Acta Linguistica Academica 71 (2024) 1–2, 67–86
DOI: 10.1556/2062.2024.00666

AKADÉMIAI KIADÓ

A rudimentary analogy-based model of how syntactic constructions emerge from linguistic experience

MÁTYÁS LAGOS*

HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary

Received: February 13, 2023 • Revised manuscript received: December 21, 2023 • Accepted: April 8, 2024

ABSTRACT

This paper highlights theoretical issues in construction grammar and presents a simple computational language model as a preliminary solution to these issues. The specific issues dealt with are the lack of explicit definition of syntactic categories and the lack of explicit proposals regarding how constructions can be learned from linguistic experience. The proposed language model, called the "analogical path model", learns two-word syntactic patterns from sentences by finding distributional analogies between word pairs. The theoretical relevance and implications of the analogical path model are discussed at the end of the paper.

KEYWORDS

construction grammar, distributional learning, analogy, language models, syntactic categories

1. INTRODUCTION

I begin by explaining the problem that this paper tries to solve. In Section 2 I will present a bigram language model that learns two-word syntactic patterns from sentences by finding distributional analogies, and in Section 3 I will discuss its relevance within theoretical linguistics.

* Corresponding author. E-mail: lagos.matyas@nytud.hun-ren.hu

1.1. Constructions

The theoretical linguistic framework called *construction grammar*, first outlined by Fillmore, Kay & O'Connor (1988), is an approach to describing linguistic regularities that was developed as an alternative to the then mainstream generative view of grammar. From the start, construction grammar was built to address the fact that grammatical constructions – syntactic patterns with some associated function, like the "passive construction" – may be *partially idiomatic*: they often have syntactic or semantic properties that are not predictable from their components (see pp. 506–510 in Fillmore et al.'s paper for examples) and it is therefore inadequate to describe them by atomistically decomposing them into general syntactic categories like "Noun" and "Verb".

This emphasis on idiomaticity has meant that in theories that adopt the construction grammatical framework, syntactic patterns tend to be thought of as things that speakers must learn from experience (like words), as opposed to generative theories where the properties of syntactic patterns are mostly thought to be consequences of some basic structural properties of humans' innate capacity for language. Goldberg & Herbst (2021), when introducing their construction grammatical approach, define a construction as "a conventional combination of form and function *that emerges from the dynamic clustering of witnessed exemplars in memory*" (p. 285, emphasis mine), the "witnessed exemplars" being the sentences heard by speakers throughout their lives. Constructions in this sense are not the syntactic patterns themselves; rather they are the form–function associations that constitute speakers' linguistic competence, allowing speakers to interpret instances of the corresponding syntactic patterns as meaningful expressions. (I will use the word *construction* in this latter sense throughout this paper.)

But looking at the way that theories within construction grammar have formally represented constructions, it is hard to get an idea of how they could emerge from linguistic experience. Take Goldberg & Herbst's (2021, 302) schematic representation of the *nice-of-you* construction (as used in the sentences *It was nice of you to help me* and *It is typical of them to be late*), which I partially reproduce in Table 1 below, omitting the semantic functions of the constituents for the sake of brevity. The top row of the table describes the syntactic categories of the constituents of the construction, and the bottom row lists the most frequent words that occur as constituents from most to least frequent (again some are omitted for brevity) based on corpus data.

Goldberg and Herbst make a compelling case for the importance of considering the idiosyncratic semantics and pragmatics of this construction when describing its usage (their section 3, pp. 291–299), but their analysis tells us little about how speakers might infer the syntactic and semantic properties of the *nice-of-you* construction from the sentences they hear. Assuming that the construction in Table 1 is part of speakers' linguistic competence raises two questions in particular:

1. What are the criteria for something to be an "AdjP" or a "V"?
2. How do speakers acquire these categories and the order in which they are supposed to occur in the construction?

It is not a serious flaw of Goldberg and Herbst's paper in particular that they do not address these questions – representations like theirs are standard in construction grammar and it is not

their goal to account for the acquisition of the *nice-of-you* construction. But the two questions do point to two issues, stated in the following two paragraphs, that make it difficult to interpret not only the explanations of particular phenomena proposed within the constructionist framework, but also the general picture of the nature of linguistic knowledge advocated by construction grammarians.

First, a schema like the one in Table 1 is supposed to be a general description that expresses the set of concrete sentences (or sentence fragments) whose well-formedness it is trying to account for. The schema could then be said to predict this set of sentences to be well-formed. In Goldberg and Herbst's case, their representation should express concrete sentences like *It was nice of you to help me* and *It would be smart of them to work together*. But representations that refer to abstract syntactic categories are not interpretable as predictions about concrete sentences unless explicit criteria are included that let us decide whether e.g. *smart* is an instance of "AdjP"; otherwise there is no way to go from the abstract to the concrete and the representations cannot be used as descriptions of patterns. The lack of explicit criteria in Goldberg and Herbst's paper is not surprising: Crystal (1966) and more recently Croft (2001, 41–45) have shown that linguists tend to use such category labels either without a definition or according to a definition based on their own arbitrarily selected criteria.

Second, without an explicit description of a process through which speakers could acquire constructions from the sentences they hear, the statement that constructions emerge from linguistic experience is a weak statement that leaves construction grammar open to valid criticism questioning its theoretical aims. As Adger (2013) says:

> [construction grammar] proponents have to provide a theory of how learning takes place so as to give rise to a constructional hierarchy [in which constructions are related to each other based on their similarities], but even book length studies on this, such as Tomasello (2003), provide no theory beyond pattern-matching combined with vague pragmatic principles of intention-reading and analogy. Tomasello's book, in particular, claims to provide a 'usage-based theory of language

Table 1. Partial reproduction of the *nice-of-you* construction from Goldberg & Herbst (2021, 302)[1]

Subj	V	AdjP	PP (of)	to-infinitive
it,	BE,	*nice, good, stupid,*		
this,	*would,*	*wrong, sweet, typical,*		
that	*might*	*silly, clever,* [...]		

[1]The verb BE is set in small caps because it refers to all forms of the lexeme, and I think the authors meant to write something like *would* BE and *might* BE under the "V" column instead of *would* and *might*.

> acquisition' but no theory is ever given, just evidence for truisms such as that children can detect patterns and that they want to communicate. (p. 473)[2]

1.2. Emergence

Construction grammarians are aware of these problems. Croft (2005), after an excellent demonstration of the uselessness of syntactic categories (pp. 277–283), proposes to abandon them altogether and to take constructions to be the primitive units of grammar (p. 283). As to how constructions may be learnable from the sentences heard by speakers, he says that they can be identified based on three things (p. 283):

1. their distinctive structures (he gives no example here, but Croft (2001, 52) refers to the "Direct Object" position that is present in active transitive constructions but not in passive constructions),
2. their unique substantive properties, like the appearance of the word *by* in passive constructions, and
3. their unique semantic properties, such as the different semantic roles of the grammatical subject in active transitive versus in passive constructions.

These are all plausible ideas. But Croft knows that his descriptions are too vague to really be usable in any explicit model of the acquisition process. At the end of his paper he concludes that important questions remain unresolved about the network of constructions constituting speakers' linguistic competence, including the question of how this network can come into being: "fundamental issues about the establishment of schemas and the interaction between frequency and similarity of utterances in constructing the network need to be addressed both theoretically and empirically" (p. 310).

Croft thinks that we should discard syntactic categories from our descriptions of constructions. I agree. Beyond the lack of explicit criteria for identifying their members there is also a "logical argument" (Croft 2001, 282) against syntactic categories: they are defined by their distributions, so we cannot define constructions i.e. distributional patterns by referring to these categories or else we get a circular definition. A category-free approach avoids such problems entirely.

But without syntactic categories, schematic formal representations like the one in Table 1 cannot be built. How can we then represent constructions? My answer is to not represent them as distinct units at all but rather to adopt the view that constructions do not exist independently of the memories of concrete sentences from which they are learned: they are nothing but holistic properties of the system into which our linguistic experience is organised. This is the sense in which I consider constructions to be *emergent*.

The challenge now is to actually give content to this view by describing how constructions – or rather the holistic properties that we can think of as constructions – could emerge from linguistic experience. I will do this in the next section, in which I consider a simple syntactic

[2]I have not read Tomasello's book in full, but I can confirm that in his sections 5.2 (pp. 161–175) and 5.3 (pp. 175–181), where he aims to outline "some pattern-finding cognitive processes" that could help us understand how children can acquire abstract linguistic representations (p. 161), he does not propose any explicitly described process that produces constructions from utterances.

pattern identification task and present a solution to it. Through this simplified model I hope to provide the beginnings of a theory that (a) avoids the issue of defining syntactic categories and (b) includes an explicitly described procedure that induces constructions from concrete sentences.

2. AN ANALOGY-BASED LANGUAGE MODEL

In order to be able to present this idea to linguists who are not familiar with bigram language models I will give a brief introduction to the notion of language modelling in subsections 2.1–2.3. I will then describe my idea for learning syntactic patterns in subsections 2.4–2.8 and evaluate it in subsections 2.9–2.10.

2.1. Language models

A *language model* assigns *probabilities* – real numbers between 0 and 1 obeying some laws – to sequences of words.[3] (I will use the term "word" in the "word form" sense as opposed to the "lexeme" sense throughout this article.) The probability assigned to a word sequence by a language model can be interpreted as the degree to which this sequence is expected by the model to occur in a text: the closer a probability is to 1, the higher degree of expectedness it represents.

We consider a particular language model to be a good model of some language if it assigns probabilities to word sequences in a way that would match their relative frequency of occurrence if we could take an infinitely large sample of sentences in this language. A good language model should thus be able to (among other things) distinguish the grammatically well-formed sequences from the grammatically ill-formed ones by assigning higher probabilities to the former kind and lower probabilities to the latter kind.

In order to correctly estimate the probabilities of word sequences, language models usually rely on statistical information obtained from a corpus of sentences in the language to be modelled, called the *training data* of the model. The kinds of statistical information that such a model records and the way that it uses this information to assign probabilities to word sequences determine how well the model can predict the probability of a word sequence in a given language.[4]

2.2. Bigram language models

The particular language model that we will be working with is a simple one: it is a *bigram* language model (see Jurafsky & Martin 2023 for a more thorough introduction) which is restricted to only be able to record information about pairs of directly adjacent words (*bigrams*).[5] This means that the only kind of statistical information that such a model has access

[3]More precisely speaking it assigns probabilities to sequences of *tokens*, but for the purposes of this article we can assume that the tokens are always word forms.

[4]Again more precisely speaking the model itself does not record or use information but is rather defined based on certain kinds of information; but keeping this distinction in mind I will refer in this article to the model doing such things.

[5]The reason for working with such a simple model is to make it as easy as possible to formulate and evaluate my idea for how syntactic patterns can emerge. Extending this idea to e.g. trigram models remains a task for the future.

to is how many times a given word has occurred directly after another word in the training data. For example after reading the sentences

- *Mary saw a movie*,
- *John heard a song*, and
- *John heard the news*,

the model will be able to recall e.g. that

- *saw* occurred after *Mary* once,
- *movie* and *song* occurred after *a* once, and
- *heard* occurred after *John* twice,

but it will not be able to recall that *movie* and *saw* occurred in the same sentence or that *song* and *heard* occurred in the same sentence.

How could such a simple model "learn" English? More concretely: Given a novel sequence of words (whose words all occur in the training data but not in this particular combination) how can this model tell if this sequence of words is a probable English sentence? Clearly it will not recognise syntactic dependencies that span over more than two words: for example it will not notice the incongruity in the word sequence *You often runs* because it has no way to know that subject–verb agreement has to hold even when the subject and the verb are not directly next to each other.

So such a model will never be able to predict sentences like a native English speaker. But it can still predict sentences with some low but above-chance-level accuracy based on observing what the most typical English word pairings are. This "degree of typicality" is expressed by the *(forward) conditional probability* of a bigram like *Mary saw*. Conditional probability is calculated based on the number of cases in which *Mary* is followed by *saw* in the training data; for example if *Mary* occurs 13 times in the training data and 5 times out of 13 it is followed by *saw* then the conditional probability of the bigram *Mary saw* is 5/13 = 0.385. Let us denote this conditional probability by the expression $P(saw|Mary\_)$ (expressing the probability of the word *saw* following the word *Mary* in a text) and the frequency function by the letter *f*, so that $P(saw \mid Mary\ \_) = \frac{f(Mary\ saw)}{f(Mary)}$.

Then when the model is asked to estimate the probability of a word sequence, it can successively compute the probability with which each word follows the previous one in the sequence based on the training data and it can multiply together these probabilities in order to get the probability of the whole sequence.

2.3. Undergeneralisation

This method of calculating the probability of a word sequence quickly runs into problems. It is an empirical fact about human languages that even in a large corpus of sentences most of the possible bigrams will occur very infrequently (as demonstrated by e.g. Yang 2010, 6) or not at all. This means that for any given word there will be many words that never occur directly after it in our training data. The conditional probability of any such pair of words is zero according to the model just described because if e.g. the bigram *Mary levitates* does not occur in the training data then $P(levitates \mid Mary\ \_) = \frac{0}{f(Mary)} = 0$ (and the estimated probability of

any sentence that contains even one bigram with zero conditional probability would be zero). In other words our model cannot make generalisations over the bigrams that is sees in the training data: if a bigram occurs in the training data then it has non-zero probability, otherwise it has zero probability.

But in order to correctly estimate the probabilities of sentences, a language model should clearly be able to generalise over the training data. This leads us to the question that my idea tries to solve: Given two words that never occur next to each other in the training data how can we tell if they can occur next to each other in any given English sentence? (There already exist so-called "smoothing" methods for generalising over bigrams; for an overview see Jurafsky & Martin (2023, 43–51) and we will see in Subsection 2.10 of this article how well some of these techniques perform compared to my proposal.)

2.4. Analogical paths

Suppose that the bigram *quite late* does not occur in the training data. What could give us reasons to believe that *late* can occur after *quite*? My answer is to look at other bigrams that *do* occur in the training data and see if they can be used as evidence for thinking that *late* can occur after *quite*. Here are the steps of searching for such analogical evidence:

1. look at each word *A* that occurs directly before *late* in the training data, as indicated by the arrow in Figure 1(a) below, and
2. look at each word *B* that occurs directly after *quite*, as in Figure 1(b), and
3. for each such word pair *A* and *B* see whether the bigram *A B* occurs in the training data; if it does as in Figure 1(c) then we can consider this to be evidence that raises the probability of the bigram *quite late*.

To see why the configuration in Figure 1(c) may work as analogical evidence supporting the bigram *quite late* it helps to work with a concrete example. Suppose that the bigrams *quite happy*, *so happy* and *so late* all occur in the training data (e.g. in the sentences *I am* ***quite happy***, *Mary was* ***so happy*** *to arrive* and *John has never been up* ***so late***). We then have the configuration shown in Figure 2.

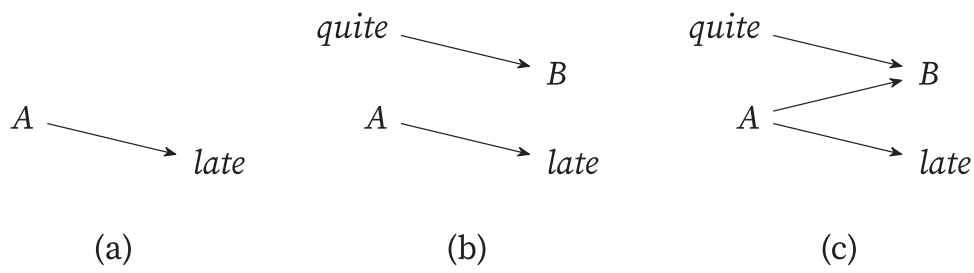

Figure 1. Searching for analogical evidence for the bigram *quite late*

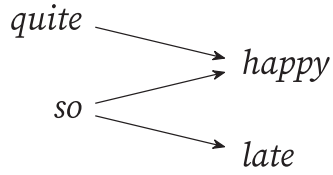

Figure 2. The bigram *so happy* as analogical basis for the bigram *quite late*

Using *so happy* as analogical basis in the above case amounts to making the following inductive generalisation:

> if two words have *at least one* common attested right context in the training data (in the case above *happy* is a common attested right context of *quite* and *so*)
>
> then *every* attested right context of one of these two words is also a possible right context of the other word (in the case above *late* is an attested right context of *so* and from this we infer that *late* is also a possible right context of *quite*).[6]

(To what extent this generalisation actually holds depends on how much we increase the probability of a bigram based on such analogical evidence; the next subsection will define how the amount of support can be calculated.)

I will call the path formed by the arrows from *quite* to *late* above an *analogical path* from *quite* to *late* because we construct the unattested bigram *quite late* by analogy to the attested bigram *so happy*. Analogical paths enable generalisation: even without the bigram *quite late* occurring in the training data we can obtain evidence that the probability of *late* following *quite* is more than zero.

We may find more than one analogical path between two words. Consider the configuration in Figure 3. Here we have two analogical paths from *quite* to *late*:

- one through *happy* and *so* (drawn in black) and
- another through *nice* and *very* (drawn in grey).

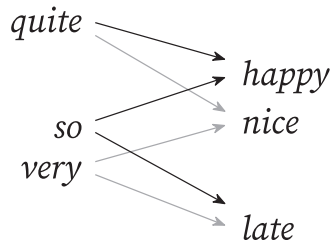

Figure 3. Multiple analogical paths from *quite* to *late*

[6]There already exists a notion in the theory of formal languages that is in some sense equivalent with this generalisation: the class of *zero-reversible* formal languages is the subclass of regular languages for which this generalisation holds, as defined and shown by Angluin (1982, 748 (Theorem 7)). The related notions of *one-reversibility* and *two-reversibility* were also applied by Berwick & Pilato (1987) for the inference of finite state grammars for certain fragments of English.

This should support assigning an even higher probability to the bigram *quite late*. Let us call the probability that is assigned to a bigram by accumulating analogical paths the *analogical probability* of the bigram; the next two subsections will define how the analogical probability of a bigram can be calculated.

2.5. Weighting the paths

This method of inferring bigrams faces the same problem that all analogy-based methods face: there are cases in which analogical inference is not justifiable. Consider what happens if the training data contains the sentences *I liked* ***you in*** *that movie*, *Put* ***it in*** *the box* and ***It runs*** *quickly*. We would have the configuration in Figure 4.

According to the method described above, this path would increase the analogical probability of the bigram *you runs* – but this bigram would be obtained by using the bigram *it in* as analogical basis. The bigram *you runs* is actually a possible bigram of English – e.g. in the sentence *The machine behind* ***you runs*** *hot* – but it seems unlikely that speakers form it on the basis of the bigram *it in*, so the configuration in Figure 4 should only minimally increase the probability of *you runs*.

And our model should in general be able to detect when an analogy is justifiable (like when we infer the bigram *quite late* from the bigram *so happy*) and when it is not (like when we infer *you runs* from *it in*). In other words, we should assign *weights* to the analogical paths in such a way that the path in Figure 5(a) below receives significantly more weight than the path in Figure 5(b). Once we have defined how the weight of an analogical path should be calculated,

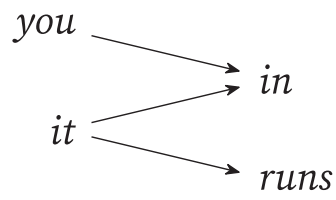

Figure 4. An undesirable analogical path

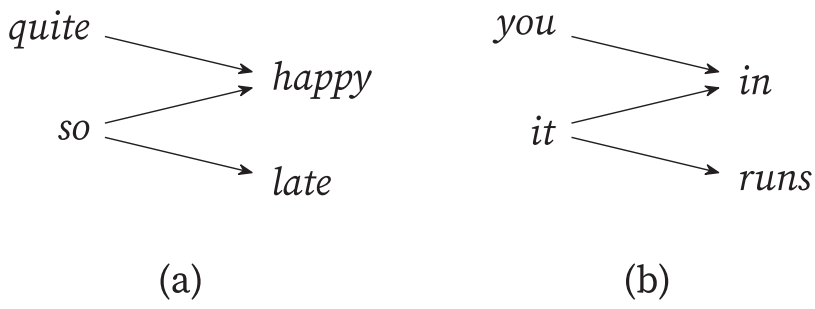

Figure 5. A desirable versus an undesirable analogical path

we can define the analogical probability of a bigram as the sum of the weights of its analogical paths.

My original weighting method was based on the idea that a frequent word like *in* being a common attested right context for two words should not be considered as justification for supposing that these two words in general behave similarly, due to the fact that frequent words, just by virtue of occurring many times, have a higher chance of being a common attested right context of two words that otherwise do not behave similarly. Therefore analogical paths consisting of frequent words should be assigned lower weights.

This method did not work as well as I had hoped. But Márton Makrai (personal communication) suggested another method that also devalues paths made up of frequent words but which is in addition sensitive to the frequencies of the bigrams that constitute the path. His solution worked much better, and importantly he showed that it actually has an intuitive probabilistic interpretation (although the interpretation that I will later present in Subsection 2.8 is slightly different from his original interpretation), so I will now define the weighting using his solution. According to his method the weight of the analogical path in Figure 5(a) – denoted by $P(quite\ late :: so\ happy)$ expressing the analogical probability of *quite late* based on the bigram *so happy* – is given by Equation (1), whose terms are explained below:

$$P(quite\ late :: so\ happy) = P(so\ happy) \cdot P(quite \mid \_\ happy) \cdot P(late \mid so\ \_). \tag{1}$$

Where:

- $P(so\ happy)$ is the *relative frequency* or the *empirical probability* of the bigram *so happy*, given by $\frac{f(so\ happy)}{N}$, where N is the total number of bigrams in the training data;
- $P(quite \mid \_\ happy)$ is the *backward conditional probability* of the bigram *quite happy*, given by $\frac{f(quite\ happy)}{f(happy)}$ and expressing the probability of *quite* preceding *happy*; and
- $P(late \mid so\ \_)$ is the *forward conditional probability* of the bigram *so late*, given by $\frac{f(so\ late)}{f(so)}$ as already defined in Subsection 2.2.

Due to the terms in the numerators and in the denumerators in the formula above, the weight of an analogical path

- increases as the frequencies of the participating bigrams increase (meaning that the model trusts frequently occurring bigrams more as analogical bases), and
- decreases as the size of the training data and the frequencies of the words forming the "middle" bigram increase (meaning that an analogical path is worth less if it is easy to find either due to there being a lot of data available or due to the path being composed of frequent words).

This method will plausibly yield a higher value for $P(quite\ late :: so\ happy)$ than for $P(you\ runs :: it\ in)$, as the conditional probabilities in the former path will likely be higher than in the latter path due to the higher frequencies of *it* and *in* compared to *so* and *happy*. (Although the bigram *it in* will likely be more frequent than the bigram *so happy*, the hope is that the conditional probabilities will make up for this – I have not been able to explain why they would, but based on the performance of the model as described in subsections 2.9–2.10 this weighting works quite well.)

2.6. The (preliminary) analogical path model

Having defined the weights of the analogical paths, the *analogical probability* of the bigram *quite late* is defined as the sum of the weights of its analogical paths, as given by the following equation where w_1 and w_2 range over the set of words occurring in the training data:[7]

$$P_{\mathrm{ANL}}(\textit{quite late}) = \sum_{w_1, w_2} P(\textit{quite late} :: w_1\ w_2). \tag{2}$$

(The next subsection will show that P_{ANL} is actually a probability distribution.) I will call language models assigning probabilities according to Equation (2) instances of the (preliminary) *analogical path model*. (I call it preliminary because in Subsection 2.10 we will make some changes to how this model assigns probabilities, although the basic principle will remain the same.)

We can now also define the *analogical conditional probability* of the bigram *quite late* as

$$P_{\mathrm{ANL}}(\textit{late} \mid \textit{quite}\ \_) = \frac{P_{\mathrm{ANL}}(\textit{quite late})}{P_{\mathrm{ANL}}(\textit{quite})}, \tag{3}$$

where $P_{\mathrm{ANL}}(\textit{quite})$ is the sum of the analogical probabilities of the set of bigrams whose first word is *quite*. (We will show in the next subsection that this is equal to the relative frequency of *quite*.)

Finally to get the analogical probability of a sentence <s> $w_1\ w_2 \ldots w_n$ </s> (where the special symbols <s> and </s> mark the beginning and the end of the sentence), we multiply the analogical conditional probabilities of its bigrams:

$$P_{\mathrm{ANL}}(\text{<s>}\ w_1\ w_2 \cdots w_n\ \text{</s>}) = P_{\mathrm{ANL}}(w_1 \mid \text{<s>}) \cdot P_{\mathrm{ANL}}(w_2 \mid w_1) \cdot \cdots \cdot P_{\mathrm{ANL}}(\text{</s>} \mid w_n), \tag{4}$$

where the underscores after the conditioning words are omitted for shortness.

Let us now turn to some formal properties of analogical probability.

2.7. Formal properties of analogical probability

We will show in this subsection that (a) the analogical probability function P_{ANL} is in fact a probability distribution over bigrams and that (b) the analogical probability $P_{\mathrm{ANL}}(w)$ of any word w is just its relative frequency $P(w)$. We will now prove (a) by showing that the analogical probabilities of all bigrams sum to 1 and then obtain (b) as a corollary.

The sum of the analogical probabilities of all bigrams can be written as follows, with x_1, x_2, w_1, w_2 ranging over the words occurring in the training data (the pairs $x_1\ x_2$ represent the bigrams for which we are trying find analogical paths, and the pairs $w_1\ w_2$ represent the bigrams through which an analogical path may connect x_1 and x_2):

$$\sum_{x_1,\, x_2,\, w_1,\, w_2} P(w_1\ w_2) \cdot P(x_1 \mid \_\ w_2) \cdot P(x_2 \mid w_1\ \_).$$

We move the summation over x_2 inside to get

[7]We do not need to require $w_1\ w_2$ to be an attested bigram because in the case of an unattested bigram $w_1\ w_2$ the expression $P(x_1\ x_2 :: w_1\ w_2)$ always evaluates to 0.

$$\sum_{x_1,\, w_1,\, w_2} P(w_1\, w_2)\cdot P(x_1 \mid \_\, w_2)\cdot \sum_{x_2} P(x_2 \mid w_1\, \_),$$

which by the definition of conditional probability reduces to just

$$\sum_{x_1,\, w_1,\, w_2} P(w_1\, w_2)\cdot P(x_1 \mid \_\, w_2).$$

We now move the summation over w_1 inside to get

$$\sum_{x_1,\, w_2} \sum_{w_1} P(w_1\, w_2)\cdot P(x_1 \mid \_\, w_2),$$

which is equal to the leftmost formula below, which we reduce with two equalities to finish the proof:

$$\sum_{x_1,\, w_2} P(w_2)\cdot P(x_1 \mid \_\, w_2) = \sum_{x_1, w_2} P(x_1\, w_2) = 1.$$

As a corollary, note that if we fix the word x_1 in the formula with which the proof above begins (i.e. if we try to express the sum of the probabilities of only those analogical paths that start from the word x_1) then in the last step we get $\sum_{w_2} P(x_1\, w_2) = P(x_1)$, so we can also conclude that $P_{\mathrm{ANL}}(w) = P(w)$ for any word w.

We have seen that P_{ANL} indeed assigns probabilities to bigrams. But probabilities of what?

2.8. Interpreting analogical probability

In order to see what is expressed by analogical probability, consider again the analogical path from *quite* to *late* through *happy* and *so*, as in Figure 6. Suppose that we wanted to generate such paths with a bigram model similarly to how we can generate a word sequence. What would be the probability with which we could generate the path in Figure 6 from *quite* to *late* as a sequence of forward and backward steps along attested bigrams?

My intuition would suggest to use the chain rule of probability along with the "Markov assumption" that the probability of the upcoming step depends only on the previous one to calculate this probability: the first word *quite* would be generated according to its relative frequency *P*(*quite*), and the upcoming steps would always be generated with the conditional probability conditioned on the previous word. Forward steps (e.g. from *quite* to *happy*) would be generated according to forward conditional probability and backward steps (e.g. from *happy* to *so*) according to backward conditional probability. The path above would then be generated with the probability

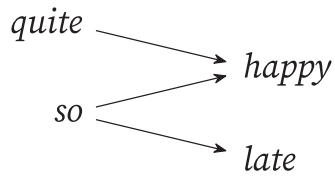

Figure 6. An analogical path

$$P(quite) \cdot P(happy \mid quite\ \_) \cdot P(so \mid \_\ happy) \cdot P(late \mid so\ \_).$$

But this can be shown to be equal to the weight that we defined for this analogical path in Equation (1). We first switch the conditioning word *quite* and the goal word *happy* in the first two terms by the definition of conditional probability to get

$$P(happy) \cdot P(quite \mid \_\ happy) \cdot P(so \mid \_\ happy) \cdot P(late \mid so\ \_)$$

then switch the middle two terms to get

$$P(happy) \cdot P(so \mid \_\ happy) \cdot P(quite \mid \_\ happy) \cdot P(late \mid so\ \_)$$

and multiply the first two terms to arrive at

$$P(so\ happy) \cdot P(quite \mid \_\ happy) \cdot P(late \mid so\ \_),$$

which is exactly $P(quite\ late :: so\ happy)$ as defined in Equation (1).

The weight of an analogical path is then nothing but the probability of generating it with a bigram model, and the analogical probability of a bigram is the probability of generating some three-step "forth-back-forth" analogical path from its first word to its second word.

But what should we do in the case of attested bigrams, which we could generate according to their relative frequencies in a single step without resorting to "forth-back-forth" analogical paths? Should we not let these single-step "forth" paths also contribute to the probability of a bigram? It would seem natural to take them into account too, especially because then the analogical path model could be considered to be a generalisation of the basic "maximum likelihood estimate" bigram model which uses only these single-step paths.

This interpretation came to my mind too late in the process of publishing this article to be able to really incorporate this idea into the model but in Subsection 2.10 we will see that it would most likely be a good idea to allow for such one-step analogical paths as well.[8]

Let us now see how well this model works. In the next subsection we will take a look at how well it can identify the intuitively plausible analogies and in Subsection 2.10 we will compare its performance in predicting sentences to some other methods of generalising over bigrams.

2.9. Intuitive testing

In this subsection we will examine whether the analogical path model can really differentiate between "good" and "bad" analogies; i.e. whether the analogical paths that receive the biggest weights from the model seem intuitively correct analogical bases from a linguist's point of view (although the conclusions of this subsection are based only on *my* intuition – the reader is of course free to disagree with my assessments).

I implemented the analogical path model as a computer program in Python 3 and trained it on the Project Gutenberg edition of Grimm's Fairy Tales (Grimm & Grimm 2001), consisting of 109,862 words. I then wrote a program that receives a bigram as input and outputs the ten best and ten worst analogical paths (ordered by weight) for this bigram according to the model. I then chose three unattested bigrams for the model to analyse. I present the results in Table 2,

[8]Another possibility would be to allow for "forth-back-forth-back-forth" paths or even "forth-forth-back-back-forth" paths etc. but this would make it very computationally expensive to find all analogical paths so I have not experimented with it.

Table 2. 10 best and 10 worst analogical paths for three unattested bigrams (the word *'s* is the contraction e.g. in the word *John's*)

	quite early		her geese		ran towards	
	A	B	A	B	A	B
Top 10 paths	*very*	*angry*	*their*	*father*	*went*	*away*
	very	*sorrowful*	*'s*	*head*	*sprang*	*out*
	very	*sorry*	*my*	*father*	*jumping*	*about*
	very	*happy*	*their*	*own*	*striding*	*about*
	very	*tired*	*my*	*own*	*striding*	*up*
	very	*high*	*my*	*husband*	*went*	*out*
	very	*pleased*	*my*	*head*	*swimming*	*about*
	very	*merry*	*my*	*mother*	*sprang*	*up*
	me	*kindly*	*their*	*mother*	*went*	*to*
	me	*sorry*	*the*	*little*	*went*	*home*
Bottom 10 paths	*,*	*the*	*the*	*full*	*went*	*,*
	<s>	*down*	*of*	*so*	*and*	*about*
	,	*round*	*'s*	*not*	*<s>*	*a*
	,	*red*	*'s*	*as*	*sprang*	*,*
	,	*a*	*the*	*daughter*	*went*	*a*
	me	*the*	*'s*	*and*	*and*	*upon*
	so	*</s>*	*my*	*will*	*came*	*</s>*
	,	*down*	*the*	*long*	*but*	*,*
	so	*a*	*of*	*when*	*and*	*for*
	,	*out*	*the*	*will*	*and*	*,*

with the analogical paths labelled by the attested bigram through which the words of the unattested bigram are connected.

Looking at the results, it is apparent that the model mostly assigns those analogical paths the biggest weights that would also be judged to be justifiable analogical bases by linguists (at least if their linguistic intuition matches mine) and vice versa: the bigrams forming the top ten paths tend to be constituted by words that are intuitively similar in their syntactic behaviour to the first and second words of the three unattested bigrams in the top row, and the bigrams forming the bottom ten paths tend to contain at least one word (or comma or sentence marker) that has little to do with the word to which it is supposed to behave similarly.

The model in fact sees quite a large difference between the best and the worst paths: among the top ten and bottom ten paths, the weights of the n-th best top paths tend to be between two to three orders of magnitude bigger than the weights of the n-th best bottom paths.

(The three unattested bigrams in Table 2 were somewhat purposefully selected to make a good first impression. There are unattested bigrams for which the analogical paths that the model finds do not match my linguistic intuition to this extent – although this might not be such a bad thing, as I will note regarding the bigram *quite sad* in Subsection 3.3. But overall the model did seem to be able to identify the intuitively most relevant attested bigrams.)

We will now compare the performance of the analogical path model in predicting English sentences with the performance of already existing methods for generalising over bigrams.

2.10. Comparative testing

The standard measure of the performance of language models is called "perplexity", which measures how well a model predicts a set of novel sentences (called the *test data*) in the language of the training data: the model that assigns the highest probabilities to such sentences will have the lowest perplexity score (for an exact definition see Jurafsky & Martin 2023, 8) and will be considered to be the best model. This measure is what we will use to compare the analogical path model to already existing bigram models, again using the Grimm's Fairy Tales corpus.

2.10.1. Amending the analogical path model. But before carrying out the tests we will have to change the way that the analogical path model estimates e.g. the probability of the word *late* occurring after the word *quite*. So far we have defined this estimate to be just the analogical conditional probability $P_{\mathrm{ANL}}(late \mid quite\ \_)$. But this will not work as is, because there will likely always be bigrams for which the model does not find any analogical paths. In one particular case in which I counted the bigrams, there were around 34,000 distinct attested bigrams in the training data and around 21 million distinct bigrams that could be combined from the words that occurred in the training data, and of these the model found analogical paths for around 15 million bigrams – which I found to be a surprisingly good result, but still problematic because any one of the 6 million zero-probability bigrams could potentially occur in the test data and would cause the analogical path model to have infinitely large perplexity.

Therefore in these cases we will have to rely on just the relative frequency or *unigram probability* of the second word *late* in order to estimate its probability. Specifically we will use a technique called *interpolation*, which amounts to always taking a weighted average of the analogical conditional probability of *quite late* and the unigram probability of *late* to estimate the probability of *late* being the upcoming word after *quite*.

And finally, as I alluded to at the end of Subsection 2.8, we will also incorporate into this weighted average the basic "maximum likelihood estimate" (MLE) conditional probability of *quite late*. I originally did this just because it improved the performance of the model, but the interpretation of analogical probability presented in Subsection 2.10 provides some theoretical justification to this addition.

The final formula for estimating the probability of a word x_2 occurring after a word x_1 with the analogical path model is then

$$P_{\mathrm{APM}}(x_2 \mid x_1\ \_) = 0.595 \cdot P(x_2 \mid x_1\ \_) + 0.4 \cdot P_{\mathrm{ANL}}(x_2 \mid x_1\ \_) + 0.005 \cdot P(x_2), \tag{5}$$

in which I set the weights by experimenting with different settings and choosing the one that achieved the best perplexity score.

2.10.2. The other models. The already existing bigram models to which I compared the analogical path model are the following:

- "baseline" smoothing, in which we just interpolate the MLE conditional probability (with weight 0.75) and the unigram probability (with weight 0.25), and
- interpolated Kneser-Ney smoothing (with absolute discount 0.75), which is described in Jurafsky & Martin (2023, 17–20).

I set the parameters of these models in the same way as the parameters of the analogical path model: finding the best setting by manual experimentation.

2.10.3. The tests and the results. I implemented the baseline smoothing model in Python 3, and for Kneser-Ney smoothing I used the Natural Language Toolkit ("NLTK", Bird, Loper & Klein 2009) implementation. I carried out ten-fold cross-validation on three subsets of the Grimm corpus: one containing around 25,000 words, another containing around 50,000 words, and the full corpus (containing around 100,000 words).[9] The results are summarised in Table 3 below, in which I took the averages of the perplexity scores over the ten tests.

The analogical path model just barely beats Kneser-Ney smoothing. The tests would have to be carried out on larger corpora to really be able to tell which of these two models is better, but I consider this result to be a positive one – interpolated Kneser-Ney smoothing is one of the best performing smoothing methods according to Jurafsky & Martin (2023, 17).

3. THEORETICAL RELEVANCE AND PROSPECTS

This section discusses the theoretical relevance of the analogical path model. In Subsection 3.1 I interpret this model as a proposal for how constructions can be learned and represented without syntactic categories, in Subsection 3.2 I argue that this model provides a way to overcome a problem that has been used to argue against the feasibility of usage-based theories of language, and in Subsection 3.3 I show how this model may be used to arrive at counter-intuitive explanations to linguistic phenomena.

Table 3. Average perplexity scores of three bigram models across varying corpus sizes

	25 K	50 K	100 K
Baseline	165.27	142.62	126.40
Kneser-Ney	141.74	124.16	112.77
Analogical path	144.48	122.85	111.11

[9]In the future I would like to carry out these tests on larger corpora; time constraints prevented me from being able to do so for this article.

3.1. Emergent constructions

At the end of the first section I promised to give content to the view that constructions emerge from linguistic experience and to show that constructions can be represented holistically and without referring to syntactic categories. I will now argue that the analogical path model fulfils these promises: the notions of "linguistic experience", "construction" and "emergence" can be naturally defined in terms of this model.

In the analogical path model, *linguistic experience* is taken to be a network made up of (a) the words we hear and (b) the arrows that connect those words that we observe to occur next to each other.

To understand what a *construction* is in this model it is best to consider what constructions are used for by human speakers: constructions enable us to interpret sentences that we have not heard before. In the analogical path model the things that enable the model to predict bigrams that it has not seen before are the analogical paths that are formed by the arrows of the network. Therefore the definition of a construction in this model will have to be based on analogical paths. (And this definition will only approximate half of the notion of "construction" in the "form–function association" sense as used in construction grammar, since no function is explicitly associated to bigrams by the analogical path model.)

Consider for example the construction that would traditionally be represented as "[Adjective + Noun]", intended to express bigrams like *calm weather*, *strange elephant* and *excellent performance*. In the analogical path model this construction could be partially represented as a collection of analogical paths from *calm* to *weather*, from *strange* to *elephant*, and from *excellent* to *performance*. This is illustrated in Figure 7 below.

This figure also suggests that a construction should not be defined as just any arbitrary collection of analogical paths. What makes a collection of analogical paths a construction is that its paths contribute a lot of analogical probability to sets of bigrams with a high degree of overlap, just like how constructions in construction grammar are assumed to allow speakers to interpret many similar sentences. What I mean by "overlap" is illustrated in Figure 7: the attested bigram *nice person* is part of analogical paths from *calm* to *weather* and from *strange* to *elephant*, and the attested bigram *good book* is part of analogical paths from *strange* to *elephant* and from *excellent* to *performance*. The bigram *strange elephant* is supported by both *nice person* and *good book*; if many other bigrams are also "co-supported" by them, then these two attested

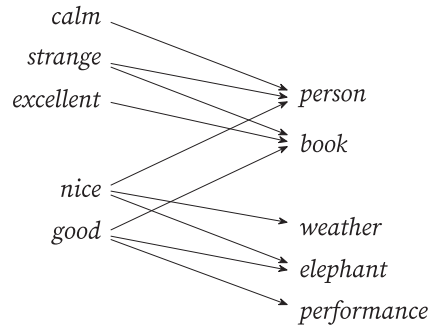

Figure 7. (Part of) the "[Adjective + Noun]" construction

bigrams can be considered to form analogical paths that are part of the same construction, and this in turn means that the non-co-supported bigrams *calm weather* and *excellent performance* can also be thought of as instances of this same construction.

A collection of analogical paths may then be said to form a construction to the extent that its paths contribute a lot of probability mass to sets of bigrams with a high degree of overlap. What is considered to be a "lot" or a "high degree" can be defined however we like in order to get a categorical definition. We might set a particular amount of analogical probability that the paths must contribute in the aggregate, and we might require that for each path in the collection there must be at least n other paths in the collection that contribute to at least m bigrams that the given path also contributes to. But I am not convinced that it is even necessary to draw a categorical boundary between constructions and non-constructions.

Finally, constructions are *emergent* in the sense that they are not considered to be units of grammar; they are nothing but holistic properties of the network of linguistic experience. And grammar itself is emergent because there is no separate module that contains grammatical constructions or rules. In particular, no reference to syntactic categories is necessary: the only things that need to be taken into account are the raw distributional properties of words.

3.2. Generalisation through memorisation

The relative success of even such a simple model is notable from the point of view of usage-based theories like (most forms of) construction grammar because it indicates that it is possible for general syntactic patterns to emerge from the memorisation of concrete word sequences without formulating rules with which to refer to abstract syntactic categories.

As I noted in Subsection 2.2, an important general property of human languages from the point of view of language acquisition is that most of the possible word combinations occur very infrequently even in large corpora. This fact has been brought up as evidence against usage-based models of linguistic competence where the memorisation and retrieval of specific word combinations is assumed to have a key role in producing and interpreting sentences. According to Yang (2010), the high degree of infrequency among word combinations "hints at inherent limitations in approaches that stress the storage of construction-specific rules or processes" (p. 18). Adger (2013), citing Yang, claims that usage-based models that try to learn constructions by generalising from memorised word combinations will never work, "simply because there is not enough data in the input to give evidence for generalisation" (p. 473).

Although the analogical path model is by no means a complete model of language acquisition, it does suggest that general patterns can in fact be learned from specific word combinations, if these word combinations are used appropriately as analogical bases.

3.3. Explanation

Consider the ten best analogical paths for the bigram *quite sad*, again based on the Grimm corpus; see Table 4. The words *very* and *so* appear in the "A" column among the top ten paths, as expected; what is less expected is that the words *grew* and *felt* and also appear multiple times. At first sight this may look like an undesirable outcome: our intuition says that these words do not behave like the word *quite* because they are verbs and *quite* is an adverb, so it would be better if they did not appear among the top paths.

Table 4. 10 best analogical paths for the bigram *quite sad*

quite sad	
A	**B**
very	*angry*
very	*sorrowful*
grew	*angry*
felt	*tired*
grew	*late*
felt	*another*
very	*sorry*
grew	*dark*
so	*loudly*
so	*happy*

But lumping these words together with all verbs hides their particular distributional properties that could explain why they might be justifiable analogical bases. Although I have not statistically analysed their distributions, it is very easily imaginable that *grew* and *felt* are more likely to precede adjectives than other verbs like e.g. *ran*, *bought* and *read*, by virtue of the former verbs' ability to occur in constructions like *Mary grew angry* and *John felt tired*, in which the adjective is predicated of the subject of the verb. (The latter verbs only have this ability in resultative constructions like *Carla ran herself tired*). They are thus similar to adverbs in their tendency to occur before adjectives, and therefore it is reasonable to use them as analogical bases when trying to decide if an adjective like *sad* can occur after an adverb like *quite*.

What we see here (if the reasoning above is correct) is that analysing the way that the analogical path model works can offer counter-intuitive explanations of why certain word sequences are well-formed – explanations that we may never have thought about without consulting the model. Alongside the intuitive and comparative tests I described at the end of the previous section, I believe this to be another important measure of the goodness of a language model: whether it can offer novel insight into linguistic phenomena.[10] (But of course to really be able to gain insight into language through the analogical path model, it will have to be developed to be able to recognise more complex syntactic patterns than those composed of two words.)

This is why it is worth developing language models such as the analogical path model: as this model always selects a subset of the training data as analogical basis when predicting a bigram, the predictions that this model makes can be interpreted as statements about the kinds of linguistic experience that make a bigram recognisable as an instance of some linguistic pattern, getting us closer to a (synchronic) explanation of the pattern.

[10]Similarly to how Skousen's (1989) Analogical Model is claimed by Skousen to explain the exceptional past tense form of the Finnish verb *sorta-* ("oppress") (Skousen 2002, 24).

ACKNOWLEDGEMENT

I thank Márton Makrai, Márton Gömöri, and the two anonymous reviewers for their contributions to this article.

REFERENCES

Adger, David. 2013. Constructions and grammatical explanation: Comments on Goldberg. Mind & Language 28(4). 466–478. https://doi.org/10.1111/mila.12027.

Angluin, Dana. 1982. Inference of reversible languages. Journal of the ACM 29(3). 741–765. https://doi.org/10.1145/322326.322334.

Berwick, Robert C. and Sam Pilato. 1987. Learning syntax by automata induction. Machine Learning 2. 9–38. https://doi.org/10.1023/A:1022860810097.

Bird, Steven, Edward Loper and Ewan Klein. 2009. Natural language processing with Python. O'Reilly Media Inc. https://www.nltk.org/book/.

Croft, William. 2001. Radical construction grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001.

Croft, William. 2005. Logical and typological arguments for radical construction grammar. In J. Östman and M. Fried (eds.) Construction grammars: Cognitive grounding and theoretical extensions. Amsterdam & Philadelphia, PA: John Benjamins Publishing Company. 273–310. https://doi.org/10.1075/cal.3.

Crystal, David. 1966. English. Lingua 17(1–2). 24–56. https://doi.org/10.1016/0024-3841(66)90003-9.

Fillmore, Charles J., Paul Kay and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. Language 64(3). 501–538. https://doi.org/10.2307/414531.

Goldberg, Adele and Thomas Herbst. 2021. The nice-of-you construction and its fragments. Linguistics 59(1). 285–318. https://doi.org/10.1515/ling-2020-0274.

Grimm, Jacob and Wilhelm Grimm. 2001. Grimms' fairy tales. Urbana, IL: Project Gutenberg. Retrieved February 9, 2023, from https://www.gutenberg.org/ebooks/2591.

Jurafsky, Dan and James H. Martin. 2023. N-gram language models. In D. Jurafsky and J. H. Martin (eds.) Speech and language processing (Ch. 3), 3rd edn. Draft. https://web.stanford.edu/~jurafsky/slp3/.

Skousen, Royal. 1989. Analogical modeling of language. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-1906-8.

Skousen, Royal. 2002. An overview of analogical modeling. In R. Skousen, D. Lonsdale and D. B. Parkinson (eds.) Analogical modeling: An exemplar-based approach to language. Amsterdam & Philadelphia, PA: John Benjamins Publishing Company. 11–27. https://doi.org/10.1075/hcp.10.

Tomasello, Michael 2003. Constructing a language: A usage-based theory of language acquisition. Harvard University Press. https://doi.org/10.2307/j.ctv26070v8.

Yang, Charles. 2010. Who's afraid of George Kingsley Zipf? Unpublished manuscript. University of Pennsylvania, Philadelphia, PA. https://www.ling.upenn.edu/~ycharles/papers/zipfnew.pdf.