What kind of grammar do LSTMs learn?

An experiment of *recurrent babbling*

Ludovica Pannitto, Aurélie Herbelot

December 18, 2020 @ BrownBag Meetings

CIMeC - University of Trento

Do RNNs learn grammar?

A popular question, relating to **productivity** and **compositionality**[1].

We propose that the evaluation of RNN grammars should be widened to include:

- the effect of the **type of input data** fed to the network
- the **theoretical paradigm** used to analyse its performance

[1] "Linguistic generalization and compositionality in modern artificial neural networks" (Baroni 2020)

Do RNNs learn grammar?

A popular question, relating to **productivity** and **compositionality**[1].

We propose that the evaluation of RNN grammars should be widened to include:

- the effect of the **type of input data** fed to the network
- the **theoretical paradigm** used to analyse its performance

[1]"Linguistic generalization and compositionality in modern artificial neural networks" (Baroni 2020)

Do RNNs learn grammar?

A popular question, relating to **productivity** and **compositionality**[1].

We propose that the evaluation of RNN grammars should be widened to include:

- the effect of the **type of input data** fed to the network
- the **theoretical paradigm** used to analyse its performance

[1] "Linguistic generalization and compositionality in modern artificial neural networks" (Baroni 2020)

How much language (L) can be learnt from a certain level of computational complexity (C) with a certain type of data (I)?

$$C \times I \xrightarrow{f} L \tag{1}$$

All aspects of the equation are of paramount importance in linguistic discussion:

complexity of the learning mechanism C - how much has to be *innate* or *hard-coded* in the function?

quality and quantity of the stimuli I - how do stimuli differ? What are the most relevant features?

language L - what is the *grammar* that best explains the language we experience?

How much language (L) can be learnt from a certain level of computational complexity (C) with a certain type of data (I)?

$$C \times I \xrightarrow{f} L \tag{1}$$

All aspects of the equation are of paramount importance in linguistic discussion:

complexity of the learning mechanism C - how much has to be *innate* or *hard-coded* in the function?

quality and quantity of the stimuli I - how do stimuli differ? What are the most relevant features?

language L - what is the *grammar* that best explains the language we experience?

Contents

1. Recurrent babbling: setup

2. Main questions

3. Results

4. Where to go next

Recurrent babbling: setup

Our Setup

- vanilla **char-LSTM** trained on a limited amount of **child-directed language**
- introduce a methodology to evaluate the **distribution of grammatical items**, focusing on the network's generated output - its *babbling*
- explore the **interaction** between meaning representations and the abstraction abilities of the network

The study is conducted on **English**.

Our Setup

- vanilla **char-LSTM** trained on a limited amount of **child-directed language**
- introduce a methodology to evaluate the **distribution of grammatical items**, focusing on the network's generated output - its *babbling*
- explore the **interaction** between meaning representations and the abstraction abilities of the network

The study is conducted on **English**.

Our Setup

- vanilla **char-LSTM** trained on a limited amount of **child-directed language**
- introduce a methodology to evaluate the **distribution of grammatical items**, focusing on the network's generated output - its *babbling*
- explore the **interaction** between meaning representations and the abstraction abilities of the network

The study is conducted on **English**.

Our Setup

- vanilla **char-LSTM** trained on a limited amount of **child-directed language**
- introduce a methodology to evaluate the **distribution of grammatical items**, focusing on the network's generated output - its *babbling*
- explore the **interaction** between meaning representations and the abstraction abilities of the network

The study is conducted on **English**.

C: Computational complexity

We train a char-LSTM on some input I_i, varying in a specific range, and make the network produce some amount of language λ_i.

$$(\text{LSTM}, I_i) \xrightarrow{a} \lambda_i \tag{2}$$

Nativist theories typically posit the need for a dedicated device for language learning while cognitive theories have argued that **general purpose memory and cognitive mechanisms** can account for the emergence of linguistic abilities.

LSTMs can be seen as domain-general attention and memory mechanisms, without any explicitly hard-coded grammatical knowledge.

I: A different input

ANNs are often trained on an input that is unrealistic in **genre** and **size**.

- child-directed language is characterized by **specific features** (e.g., *repetitiousness*) that are not present in the most widely used corpora
- it has been estimated that, by the age of 3, welfare children have heard about 10 millions words while the average working-class child has heard around 30 millions, and the variation depends on **many factors**

We evaluate three different language sources: **CHILDES**, **OpenSubtitles** (movie and TV series subtitles) and **Simple English Wikipedia**.

I: A different input

ANNs are often trained on an input that is unrealistic in **genre** and **size**.

- child-directed language is characterized by **specific features** (e.g., *repetitiousness*) that are not present in the most widely used corpora
- it has been estimated that, by the age of 3, welfare children have heard about 10 millions words while the average working-class child has heard around 30 millions, and the variation depends on **many factors**

We evaluate three different language sources: **CHILDES**, **OpenSubtitles** (movie and TV series subtitles) and **Simple English Wikipedia**.

I: A different input

ANNs are often trained on an input that is unrealistic in **genre** and **size**.

- child-directed language is characterized by **specific features** (e.g., *repetitiousness*) that are not present in the most widely used corpora
- it has been estimated that, by the age of 3, welfare children have heard about 10 millions words while the average working-class child has heard around 30 millions, and the variation depends on **many factors**

We evaluate three different language sources: **CHILDES**, **OpenSubtitles** (movie and TV series subtitles) and **Simple English Wikipedia**.

L: Features of generated language

Catenae[2], are characterized as fundamental **meaning-bearing units**, in line with the theoretical tenets of constructionist theories[3], thus being ideal candidates for populating our lexicon (or *Constructicon*).

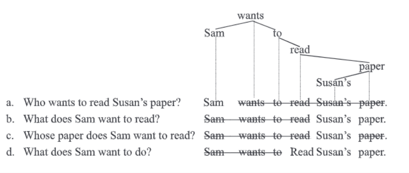

Figure 1: Image showing cases of ellipses, from *Constructions are catenae: Construction grammar meets dependency grammar* (Osborne and Groß 2012)

---

[2] "Catenae: Introducing a novel unit of syntactic analysis" (Osborne, Putnam, and Groß 2012)

[3] *Constructions at work: The nature of generalization in language* (Goldberg 2006)

L: Features of generated language

Catenae[2], are characterized as fundamental **meaning-bearing units**, in line with the theoretical tenets of constructionist theories[3], thus being ideal candidates for populating our lexicon (or *Constructicon*).

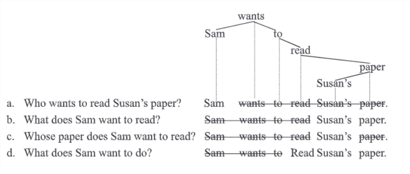

Figure 1: Image showing cases of ellipses, from *Constructions are catenae: Construction grammar meets dependency grammar* (Osborne and Groß 2012)

[2] "Catenae: Introducing a novel unit of syntactic analysis" (Osborne, Putnam, and Groß 2012)
[3] *Constructions at work: The nature of generalization in language* (Goldberg 2006)

Catenae

Definition of *Catena*:

"a word, or a combination of words which is continuous with respect to dominance"

Figure 2: Dependency representation for the sentence: *Mary had a little lamb*

- Mary had lamb
- had a lamb
- little lamb
- Mary had NOUN
- nsubj VERB dobj

The number and composition of *catenae* depends on **how elements are arranged** in the structure of the dependency tree.

Catenae

Definition of *Catena*:

"a word, or a combination of words which is continuous with respect to dominance"

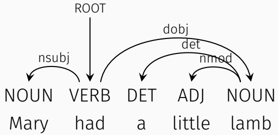

Figure 2: Dependency representation for the sentence: *Mary had a little lamb*

- Mary had lamb
- had a lamb
- little lamb
- Mary had NOUN
- nsubj VERB dobj

The number and composition of *catenae* depends on **how elements are arranged** in the structure of the dependency tree.

Catenae

Definition of *Catena*:

"a word, or a combination of words which is continuous with respect to dominance"

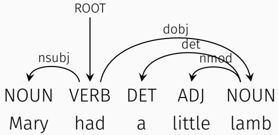

Figure 2: Dependency representation for the sentence: *Mary had a little lamb*

- Mary had lamb
- had a lamb
- little lamb
- Mary had NOUN
- nsubj VERB dobj

The number and composition of *catenae* depends on **how elements are arranged** in the structure of the dependency tree.

Main questions

Main questions

Q1: To what extent is the network able to generate **new** language?

- We expect the network to reproduce the **statistical regularities** of the input, we further investigate what kind of regularities are acquired and how do the language models differ.

Q2: On what conditions is the network able to generalize its *grammatical* knowledge?

- We can state that the network has learned some grammar once it is able to use an acquired pattern in a **productive** and **creative** way.
- We expect this generalization ability to evolve during training and the **distributional properties** of patterns to be in relation with the grammatical abilities of the network at various stages of learning.

Main questions

Q1: To what extent is the network able to generate **new** language?

- We expect the network to reproduce the **statistical regularities** of the input, we further investigate what kind of regularities are acquired and how do the language models differ.

Q2: On what conditions is the network able to generalize its *grammatical* knowledge?

- We can state that the network has learned some grammar once it is able to use an acquired pattern in a **productive** and **creative** way.
- We expect this generalization ability to evolve during training and the **distributional properties** of patterns to be in relation with the grammatical abilities of the network at various stages of learning.

Main questions

Q1: To what extent is the network able to generate **new** language?

- We expect the network to reproduce the **statistical regularities** of the input, we further investigate what kind of regularities are acquired and how do the language models differ.

Q2: On what conditions is the network able to generalize its *grammatical* knowledge?

- We can state that the network has learned some grammar once it is able to use an acquired pattern in a **productive** and **creative** way.
- We expect this generalization ability to evolve during training and the **distributional properties** of patterns to be in relation with the grammatical abilities of the network at various stages of learning.

Pipeline

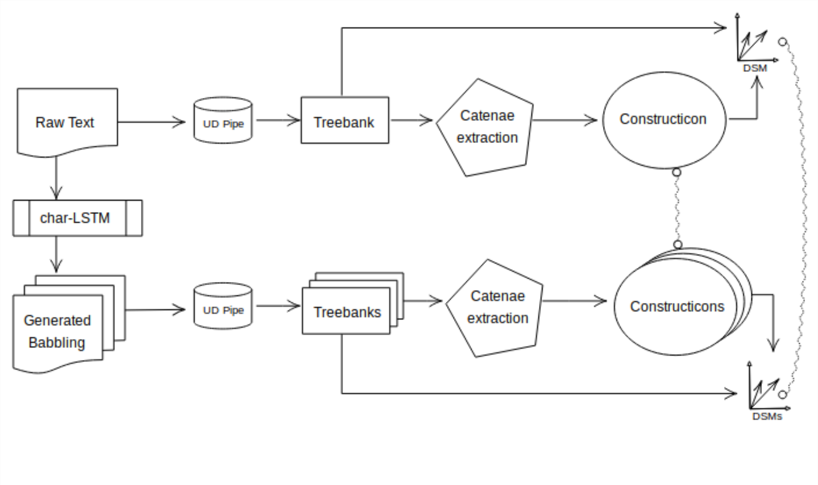

Figure 3: A summary of the work pipeline

Catenae extraction

catena	frequency	mi
largest mi		
@nsubj @root	294.59K	633.93K
_DET _NOUN	189.97K	552.32K
_VERB @obj	190.72K	520.82K
_PRON _VERB	271.44K	503.17K
@nsubj _AUX @root	129.60K	478.86K
smallest mi		
_PRON @nsubj	17.50K	-35.54K
@root @nsubj	27.61K	-34.89K
@nsubj _PRON	11.63K	-30.47K
_VERB @nsubj	12.79K	-26.82K
_AUX _PRON	15.75K	-26.67K

Table 1: Examples of catenae extracted from CHILDES. Largest and smallest mutual information are reported, in top and bottom tier of the table respectively.
Part of Speech are prefixed by "_" and syntactic relations are prefixed by "@"

Results

Q1: What do ANNs approximate?

We evaluated **Spearman** ρ among the top 10K catenae extracted from the input and from each *babbling* stage produced by the LSTM.

Our analysis shows that the network has acquired statistical **regularities at the level of grammatical patterns**, and is able to use them productively to generate **novel** language fragments that **adhere to the same distribution as the input**.

Catenae extracted from babblings almost perfectly correlate with those extracted from the same input, but correlation values are quite **loose for out-of-domain pairs**.

Q1: What do ANNs approximate?

We evaluated **Spearman** ρ among the top 10K catenae extracted from the input and from each *babbling* stage produced by the LSTM.

Our analysis shows that the network has acquired statistical **regularities at the level of grammatical patterns**, and is able to use them productively to generate **novel** language fragments that **adhere to the same distribution as the input**.

Catenae extracted from babblings almost perfectly correlate with those extracted from the same input, but correlation values are quite **loose for out-of-domain pairs**.

Q1: What do ANNs approximate?

We evaluated **Spearman** ρ among the top 10K catenae extracted from the input and from each *babbling* stage produced by the LSTM.

Our analysis shows that the network has acquired statistical **regularities at the level of grammatical patterns**, and is able to use them productively to generate **novel** language fragments that **adhere to the same distribution as the input**.

Catenae extracted from babblings almost perfectly correlate with those extracted from the same input, but correlation values are quite **loose for out-of-domain pairs**.

Q1: What do ANNs approximate?

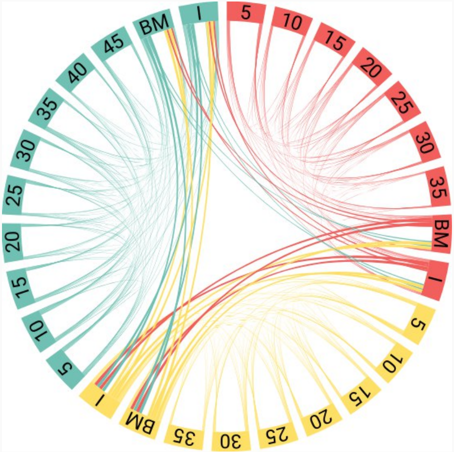

Figure 4: The thickness of the connections is **inversely** proportional to correlation. OpenSubtitles is shown in green on the left of the plot, CHILDES in red in the top right and Simple Wikipedia in yellow at the bottom.

Q2: Meaning and abstraction

The case of *[SBJ V OBJ OBJ2]* [4]

The meaning of the ditransitive pattern emerges from its strong association with ***give*** in child-directed speech: part of the meaning of *give* remains attached to the construction.

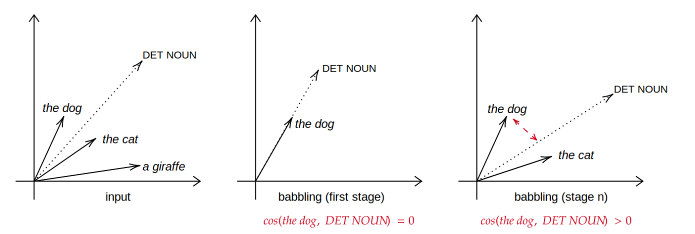

Figure 5: The network is supposed to capture stereotypical instances at early stages of learning and the productivity of the pattern will increase during training

[4] *Constructions at work: The nature of generalization in language* (Goldberg 2006)

Q2: Meaning and abstraction

The case of *[SBJ V OBJ OBJ2]* [4]

The meaning of the ditransitive pattern emerges from its strong association with ***give*** in child-directed speech: part of the meaning of *give* remains attached to the construction.

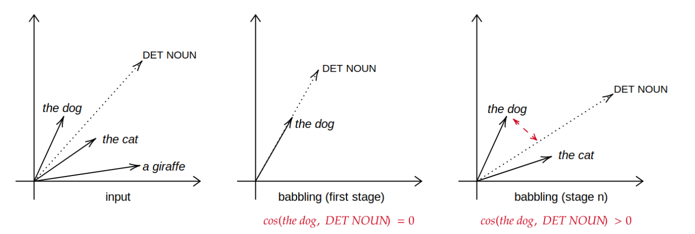

Figure 5: The network is supposed to capture stereotypical instances at early stages of learning and the productivity of the pattern will increase during training

[4] *Constructions at work: The nature of generalization in language* (Goldberg 2006)

Q2: Meaning and abstraction

cat_1	cat_2	input	5	10	...	30	35	shift
a minute	a _NOUN	0.28	0.71	0.51	...	0.37	0.34	0.37
a minute	a @root	0.13	0.49	0.37	...	0.22	0.20	0.30
you _VERB it	_PRON @root @expl	0.10	0.46	0.28	...	0.17	0.21	0.25
you _VERB you	you _VERB @iobj	0.28	0.68	0.56	...	0.42	0.43	0.25
we can _VERB	_PRON can @root	0.51	0.79	0.74	...	0.61	0.57	0.22

Table 2: Pairs of catenae (cat_1, cat_2), their cosine similarity in the space obtained from CHILDES and in the spaces obtained from intermediate *babbling* stages.
The last column shows the difference between cosine similarity at epoch 5 and cosine similarity at epoch 35.

Q2: Meaning and abstraction

Hypotheses:

- pairs with very **high input similarity** are unlikely to exhibit abstraction: the *catena* that is part of the *Constructicon* is the least abstract one, and there is **no need** for the more abstract category - i.e., non productive idioms like *talk through your hat* vs. *talk through your N*
- **low similarity** pairs, on the other hand, may simply contain **unrelated *catenae*** - i.e., too generic associations, like *the dog* vs *DET NOUN*

Instead, given pairs (cat_1, cat_2) with cat_1 being a less abstract instance of cat_2, we expect the highest shifts to happen at **intermediate levels of similarities in the input distributional space**.

Q2: Meaning and abstraction

Hypotheses:

- pairs with very **high input similarity** are unlikely to exhibit abstraction: the *catena* that is part of the *Constructicon* is the least abstract one, and there is **no need** for the more abstract category - i.e., non productive idioms like *talk through your hat* vs. *talk through your N*
- **low similarity** pairs, on the other hand, may simply contain **unrelated *catenae*** - i.e., too generic associations, like *the dog* vs *DET NOUN*

Instead, given pairs (cat_1, cat_2) with cat_1 being a less abstract instance of cat_2, we expect the highest shifts to happen at **intermediate levels of similarities in the input distributional space**.

Q2: Meaning and abstraction

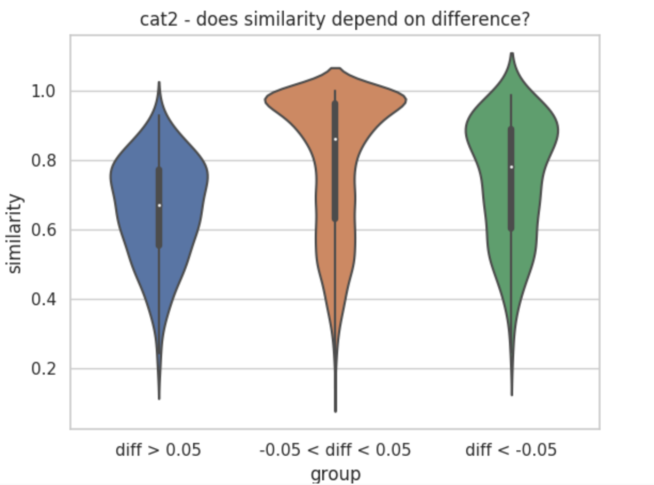

Figure 6: Distribution of average cosine similarities for the three groups of cat_2, showing low, intermediate and high average shifts respectively.

Where to go next

Wrap-up

ANNs approximate the **distribution of constructions** at a quite refined level, even when trained over a bare 3M words from the CHILDES corpus.

We can follow paths of abstraction by putting our **grammar formalism** in a vector space.

- no sharp distinction between **lexicon** and **grammar** $\rightarrow$ different items can therefore be compared, irrespective of their lexical nature
- no assumption about the **stability** of the constructicon $\rightarrow$ what is relevant for productivity at the earliest stages of learning might become superfluous later on
- all items are **form-meaning** pairs $\rightarrow$ i.e., constructions
- **distributional semantics** is used both as a quantitative tool and as a usage-based cognitive hypothesis[5] $\rightarrow$ in line with the view of constructions as "*invitations to form categories*"[6]

---

[5] "Distributional semantics in linguistic and cognitive research" (Lenci 2008)
[6] *Explain me this: Creativity, competition, and the partial productivity of constructions* (Goldberg 2019)

Wrap-up

ANNs approximate the **distribution of constructions** at a quite refined level, even when trained over a bare 3M words from the CHILDES corpus.

We can follow paths of abstraction by putting our **grammar formalism** in a vector space.

- no sharp distinction between **lexicon** and **grammar** $\rightarrow$ different items can therefore be compared, irrespective of their lexical nature
- no assumption about the **stability** of the constructicon $\rightarrow$ what is relevant for productivity at the earliest stages of learning might become superfluous later on
- all items are **form-meaning** pairs $\rightarrow$ i.e., constructions
- **distributional semantics** is used both as a quantitative tool and as a usage-based cognitive hypothesis[5] $\rightarrow$ in line with the view of constructions as "*invitations to form categories*"[6]

[5] "Distributional semantics in linguistic and cognitive research" (Lenci 2008)
[6] *Explain me this: Creativity, competition, and the partial productivity of constructions* (Goldberg 2019)

Wrap-up

ANNs approximate the **distribution of constructions** at a quite refined level, even when trained over a bare 3M words from the CHILDES corpus.

We can follow paths of abstraction by putting our **grammar formalism** in a vector space.

- no sharp distinction between **lexicon** and **grammar** → different items can therefore be compared, irrespective of their lexical nature
- no assumption about the **stability** of the constructicon → what is relevant for productivity at the earliest stages of learning might become superfluous later on
- all items are **form-meaning** pairs → i.e., constructions
- **distributional semantics** is used both as a quantitative tool and as a usage-based cognitive hypothesis[5] → in line with the view of constructions as "*invitations to form categories*"[6]

[5] "Distributional semantics in linguistic and cognitive research" (Lenci 2008)
[6] *Explain me this: Creativity, competition, and the partial productivity of constructions* (Goldberg 2019)

The whole plan

- We have looked at how the network is using language, as a proxy of its grammatical competence
- We want to investigate to what extent is the network reasoning ”constructively” rather than ”generatively” and vice versa (i.e., to test the assumption: ”*is it possible to tell apart the shape of the input from the grammar itself?*”)
    - (a) *The smaller they are, the faster they cook*, (b) *The more you give, the more you get*, (c) *Cookies were smaller this time and faster to cook* - is (a) more similar to (b) than to (c)?
    - (a) *The boy sneezed the foam off the cappuccino*, (b) *The dog barked me out of the room*, (c) *Foam boy the off the cappuccino* - is (a) more similar to (b) than to (c)?
- We plan to manipulate specific features of the input language

The whole plan

- We have looked at how the network is using language, as a proxy of its grammatical competence
- We want to investigate to what extent is the network reasoning ”constructively” rather than ”generatively” and vice versa (i.e., to test the assumption: ”*is it possible to tell apart the shape of the input from the grammar itself?*”)
    - (a) *The smaller they are, the faster they cook*, (b) *The more you give, the more you get*, (c) *Cookies were smaller this time and faster to cook* - **is (a) more similar to (b) than to (c)?**
    - (a) *The boy sneezed the foam off the cappuccino*, (b) *The dog barked me out of the room*, (c) *Foam boy the off the cappuccino* - **is (a) more similar to (b) than to (c)?**
- We plan to manipulate specific features of the input language

The whole plan

- We have looked at how the network is using language, as a proxy of its grammatical competence
- We want to investigate to what extent is the network reasoning ”constructively” rather than ”generatively” and vice versa (i.e., to test the assumption: ”*is it possible to tell apart the shape of the input from the grammar itself?*”)
  - (a) *The smaller they are, the faster they cook*, (b) *The more you give, the more you get*, (c) *Cookies were smaller this time and faster to cook* - **is (a) more similar to (b) than to (c)?**
  - (a) *The boy sneezed the foam off the cappuccino*, (b) *The dog barked me out of the room*, (c) *Foam boy the off the cappuccino* - **is (a) more similar to (b) than to (c)?**
- We plan to manipulate specific features of the input language

Thank you!

References

Baroni, Marco (2020). "Linguistic generalization and compositionality in modern artificial neural networks". In: *Philosophical Transactions of the Royal Society B* 375.1791, p. 20190307.

Goldberg, Adele E (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

— (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Lenci, Alessandro (2008). "Distributional semantics in linguistic and cognitive research". In: *Italian journal of linguistics* 20.1, pp. 1–31.

Osborne, Timothy and Thomas Groß (2012). *Constructions are catenae: Construction grammar meets dependency grammar*.

Osborne, Timothy, Michael Putnam, and Thomas Groß (2012). "Catenae: Introducing a novel unit of syntactic analysis". In: *Syntax* 15.4, pp. 354–396.