IN5550: Neural Methods in Natural Language Processing
Sub-lecture 5.1
Distributional and distributed

Andrey Kutuzov

University of Oslo

21 February 2023

Contents

Bag-of-words problems

Simple Bag-of-Words approaches (without representation learning) do not take into account semantic relationships between linguistic entities.

Bag-of-words problems

Simple Bag-of-Words approaches (without representation learning) do not take into account semantic relationships between linguistic entities.

No way to detect semantic similarity between documents which do not share words:

- *The war was devastating for the region.*
- *This military conflict left the country in ruins.*

Distributional and Distributed

Bag-of-words problems

Simple Bag-of-Words approaches (without representation learning) do not take into account semantic relationships between linguistic entities.

No way to detect semantic similarity between documents which do not share words:

- *The war was devastating for the region.*
- *This military conflict left the country in ruins.*

It means we need more sophisticated semantically-aware methods.

Bag-of-words problems

Simple Bag-of-Words approaches (without representation learning) do not take into account semantic relationships between linguistic entities.

No way to detect semantic similarity between documents which do not share words:

- *The war was devastating for the region.*
- *This military conflict left the country in ruins.*

It means we need more sophisticated semantically-aware methods.

Like distributional word embeddings.

Distributional and Distributed

Distant memory from the last lecture

- '*Generalizations: similar words get similar representations in the embedding layer*'

Distributional and Distributed

Distant memory from the last lecture

- '*Generalizations: similar words get similar representations in the embedding layer*'
- Neural language models learn vector representations for words as a byproduct of their training process.

Distant memory from the last lecture

- '*Generalizations: similar words get similar representations in the embedding layer*'
- Neural language models learn vector representations for words as a byproduct of their training process.
- These representations are similar for semantically similar words.

Distant memory from the last lecture

- ‘*Generalizations: similar words get similar representations in the embedding layer*’
- Neural language models learn vector representations for words as a byproduct of their training process.
- These representations are similar for semantically similar words.
- Good word embeddings from an auxiliary task:
  - Language models (LMs) are trained on raw texts, no manual annotation needed.

Distributional and Distributed

Distant memory from the last lecture

- '*Generalizations: similar words get similar representations in the embedding layer*'
- Neural language models learn vector representations for words as a byproduct of their training process.
- These representations are similar for semantically similar words.
- Good word embeddings from an auxiliary task:
  - Language models (LMs) are trained on raw texts, no manual annotation needed.
  - One can train an LM on the texts collected from the whole Internet.
    - Internet Archive
    - CommonCrawl
    - etc

Distributional and Distributed

Distant memory from the last lecture

- '*Generalizations: similar words get similar representations in the embedding layer*'
- Neural language models learn vector representations for words as a byproduct of their training process.
- These representations are similar for semantically similar words.
- Good word embeddings from an auxiliary task:
  - Language models (LMs) are trained on raw texts, no manual annotation needed.
  - One can train an LM on the texts collected from the whole Internet.
    - Internet Archive
    - CommonCrawl
    - etc

How come that we can get good word embeddings without any manually annotated data?

All of this week' sub-lectures in one slide

- Vector space models of meaning based on distributional information are not something new [Turney et al., 2010].

All of this week' sub-lectures in one slide

- Vector space models of meaning based on distributional information are not something new [Turney et al., 2010].
- But around 2011-2013, such representations trained using machine learning became extremely popular in NLP.

All of this week' sub-lectures in one slide

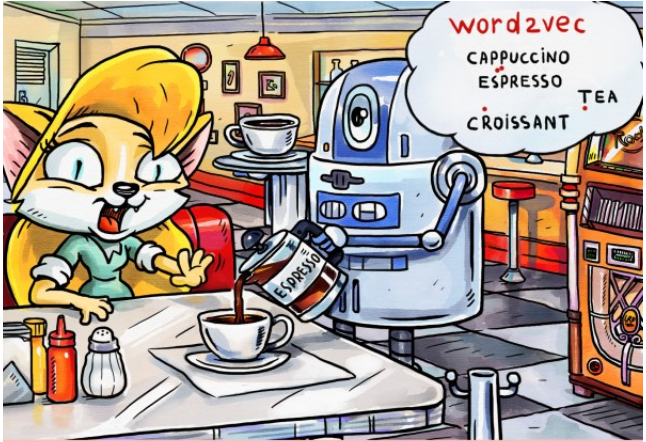
- Vector space models of meaning based on distributional information are not something new [Turney et al., 2010].
- But around 2011-2013, such representations trained using machine learning became extremely popular in NLP.
- Commonly used in research and large-scale industry projects (web search, opinion mining, tracing events, plagiarism detection, document collections management, etc.)

Distributional and Distributed

All of this week' sub-lectures in one slide

- Vector space models of meaning based on distributional information are not something new [Turney et al., 2010].
- But around 2011-2013, such representations trained using machine learning became extremely popular in NLP.
- Commonly used in research and large-scale industry projects (web search, opinion mining, tracing events, plagiarism detection, document collections management, etc.)
- All this is based on their ability to efficiently predict semantic similarity between linguistic entities (in particular, words).
- Semantic information is distributed across word vectors, making them non-interpretable.

Distributional and Distributed



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

(by Dmitry Malkov)

Distributional hypothesis

OK, why does it work at all?

OK, why does it work at all?

Tiers of linguistic analysis

Distributional hypothesis

OK, why does it work at all?

Tiers of linguistic analysis

Computational approaches model various tiers of language:

- graphematics – how words are spelled,

Distributional hypothesis

OK, why does it work at all?

Tiers of linguistic analysis

Computational approaches model various tiers of language:

- graphematics – how words are spelled,
- phonetics – how words are pronounced,

Distributional hypothesis

OK, why does it work at all?

Tiers of linguistic analysis

Computational approaches model various tiers of language:

- graphematics – how words are spelled,
- phonetics – how words are pronounced,
- morphology – how words inflect,

Distributional hypothesis

OK, why does it work at all?

Tiers of linguistic analysis

Computational approaches model various tiers of language:

- graphematics – how words are spelled,
- phonetics – how words are pronounced,
- morphology – how words inflect,
- syntax – how words interact in sentences,

OK, why does it work at all?

Tiers of linguistic analysis

Computational approaches model various tiers of language:

- graphematics – how words are spelled,
- phonetics – how words are pronounced,
- morphology – how words inflect,
- syntax – how words interact in sentences,
- pragmatics – how sentences serve communicative purposes of human beings.

Distributional hypothesis

To model means to capture important features of some phenomenon. For example, in a phrase *'The judge sits in the court'*, the word *'judge'*:

Distributional hypothesis

To model means to capture important features of some phenomenon. For example, in a phrase *'The judge sits in the court'*, the word *'judge'*:

1. consists of 3 phonemes [j e j];

Distributional hypothesis

To model means to capture important features of some phenomenon. For example, in a phrase *'The judge sits in the court'*, the word *'judge'*:

1. consists of 3 phonemes [j e j];
2. is a singular noun in the nominative case;

Distributional hypothesis

To model means to capture important features of some phenomenon. For example, in a phrase *'The judge sits in the court'*, the word *'judge'*:

1. consists of 3 phonemes [j e j];
2. is a singular noun in the nominative case;
3. is a nominal subject dependent of the word 'sits' in the syntactic tree of our sentence.

Distributional hypothesis

To model means to capture important features of some phenomenon. For example, in a phrase *'The judge sits in the court'*, the word *'judge'*:

1. consists of 3 phonemes [j e j];
2. is a singular noun in the nominative case;
3. is a nominal subject dependent of the word 'sits' in the syntactic tree of our sentence.

Such representations describe many important features of the word *'judge'*.
But not meaning (semantics).

Distributional hypothesis

To model means to capture important features of some phenomenon. For example, in a phrase *'The judge sits in the court'*, the word *'judge'*:

1. consists of 3 phonemes [j e j];
2. is a singular noun in the nominative case;
3. is a nominal subject dependent of the word 'sits' in the syntactic tree of our sentence.

Such representations describe many important features of the word *'judge'*.
But not meaning (semantics).

Question

Are these representations discrete or continuous?

Distributional hypothesis

How to represent meaning?

Distributional hypothesis

How to represent meaning?

- Semantics is difficult to represent formally.

Distributional hypothesis

How to represent meaning?

- Semantics is difficult to represent formally.
- Words which are similar in their meaning should possess mathematically similar representations independent of their surface forms.

Distributional hypothesis

How to represent meaning?

- Semantics is difficult to represent formally.
- Words which are similar in their meaning should possess mathematically similar representations independent of their surface forms.
- *'Judge'* must be similar to *'court'* but not to *'kludge'*...
- ...even though their surface form suggests the opposite.

Distributional hypothesis

How to represent meaning?

- Semantics is difficult to represent formally.
- Words which are similar in their meaning should possess mathematically similar representations independent of their surface forms.
- *'Judge'* must be similar to *'court'* but not to *'kludge'*...
- ...even though their surface form suggests the opposite.
- Why so?

Distributional hypothesis

Arbitrariness of a linguistic sign

Distributional hypothesis

Arbitrariness of a linguistic sign

Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]

Distributional hypothesis

Arbitrariness of a linguistic sign

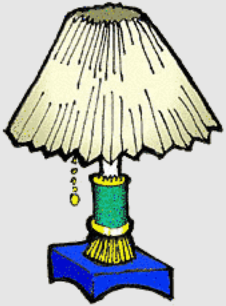
Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]
The concept of *'Lantern'* can be expressed by any sequence of letters or sounds in different languages:

Distributional hypothesis

Arbitrariness of a linguistic sign

Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]

The concept of *'Lantern'* can be expressed by any sequence of letters or sounds in different languages:



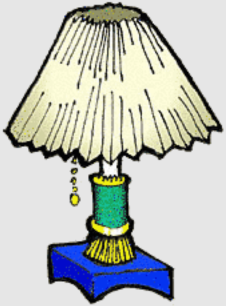
- lantern

Distributional hypothesis

Arbitrariness of a linguistic sign

Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]

The concept of *'Lantern'* can be expressed by any sequence of letters or sounds in different languages:

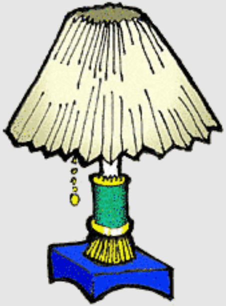


- lantern
- lykt

Arbitrariness of a linguistic sign

Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]

The concept of *'Lantern'* can be expressed by any sequence of letters or sounds in different languages:



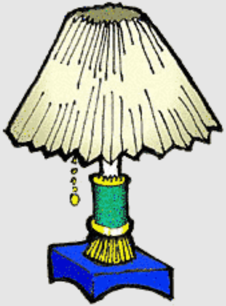
- lantern
- lykt
- лампа

Distributional hypothesis

Arbitrariness of a linguistic sign

Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]

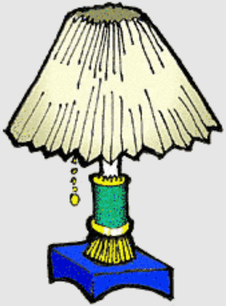
The concept of *'Lantern'* can be expressed by any sequence of letters or sounds in different languages:



- lantern
- lykt
- лампа
- lucerna

Arbitrariness of a linguistic sign

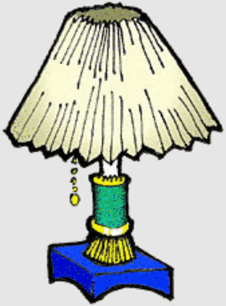
Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]
The concept of *'Lantern'* can be expressed by any sequence of letters or sounds in different languages:



- lantern
- lykt
- лампа
- lucerna
- гэрэл

Arbitrariness of a linguistic sign

Unlike in road signs, there's no direct link between form and meaning in words [Saussure, 1916]
The concept of *'Lantern'* can be expressed by any sequence of letters or sounds in different languages:



- lantern
- lykt
- лампа
- lucerna
- гэрэл
- ...

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?

And how can our ML models get this information?

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?

And how can our ML models get this information?

Possible data sources

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?

And how can our ML models get this information?

Possible data sources

Methods of computationally representing semantics in natural languages fall into 2 large groups:

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?
And how can our ML models get this information?

Possible data sources

Methods of computationally representing semantics in natural languages fall into 2 large groups:

1. Manually building semantic networks or ontologies (knowledge-based approach). Works top-down: from abstractions to real texts.

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?
And how can our ML models get this information?

Possible data sources

Methods of computationally representing semantics in natural languages fall into 2 large groups:

1. Manually building semantic networks or ontologies (knowledge-based approach). Works top-down: from abstractions to real texts. For example, WordNet [Miller, 1995].

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?
And how can our ML models get this information?

Possible data sources

Methods of computationally representing semantics in natural languages fall into 2 large groups:

1. Manually building semantic networks or ontologies (knowledge-based approach). Works top-down: from abstractions to real texts. For example, WordNet [Miller, 1995].
2. Extracting semantics from usage patterns in text corpora (distributional approach). Works bottom-up: from real texts to abstractions.

Distributional hypothesis

How do we humans know that '*lantern*' and '*lamp*' have similar meaning? What is meaning, after all?

And how can our ML models get this information?

Possible data sources

Methods of computationally representing semantics in natural languages fall into 2 large groups:

1. Manually building semantic networks or ontologies (knowledge-based approach). Works top-down: from abstractions to real texts. For example, WordNet [Miller, 1995].
2. Extracting semantics from usage patterns in text corpora (distributional approach). Works bottom-up: from real texts to abstractions.

The second approach is behind 'word embeddings' (and most modern NLP).

Distributional hypothesis

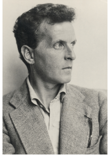
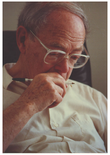
Hypothesis: meaning is actually a sum of contexts.

Distributional differences will always be enough to explain semantic differences:

Distributional hypothesis

Hypothesis: meaning is actually a sum of contexts.

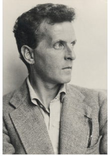
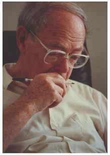
Distributional differences will always be enough to explain semantic differences:


- Words with similar typical contexts have similar meaning.

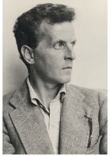
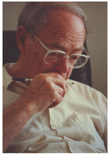
Distributional hypothesis

Hypothesis: meaning is actually a sum of contexts.
Distributional differences will always be enough to explain semantic differences:


- Words with similar typical contexts have similar meaning.
- First formulated by:
  - philosopher Ludwig Wittgenstein (1930s);
  - linguists Zelig Harris [Harris, 1954] and John Firth.

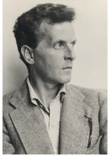
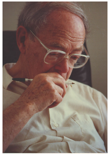
Hypothesis: meaning is actually a sum of contexts.
Distributional differences will always be enough to explain semantic differences:


- Words with similar typical contexts have similar meaning.
- First formulated by:
  - philosopher Ludwig Wittgenstein (1930s);
  - linguists Zelig Harris [Harris, 1954] and John Firth.
- *'You shall know a word by the company it keeps'* [Firth, 1957]

Distributional hypothesis

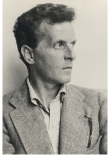
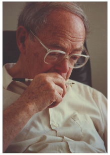
Hypothesis: meaning is actually a sum of contexts.
Distributional differences will always be enough to explain semantic differences:


- Words with similar typical contexts have similar meaning.
- First formulated by:
  - philosopher Ludwig Wittgenstein (1930s);
  - linguists Zelig Harris [Harris, 1954] and John Firth.
- *'You shall know a word by the company it keeps'* [Firth, 1957]
- More details in [Brunila and LaViolette, 2022].

Distributional hypothesis

Hypothesis: meaning is actually a sum of contexts.
Distributional differences will always be enough to explain semantic differences:


- Words with similar typical contexts have similar meaning.
- First formulated by:
  - philosopher Ludwig Wittgenstein (1930s);
  - linguists Zelig Harris [Harris, 1954] and John Firth.
- *'You shall know a word by the company it keeps'* [Firth, 1957]
- More details in [Brunila and LaViolette, 2022].
- Distributional semantics models (DSMs) get information from lexical co-occurrences in large natural corpora.

Distributional hypothesis

Contexts for '*tea*':

establishments, besides two livery stables, a	tea They never boasted of Robert Acton, nor indulg
en things, because their methods of family to	tea at once. pose, which you carry so well
, as, indeed, A waiter comes in with the	tea. He places the tray on the table. Jasper
let me always remain here.' “I prefer weak	tea!” cried Daisy, and she went off with the
hell. I should think you had drunk enough	tea in Chin a. life; it is a failure,
. Not a bit. Come in and have some	tea. Stay to dinner. every year. Don't persist
responsibility. And greatly as we enjoyed our	tea Crusoe island. Then there's the religious diff
your naturally liking me. (She is and had	tea in the evening. Afraid though as he was
] Tell them I shan't be home to	tea, will you, LADY BRITOMART. I must get the
, that was Mr. McComas will not come to	tea, ma'am: he has gone to call upon
, or asked you to have a cup of	tea. It's not human. ugly woman must have
woman can hardly know one places Gilbey's	tea on the table before him]. The lady that
of my – my hopes.' BROADBENT. He'll want	tea. Let us have some. BURGE-LUBIN [_resolutely ge
: THE MANAGER. Can I take any order? Some	tea? would THE SHE-ANCIENT. Speak, Arjillax: you w
to Tramp.} Will you drink a sup of	tea with myself and the the happiest person in
are trying to sleep." the evening after your	tea. “Better still–then there you are!” And Streth
GUINNESS. I'll go get you some fresh	tea, ducky. [She takes up the its burden, is
sional men, artists, and even with laborers	tea services out and made the people who had
the Lutches and Mrs. Rance the attendance at	tea just in the right place on the west
she came over to the great house to	tea. She had let the proposal that she should
. The Baroness found it amusing to go to	tea; she dressed as if for dinner. The tea-
tea; she dressed as if for dinner. The	tea-table offered an anomalous and picturesque rep
would be dead in two years, as the	tea-table. Be serious, Felix. You forget that I

Distributional hypothesis

Contexts for ‘*tea*’:

establishments, besides two livery stables, a	tea They never boasted of Robert Acton, nor indulg
en things, because their methods of family to	tea at once. pose, which you carry so well
, as, indeed, A waiter comes in with the	tea. He places the tray on the table. Jasper
let me always remain here.' “I prefer weak	tea!” cried Daisy, and she went off with the
hell. I should think you had drunk enough	tea in Chin a. life; it is a failure,
. Not a bit. Come in and have some	tea. Stay to dinner. every year. Don't persist
responsibility. And greatly as we enjoyed our	tea Crusoe island. Then there's the religious diff
your naturally liking me. (She is and had	tea in the evening. Afraid though as he was
] Tell them I shan't be home to	tea, will you, LADY BRITOMART. I must get the
, that was Mr. McComas will not come to	tea, ma'am: he has gone to call upon
, or asked you to have a cup of	tea. It's not human. ugly woman must have
woman can hardly know one places Gilbey's	tea on the table before him]. The lady that
of my – my hopes.' BROADBENT. He'll want	tea. Let us have some. BURGE-LUBIN [_resolutely ge
: THE MANAGER. Can I take any order? Some	tea? would THE SHE-ANCIENT. Speak, Arjillax: you w
to Tramp.} Will you drink a sup of	tea with myself and the the happiest person in
are trying to sleep." the evening after your	tea. “Better still–then there you are!” And Streth
GUINNESS. I'll go get you some fresh	tea, ducky. [She takes up the its burden, is
sional men, artists, and even with laborers	tea services out and made the people who had
the Lutches and Mrs. Rance the attendance at	tea just in the right place on the west
she came over to the great house to	tea. She had let the proposal that she should
. The Baroness found it amusing to go to	tea; she dressed as if for dinner. The tea-
tea; she dressed as if for dinner. The	tea-table offered an anomalous and picturesque rep
would be dead in two years, as the	tea-table. Be serious, Felix. You forget that I

Contexts for ‘*coffee*’:

prompt his an incident in my life as	coffee for breakfast. Of course, hes too _Two fig
ek her out all courteously, PETKOFF (over his	coffee and cigaret). I don't believe in going
could be done, too,' he remarked, sipping his	coffee. 'Bury him in some sort,' I explained. 'One
, of us. I should like a cup of	coffee. MICHAEL. If you'd come in better hours,
UKA (innocently). Perhaps you would like some	coffee, sir? DISCOVERY ANTICIPATED BY DIVINATION s
her in public because he has fallen head	coffee-colored heathens and pestilential white agi
manners for he was novels, broken backed,	coffee stained, torn and '''This was the last time
stretches her hand across the table for the	coffee pot.) welcome, an expression which drops in
ittle sitting-room, and cigarettes, after the	coffee, had been permitted by the ladies, and in
had just given me a pannikin of hot	coffee....Slapped it down there, on my chest–bange
a heavy roll coming; tried to save my	coffee, burnt my fingers....and fell out of my
I'll have a claret cup instead of	coffee. Put some first night that we've come
t of trouble travelling. And then, with fresh	coffee, a clean cup, and a brandy bottle on
. Your word had such weight with me!” fresh	coffee? He gave his friend a glance as to
wont press you. `Try a weed with your	coffee. Local tobacco. The black coffee you get at

Representing words with vectors

- Your neural classifiers in Obligatory 1 implicitly learned vector representations for words (embeddings).
- In practice, representing word meaning with vectors was first popularized in psychology by [Osgood et al., 1964]...
- ...then developed by many others.

Representing words with vectors

- Your neural classifiers in Obligatory 1 implicitly learned vector representations for words (embeddings).
- In practice, representing word meaning with vectors was first popularized in psychology by [Osgood et al., 1964]...
- ...then developed by many others.

- Word vectors can be created manually...
- ...but in most cases, corpus-driven distributional methods are much more efficient.

Representing words with vectors

Componential analysis: manual creation of word vectors

Kinship terms	[MALE]	[ASCEND]	[DESCEND]	[LINEAL]
Father	+	+	-	+
Mother	-	+	-	+
Uncle	+	+	-	-
Aunt	-	+	-	-
Brother	+	-	-	+
Sister	-	-	-	+
Son	+	-	+	+
Daughter	-	-	+	+
Nephew	+	-	+	-
Niece	-	-	+	-
Causin	+/-	-	-	-

[Widyastuti, 2010]

We will not do this. We will use distributional vector models *(next sub-lecture 5.2)*.

References I

Brunila, M. and LaViolette, J. (2022).
What company do words keep? revisiting the distributional semantics of J.R. Firth & Zellig Harris.
In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4403–4417, Seattle, United States. Association for Computational Linguistics.

Firth, J. (1957).
A synopsis of linguistic theory, 1930-1955.
Blackwell.

Harris, Z. S. (1954).
Distributional structure.
Word, 10(2-3):146–162.

References II

Miller, G. (1995).
Wordnet: a lexical database for English.
Communications of the ACM, 38(11):39–41.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1964).
The measurement of meaning.
University of Illinois Press.

Saussure, F. d. (1916).
Course in general linguistics.
Duckworth.

Turney, P., Pantel, P., et al. (2010).
From frequency to meaning: Vector space models of semantics.
Journal of artificial intelligence research, 37(1):141–188.

References III

Widyastuti, S. (2010).
Componential analysis of meaning: Theory and applications.
Journal of English and Education, 4(1):116–128.