Chuang & R. Harald Baayen. 2023b. How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *Cognitive Psychology* 146. 101598.

Heitmeier, Maria, Yu-Ying Chuang & R. Harald Baayen. 2024. *The discriminative lexicon: Theory and implementation in the Julia package JudiLing*. Cambridge: Cambridge University Press, to appear.

Higuchi, Saya, Sander M. Bohté & Sebastian Otte. 2024a. Understanding the convergence in balanced resonate-and-fire neurons. *arXiv preprint arXiv:2406.00389*.

Higuchi, Saya, Sebastian Kairat, Sander M. Bohte & Sebastian Otte. 2024b. Balanced resonate-and-fire neurons. *arXiv preprint arXiv:2402.14603*.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8). 1735–1780.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals & Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Horton, Robin. 1967. African traditional thought and western science. *Africa* 37(2). 155–187.

Johnson, Keith. 1997. The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics* 50. 101–113.

Jordan, Michael. 1986. Serial order: A parallel distributed processing approach. Technical report, June 1985-March 1986. Technical report, California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science.

Keuleers, Emmanuel, Michael Stevens, Pawel Mandera & Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology* 8. 1665–1692.

Koesling, Kristina, Gero Kunter, R. Harald Baayen & Ingo Plag. 2012. Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech* 56(4). 529–554.

Kučera, Henry & Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Levshina, Natalia. 2022. Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft* 41(1). 179–205.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177.

Lieber, Rochelle. 2004. *Morphology and lexical semantics*, vol. 104. Cambridge: Cambridge University Press.

Maaten, Laurens v. d. & Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9. 2579–2605.

MacWhinney, Brian. 2000. The childes project. *Tools for Analyzing Talk. Part*, 1.

Magnuson, James S., Heejo You, Sahil Luthra, Monica Li, Hosung Nam, Monty Escabi, Kevin Brown, Paul D. Allopenna, Rachel M. Theodore, Nicholas Monto & Jay G. Rueckl. 2020. Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive Science* 44(4). e12823.

Mandelkern, Solomon. 1896. *Veteris Testamenti concordantiae'hebraice atque chaldaice… Concinnavit Solomon Mandelkern…*, vol. 1. Graz: Veit et Company.

Marsolek, Chad J. 2008. What antipriming reveals about priming. *Trends in Cognitive Sciences* 12(5). 176–181.

McClelland, James L. & Karalyn Patterson. 2002a. Rules or connections in past-tense inflections: What does the evidence rule out. *Trends in Cognitive Sciences* 6(11). 465–472.

McClelland, James L. & Karalyn Patterson. 2002b. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* 6(11). 465–472.

McClelland, James L. & Karalyn Patterson. 2002c. 'words or rules' cannot exploit the regularity in exceptions: Reply to Pinker and Ullman. *Trends in Cognitive Sciences* 6(11). 464–465.

McClelland, James L. & David E. Rumelhart (eds.). 1986. *Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models*. Cambridge, Mass: MIT Press.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. 3111–3119.

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević & R. Harald Baayen. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4). 507–526.

Nikolaev, Alexander, Yu-Ying Chuang & R. Harald Baayen. 2023. A generating model for Finnish nominal inflection using distributional semantics. *The Mental Lexicon* 17(3). 368–394.

Oganian, Yulia, Ilina Bhaya-Grossman, Keith Johnson & Edward F. Chang. 2023. Vowel and formant representation in the human auditory speech cortex. *Neuron* 111(13). 2105–2118.

Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Lou Boves, Jean-Pierre Martens, Michael Moortgat & R. Harald Baayen. 2002. Experiences from the spoken Dutch corpus project. In Manuel Gonz ez Rodriguez & Carmen Paz Su ez Araujo (eds.), *Proceedings of the third international conference on language resources and evaluation*, 340–347. ELRA.

Pal, Koyena, Jiuding Sun, Andrew Yuan, Byron C. Wallace & David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. *arXiv preprint arXiv:2311.04897*. https://doi.org/10.18653/v1/2023.conll-1.37.

Pepperberg, Irene M. 2020. The comparative psychology of intelligence: Some thirty years later. *Frontiers in Psychology* 11. 531634.

Pinheiro, José C. & Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. Statistics and Computing. New York: Springer.

Pinker, Steven & Michael Ullman. 2002. Combination and structure, not gradedness, is the issue. *Trends in Cognitive Sciences* 6. 472–474.

Pitt, Mark, Keith Johnson, Elizabeth Hume, Scott Kiesling & William Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45(1). 89–95.

Plag, Ingo, Julia Homann & Gero Kunter. 2017. Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics* 53(1). 181–216.

Plag, Ingo, Lea Kawaletz, Sabine Arndt-Lappe & Rochelle Lieber. 2023. Analogical modeling of derivational semantics. Two case studies. In Kotowski, Sven and PLAG, Ingo (eds.). *The semantics of derivational morphology: Theory, methods, evidence*. Berlin, Walter de Gruyter. 103–141.

Polomé, Edgar C. 1967. *Swahili language handbook*. Language Handbook Series. Washington, D.C: Center for Applied Linguistics.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Radford, Alec, Karthik Narasimhan, Tim Salimans & Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report. OpenAI.

Rescorla, Robert A. & Allan R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Abraham H. Black & William F. Prokasy (eds.), *Classical conditioning II: Current research and theory*, 64–99. New York: Appleton Century Crofts.

Romain, Laurence, Adnane Ez-zizi, Petar Milin & Divjak Divjak. 2022. What makes the past perfect and the future progressive? Experiential coordinates for a learnable, context-based model of tense and aspect. *Cognitive Linguistics* 33(2). 251–289.

Rumelhart, David E. & James L. McClelland (eds.). 1986. *Parallel distributed processing. Explorations in the microstructure of cognition. Vol. 1: Foundations*. Cambridge, Mass: MIT Press.

Rumelhart, David E., Geoffrey Hinton & Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323. 533–536.

Scharenborg, Odette. 2008. Modelling fine-phonetic detail in a computational model of word recognition. In *The 9th annual conference of the international speech communication association*, 1473–1476. ISCA Archive.

Schwartz, R., Jesse Dodge, Noah A. Smith & Oren Etzioni. 2020. Green ai. *Communications of the ACM* 63(12). 54–63.

Seidenberg, Mark S. & Laura M. Gonnerman. 2000. Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences* 4(9). 353–361.

Shafaei-Bajestan, Elnaz, Peter Uhrig & R. Harald Baayen. 2022. Making sense of spoken plurals. *The Mental Lexicon* 17(3). 337–367.

Shafaei-Bajestan, Elnaz, Masoumeh Moradipour-Tari, Peter Uhrig & R. Harald Baayen. 2023. Ldl-auris: A computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience* 38(4). 509–536.

Shafaei-Bajestan, Elnaz, Masoumeh. Moradipour-Tari, Peter Uhrig & R. Harald Baayen. 2024. The pluralization palette: Unveiling semantic clusters in English nominal pluralization through distributional semantics. *Morphology*. https://doi.org/10.1007/s11525-024-09428-9.

Shahmohammadi, Hassan, Maria Heitmeier, Shafaei Bajestan, Hendrik P.A. Lensch & R. Harald Baayen. 2023. Language with vision: A study on grounded word and sentence embeddings. *Behavior Research Methods*. 1–25.

Skinner, Burrhus F. 1957. *A functional analysis of verbal behavior*. New York: Appleton-Century-Crofts.

Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Kluwer.

Slobodchikoff, Con, William Briggs & Patricia Dennis. 2009. Decoding the information contained in the alarm calls of gunnison prairie dogs. *Journal of the Acoustical Society of America* 125(4 Suppl). 2739.

Strubell, Emma, Ananya Ganesh & Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(09). 13693–13696.

Stupak, Inna V. & R. Harald Baayen. 2022. An inquiry into the semantic transparency and productivity of German particle verbs and derivational affixation. *The Mental Lexicon* 17(3). 422–457.

Taatgen, Niels A., Christian Lebiere & John R. Anderson. 2006. In Sun, Ron (ed.), *Modeling paradigms in ACT-R. Cognition and multi-agent interaction: From cognitive modeling to social simulation*, 29–52. Cambridge: Cambridge University Press.

Talmy, Leonard. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Description* 3. 57–149.

Thorndike, Edward L. & Irving Lorge. 1944. *A teacher's word book of 30.000 words*. New York: Columbia University Press.

Uhrig, Peter. 2018. Newsscape and the distributed little red hen lab – a digital infrastructure for the large-scale analysis of tv broadcasts. In Anne-Julia Zwierlein, Jochen Petzold, Katharina Böhm & Martin Decker (eds.), *Anglistentag 2017 in regensburg: Proceedings. Proceedings of the conference of the German association of university teachers of English*, 99–114. Trier: Wissenschaftlicher Verlag Trier.

Uit den Boogaart, P. C. (ed.). 1975. *Woordfrequenties in Gesproken en Geschreven Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30.

Wood, Simon N. 2017. *Generalized additive models*. New York: Chapman & Hall/CRC.

Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, James Huang, Charles Bai. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* 4. 795–813.

Wu, Wilson, John X. Morris & Lionel Levine. 2024. Do language models plan ahead for future tokens? *arXiv preprint arXiv:2404.00859*.

Yin, Bojian, Federico Corradi & Sander M. Bohté. 2021. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence* 3(10). 905–913.

Yule, G. Udney. 1944. *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zhu, Xunyi, Jian Li, Yong Liu, Can Ma & Weiping Wang. 2023. A survey on model compression for large language models. arXiv preprint arXiv:2308.07633.

Zipf, George K. 1949. *Human behavior and the principle of the least effort. An introduction to human ecology*. New York: Hafner.