Computational methods in gender linguistic research: Distributional semantics and discriminative learning

Dr Dominic Schmitz
English Language and Linguistics
Heinrich Heine University Düsseldorf

dominic.schmitz@hhu.de
dmncschmtz.bsky.social

Complementary perspectives

Computational methods
- Model structure in high-dimensional linguistic data
- Derive semantic representations
- Simulate lexical processing
- Generate testable, quantitative predictions

Corpus methods
- Reveal large-scale usage patterns
- Identify frequencies, collocations, genre effects
- Track diachronic change in gendered expressions
- Make biases visible in naturally occurring data

Experimental methods
- Probe real-time comprehension and production
- Measure reaction times, choices, eye movements
- Separate grammatical cues from social cues
- Provide causal evidence through controlled manipulation

→ Corpus approaches show *what* patterns exist

→ Experimental approaches show *how* speakers process them

→ Computational approaches help explain *why* they arise and *what* they predict

Distributional semantics

- Meaning is reflected in patterns of use
- Harris (1954)

  "Words that occur in similar contexts tend to have similar meanings"

- Firth (1957)

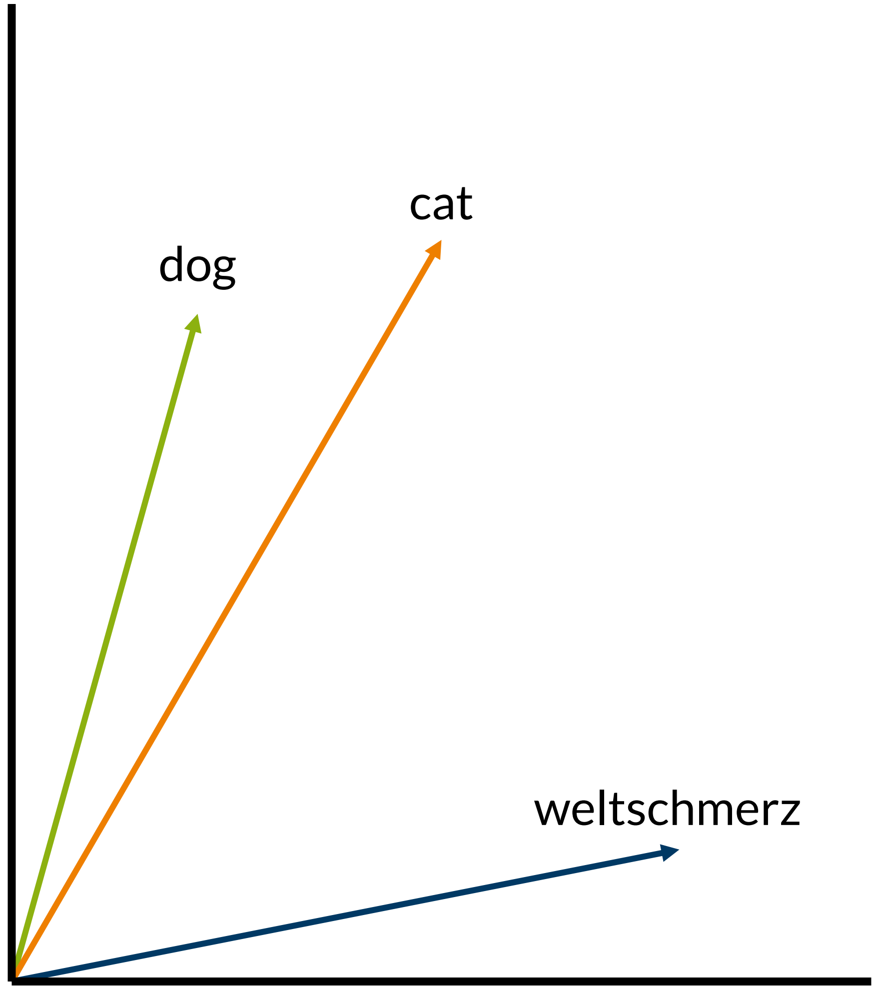
  "You shall know a word by the company it keeps"

- Distributional models reverse-engineer meaning by tracking contextual regularities in large corpora
- Output: high-dimensional semantic vectors representing patterns of linguistic behaviour

Boleda (2020)

Vector space models of meaning

- Each word becomes a point in a multidimensional semantic space
- Dimensions capture statistical properties of contexts (often not interpretable individually)
- Semantic similarity corresponds to geometric closeness (e.g. cosine similarity)
- Distances encode graded, continuous semantic relatedness



Boleda (2020)

What counts as "context"?

- Context can mean many things
  - words in a window (±*n* words)
  - sentence, paragraph, document
  - syntactic relations
  - multimodal contexts (e.g. visual information)
- Different models operationalise context differently, leading to different semantic spaces
- Key idea: semantically similar words have overlapping context distributions

Boleda (2020)

Defining context

```
The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]
```

Thater (2011)

Defining context

The **apple** is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]

Thater (2011)

Defining context

The apple is the pomaceous fruit of the **apple** tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]

Thater (2011)

Defining context

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of **apples**, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated, and its fruits have been improved through selection over thousands of years. […]

Thater (2011)

Defining context

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics. […] The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown. […] The apple tree was perhaps the earliest tree to be cultivated,

selection o

	tree	fruit	forms	perhaps	apple	…
…						
apple	3	2	1	1	0	…
tree	0	1	1	1	3	…
…						

Thater (2011)

Defining context

The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family (Rosaceae). […] There are more than 7,500 known cultivars of apples, resulting in a range of […] forms a tree that is […] etres (9.8 to 39 ft) […] y crown. […] The apple tre[…] perhaps the earliest tree to be cultivated, […] selection o[…]

rows
the vocabulary we are interested in

columns
"features" of the vocabulary, i.e. semantic dimensions

	tree	**fruit**	**forms**	**perhaps**	**apple**	**...**
...						
apple	3	2	1	1	0	...
tree	0	1	1	1	3	...
...						

Thater (2011)

From co-occurrence to vectors

- Start with a large vocabulary and context inventory
- Build a matrix: rows = target words, columns = contextual features
- Fill the matrix with counts or derived statistics
- Each row → a word vector, e.g. $v_{apple} = \langle 3, 2, 1, 1, 0, \dots \rangle$

	tree	**fruit**	**forms**	**perhaps**	**apple**	**...**
...						
apple	3	2	1	1	0	...
tree	0	1	1	1	3	...
...						

Thater (2011)

Design choices in distributional models 1

Pre-processing

- Do we use word-forms (*teachers*, *teacher*, *Teacher*) or simplified forms (teacher)?
- Do we keep function words (*the*, *of*, *to*) or remove them?
- How do we treat punctuation, compounds, or multi-word expressions?

Context

- A fixed window (e.g. the five words around the target)?
- Whole sentences or paragraphs?
- Grammatically defined relations (e.g. subject–verb–object links)?

Design choices in distributional models 2

Associative strength

- Some methods simply count co-occurrences
- Others give more weight to informative contexts (rare but meaningful associations)
- Some methods learn these weights automatically

Model type

- **Count-based models**
  Build a large co-occurrence table and transform it
- **Predictive models**
  Learn vectors by predicting missing words from context
- **Subword models**
  Include information from letter sequences to handle morphological richness

Predictive models: *word2vec*

- Instead of counting co-occurrences, predictive models learn meaning by **guessing** words
- The model sees sentences and repeatedly asks itself questions like "Given this context, which word is likely to appear here?"
- To guess well, the model must learn which words occur in similar situations
- Words with similar contexts end up having similar vectors

The ______ grows on trees.

Mikolov et al. (2013)

Subword models: *fastText*

- Many languages have rich morphology: *Student*, *Studentin*, *Studierende*, …
- Traditional models treat each word as unrelated, even if they clearly share meaning
- *fastText* improves this by breaking words into small letter chunks
- These 'subword' pieces also get vectors, and a word's meaning is built from these pieces, e.g. $v_{student\_n3} = \langle stu, tud, ude, den, ent \rangle$
- Result
  - better handling of rare (and even novel) forms
  - better handling of inflected words
  - more realistic similarity between related forms

Bojanowski et al. (2016)

Subword models: *fastText*

- Example: *Student*
  - $v_{student\_n2} = \langle st, tu, ud, de, en, nt \rangle$
  - $v_{student\_n3} = \langle stu, tud, ude, den, ent \rangle$
  - $v_{student\_n4} = \langle stud, tude, uden, dent \rangle$
  - $v_{student\_n5} = \langle stude, tuden, udent \rangle$
  - $v_{student\_n6} = \langle studen, tudent \rangle$
  - $v_{student\_full} = \langle student \rangle$
  - $v_{student} = \left\langle \begin{matrix} v_{student\_n2} + v_{student\_n3} + v_{student\_n4} + v_{student\_n5} + v_{student\_n6} + \\ v_{student\_full} \end{matrix} \right\rangle$

Bojanowski et al. (2016)

A problem: one form, two meanings

- German gives us a methodological gift and a headache at the same time
  - Gift: masculine role nouns are beautifully regular
  - Headache: that regularity hides two different meanings

    Arbeiter = ‘male worker’ (specific masculine)

    Arbeiter = ‘worker of any/unknown gender’ (generic masculine)

- *fastText* doesn’t know this, it only sees the spelling *Arbeiter*
- So, both meanings collapse into one vector $v_{Arbeiter}$

Schmitz (2024)

A solution: instance vectors

- Following Lapesa et al. (2018) we can compute instance vectors
  - Instead of one vector per word type, compute one vector per token based on the actual words around it
- That is,
  - Take a target token (e.g. *Arbeiter*)
  - Take the *n* content words before and after it
  - Average their *fastText* vectors
  - The result = an instance vector capturing the meaning in this sentence
- Instance vectors are thus a method of contextual semantic disambiguation that remains purely distributional, without resorting to grammars or lexicons

Schmitz (2024)

A study: male bias of generic masculines

1. **Corpus**

   30,000 manually annotated attestations of role nouns (generic vs specific use)

2. **Target paradigms**

   76 role nouns from Gabriel et al. (2008)

3. **Context vectors**

   Pre-trained German *fastText* vectors (subword-based)

4. **Instance vector computation**

   For each target token, compute one instance vector at window sizes $n = 2, 5, 8$

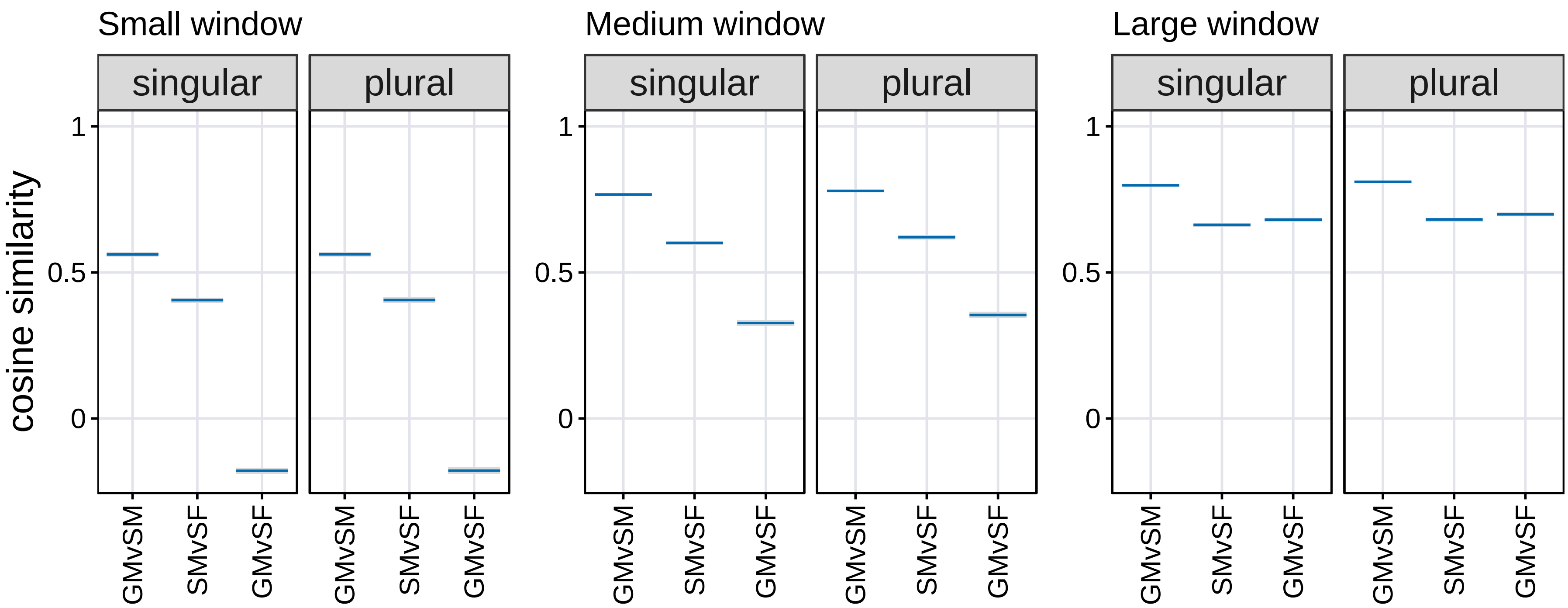
5. **Analysis**

   Compute cosine similarities within paradigms (generic $\leftrightarrow$ specific masc; generic $\leftrightarrow$ specific fem; specific masc $\leftrightarrow$ specific fem)

   Model using beta-regression

Schmitz (2024)

A study: male bias of generic masculines



- Across number and window sizes, generic masculines are always more similar to specific masculines than to specific feminines

Schmitz (2024)

A study: male bias of generic masculines

- A single *fastText* vector cannot solve the generic/specific ambiguity
- But instance vectors can
- Instance vectors allow us to
  - treat each token as its own semantic event,
  - recover the contextual meaning of *Arbeiter* as used in that sentence,
  - and measure similarity patterns that reveal systematic male bias in generic uses

Schmitz (2024)

From distributional semantics to discriminative learning

- Distributional models (like *fastText*) capture semantic similarity through patterns of co-occurrence
- They give us representations, but not learning
  - How do speakers actually acquire these mappings?
  - How do form cues lead to meaning in comprehension?
  - How do meanings activate forms in production?
- Discriminative learning models address exactly this
- They implement error-driven learning, cue competition, and discriminability, providing a cognitive model of how lexical knowledge emerges from usage patterns

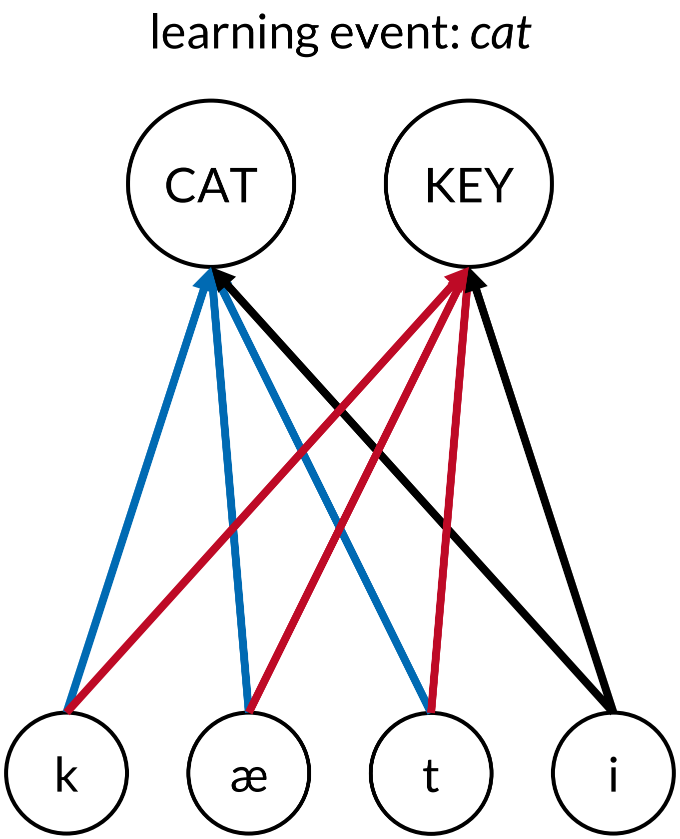
Chuang & Baayen (2021)

Discriminative learning: the basic idea

- Learning is **error-driven**: what is learned is the difference between predicted and observed outcomes
- **Cues** compete to predict **outcomes**
- Correct predictions strengthen cue → outcome links; incorrect predictions weaken cue → outcome links
- Learning is incremental, usage-based, and continuous across the lifespan
- Lexical knowledge emerges from these learned mappings

Baayen et al. (2011); Chuang & Baayen (2021)

Cue competition in lexical learning



Chuang & Baayen (2021)

The learning mechanism: error-driven updating

- Discriminative learning follows the Rescorla–Wagner model of associative learning (Rescorla & Wagner, 1972)
- Learning is error-driven: learning happens when reality does not match the model's prediction

Baayen et al. (2011); Chuang & Baayen (2021)

The learning mechanism: error-driven updating

- Imagine the model is trying to learn that the sound /kæt/ means CAT, not KEY or CUP

Step 1: Look at the cues in the input

- When the model hears /kæt/, it identifies the cues: /k/, /æ/, /t/
- These cues each have weights pointing to many possible meanings (cat, key, cap, …)

Chuang & Baayen (2021)

The learning mechanism: error-driven updating

- Imagine the model is trying to learn that the sound /kæt/ means CAT, not KEY or CUP

Step 2: Add up the current evidence for each meaning

- The model adds the cue → meaning weights:
  - maybe /k/ gives weak support for CAT and KEY,
  - /æ/ gives strong support for CAT,
  - /t/ gives modest support for CAT
- This creates the model's **prediction**:

  CAT might get a medium score, KEY a low score, etc.

Chuang & Baayen (2021)

The learning mechanism: error-driven updating

- Imagine the model is trying to learn that the sound /kæt/ means CAT, not KEY or CUP

Step 3: Compare prediction with the actual outcome

- Reality is: **CAT = correct**, all others = incorrect
- If the model did **not predict CAT strongly enough**, this is an *error*
- If the model gave **too much support to KEY**, that is also an *error*

Chuang & Baayen (2021)

The learning mechanism: error-driven updating

- Imagine the model is trying to learn that the sound /kæt/ means CAT, not KEY or CUP

Step 4: Adjust the weights to reduce this error next time

- Strengthen the weights from the present cues (/k/, /æ/, /t/) **towards CAT**, because CAT should have been predicted more strongly
- Weaken the weights from these same cues **towards KEY, CAP, CUP**, because the model wrongly predicted them
- After many such steps, the cue → meaning weights come to reflect the statistical structure of the input

Chuang & Baayen (2021)

NDL as a model of semantic structure

- Each outcome has a vector of incoming cue weights
- If we specify outcomes and cues to be words, we compute word embeddings
- Each embedding reflects the cues that reliably predict the meaning of its outcome
- Outcomes with similar patterns of predictive cues are semantically similar
- NDL thus provides a purely usage-based semantic space aligned with psychological learning principles

Baayen et al. (2011); Chuang & Baayen (2021)

Why NDL is useful beyond theory

- **Transparent**: mathematically equivalent to linear regression
- **Flexible** cue choices: orthography, phonology, morphology, syntax, prosody, social meaning signals
- **No morpheme segmentation or rule system required** — structure emerges from data
- **Connects semantics and processing**
    - predicts activation, competition, confusability, semantic neighbourhood effects
    - can feed into behavioural models (RTs, choices, acoustic durations)

Baayen et al. (2011); Chuang & Baayen (2021)

Applying NDL: modelling generic and specific meanings

- German gives us a methodological gift and a headache at the same time
  - Gift: masculine role nouns are beautifully regular
  - Headache: that regularity hides two different meanings

    Arbeiter = 'male worker' (specific masculine)

    Arbeiter = 'worker of any/unknown gender' (generic masculine)
- NDL lets us treat the features of these forms as distinct cues/outcomes with separate semantic representations
- By learning from real corpus data, the model reconstructs how the two meanings differ in their semantic neighbourhoods

Schmitz et al. (2023)

Applying NDL: modelling generic and specific meanings

1. **Corpus**

   30,000 manually annotated attestations of role nouns (generic vs specific use)

2. **Target paradigms**

   76 role nouns from Gabriel et al. (2008)

3. **NDL setup**

   - Cues and outcomes: all content words in a sentence, reduced to base form, and grammatical gender, number, generic/specific
   - Each sentence is a learning event; all cues → each outcome present in that sentence

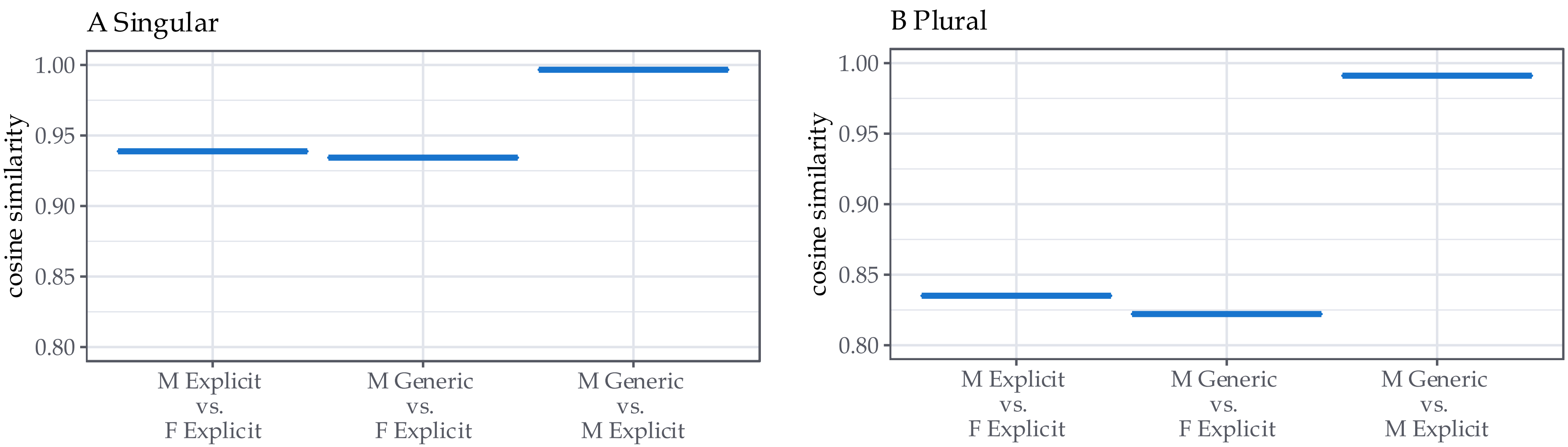
4. **Vectors**

   - Sum of parts vectors for target words, e.g.

$$v_{Arbeiter\_masc\_sg\_g} = v_{Arbeiter} + v_{masculine} + v_{singular} + v_{generic}$$

Schmitz et al. (2023)

Applying NDL: modelling generic and specific meanings



- Across number, generic masculines are more similar to specific masculines than to specific feminines

Schmitz et al. (2023)

From learned semantics to lexical processing

- NDL and other algorithms of distributional semantics give us semantic embeddings: vectors describing how cues relate to outcomes
- But language processing involves (at least) two mappings
  - Comprehension: form → meaning
  - Production: meaning → form
- Linear discriminative learning (LDL) models both mappings directly
- LDL uses the same learning principles as NDL but extends them to vector semantics and continuous mappings

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

Baayen et al. (2019); Chuang & Baayen (2021)

Form vectors

- LDL represents a word's form as a binary vector encoding which sublexical cues it contains
- Standard cues are *n*-grams
- Each unique *n*-gram across the lexicon becomes a column in the form matrix C; each word is a row, marked with 1s where its *n*-grams occur
- This avoids assuming phonemes: speech is contextual, gradients matter, and *n*-grams capture this better than discrete phonemes
- C can be built from orthography, phonology, syllables, or even acoustic vectors
- Because only a few cues are present per word, C is a sparse matrix, optimised for efficient computation

Baayen et al. (2019); Chuang & Baayen (2021)

Form vectors

- Example: *Student*, *Studentin*, *Ärztin*, *resistent*

	#st	stu	tud	ude	den	ent	nt#	nti	tin	in#
student	1	1	1	1	1	1	1	0	0	0
studentin	1	1	1	1	1	1	0	1	1	1
ärztin	0	0	0	0	0	0	0	0	1	1
resistent	0	0	0	0	0	1	1	0	0	0

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

Baayen et al. (2019); Chuang & Baayen (2021)

Meaning vectors

- Meanings are represented by semantic vectors
- Vectors come from NDL or other distributional methods

	D1	D2	D3	D4	D5	...
student	0.2	0.4	0.3	0.9	0.8	...
studentin	0.1	0.5	0.2	0.8	0.7	...
ärztin	0.9	0.1	0.1	0.3	0.2	...
resistent	0.5	0.9	0.8	0.4	0.4	...

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

Baayen et al. (2019); Chuang & Baayen (2021)

Linear mappings

- Linear mappings allow transparent, interpretable learning
- They implement discriminative learning in vector spaces
- Efficient enough to scale to full lexicons (tens of thousands of words)
- Empirically successful in modelling
  - lexical decision (Chuang et al. 2020),
  - auditory recognition (Arnold et al. 2017),
  - morphological processing (Baayen & Smolka 2020),
  - semantic priming (Baayen & Smolka 2020),
  - subphonemic durational differences (Schmitz et al. 2021),
  - and more

Baayen et al. (2019); Chuang & Baayen (2021)

Linear mappings: comprehension

- The model learns which form features reliably point to which areas of meaning space
- LDL learns a transformation matrix F so that $S = C \cdot F$
- Because S and C are high-dimensional, $C \cdot F$ never results in S, but in $\hat{S}$
- $\hat{S}$ is the best approximation to S possible
- $\hat{S}$ reflects the outcome of the comprehension process, i.e. differences between S and $\hat{S}$ represent the doings of the simulated mental lexicon
- Based on $\hat{S}$, meaningful measures based on semantics in the mental lexicon can be derived

Baayen et al. (2019); Chuang & Baayen (2021)

Linear mappings: production

- The model learns which pieces of form are most likely for what it wants to say
- LDL learns a transformation matrix G so that $C = S \cdot G$
- Because S and C are high-dimensional, $S \cdot G$ never results in C, but in $\hat{C}$
- $\hat{C}$ is the best approximation to C possible
- $\hat{C}$ reflects the outcome of the comprehension process, i.e. differences between C and $\hat{C}$ represent the doings of the simulated mental lexicon
- Based on $\hat{C}$, meaningful measures based on forms in the mental lexicon can be derived

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

Baayen et al. (2019); Chuang & Baayen (2021)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience

- Form vectors in C, meaning vectors in S
- Comprehension via F for $\hat{S}$, production via G for $\hat{C}$
- Once trained, LDL can
  - map form $\rightarrow$ meaning
  - map meaning $\rightarrow$ form
  - compute different measures based on comprehended semantics and produced form

Baayen et al. (2019); Chuang & Baayen (2021)

Applying LDL: generic masculines in the mental lexicon

C matrix

- Orthographic trigrams used as form cues
- Each word type corresponds to a row; each trigram cue is a column
- Sparse, binary coding: cue present = 1, absent = 0

S matrix

- The NDL semantic embeddings for each word serve as meaning vectors
- These embeddings capture the learned contextual semantics

Mappings

- Comprehension: form → meaning
- Production: meaning → form

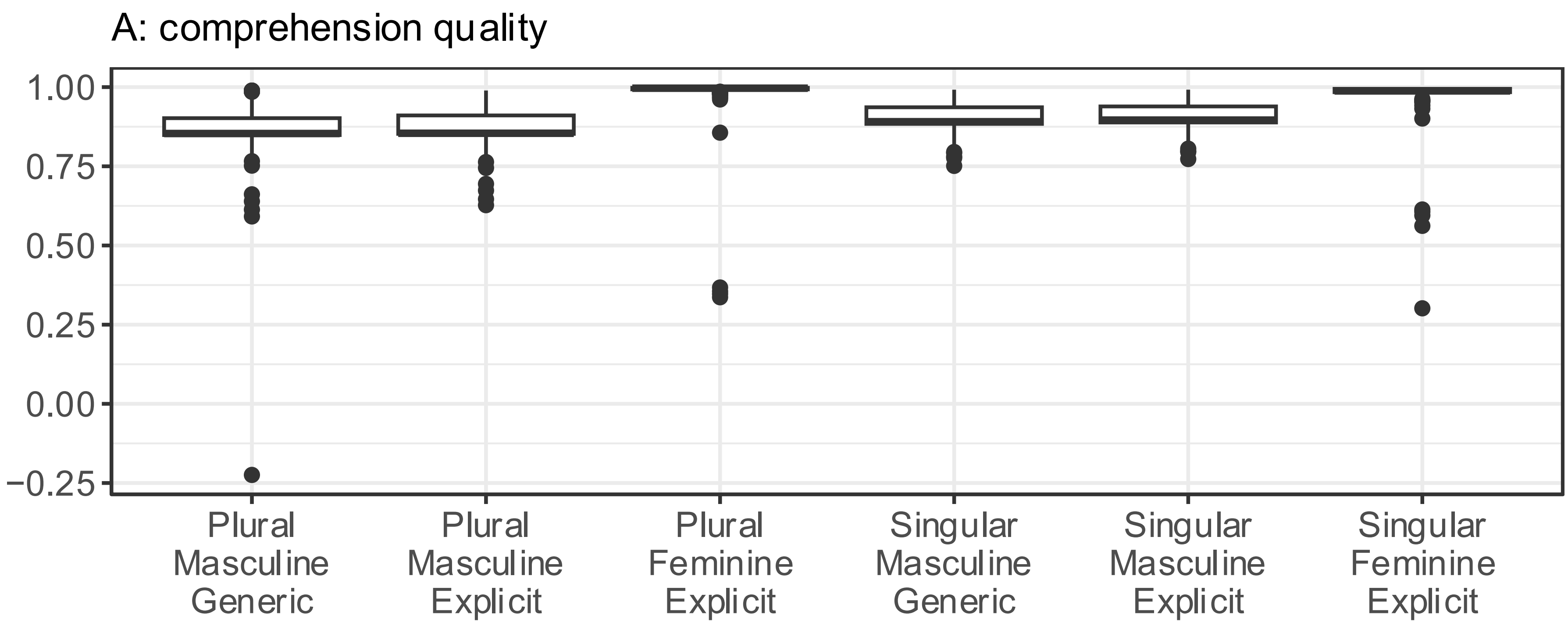
Schmitz et al. (2023)

Applying LDL: generic masculines in the mental lexicon

Measure 1: comprehension quality

- How well is the input semantic vector comprehended?

  = correlation of input vector and comprehended vector



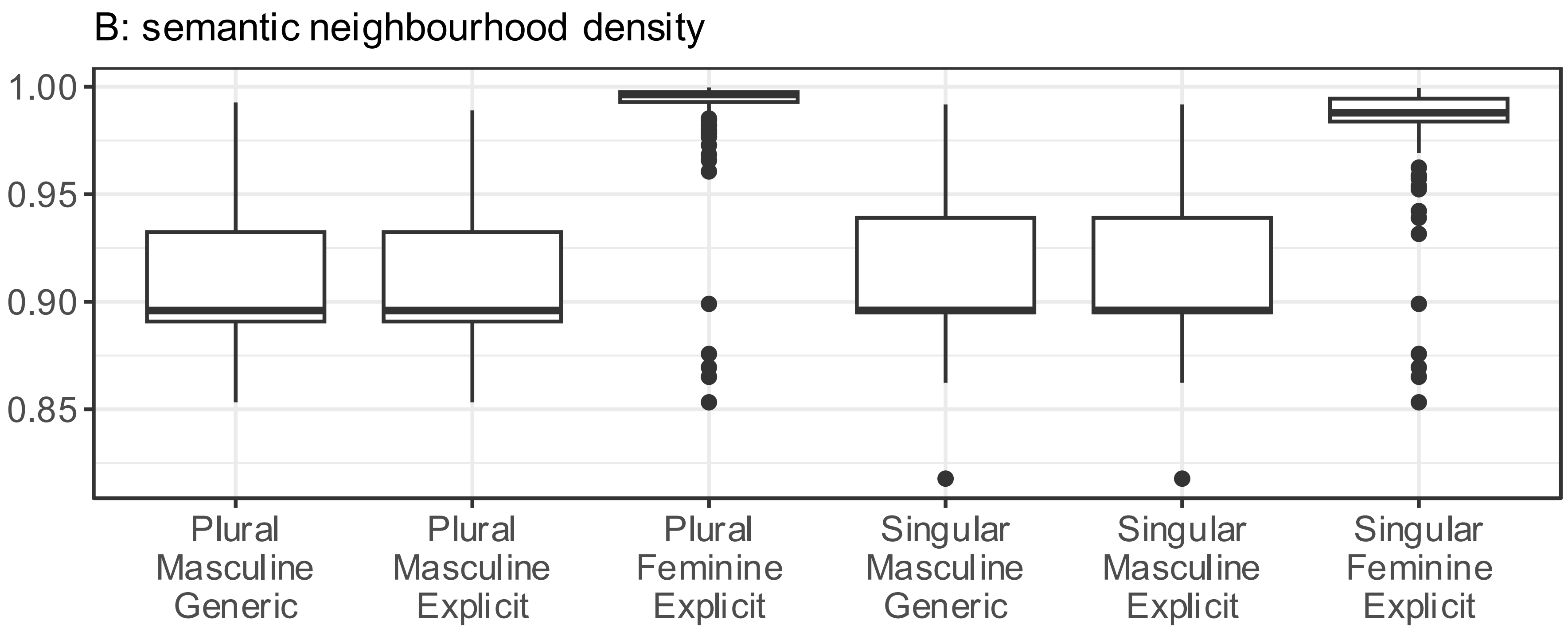
Schmitz et al. (2023)

Applying LDL: generic masculines in the mental lexicon

Measure 2: semantic neighbourhood density

- How dense is the semantic neighbourhood of a target?

  = mean correlation of 10 nearest neighbours



Schmitz et al. (2023)

Applying LDL: generic masculines in the mental lexicon

Measure 3: semantic activation diversity

- How strongly are semantic dimensions activated by the target?

  = Euclidean norm of the comprehended semantic vector

Schmitz et al. (2023)

Limitations of the studies so far

- In the NDL/LDL study, we treated **genericity** as if it were a **grammatical exponent** like number or grammatical gender, although it isn't one
- In the instance vector study, **targets** were represented by instance vectors, **non-targets** were not
- This introduces a **representational asymmetry**: we force distinctions on the target forms that the rest of the sentence does not encode

Schmitz et al. (2023); Schmitz (2024)

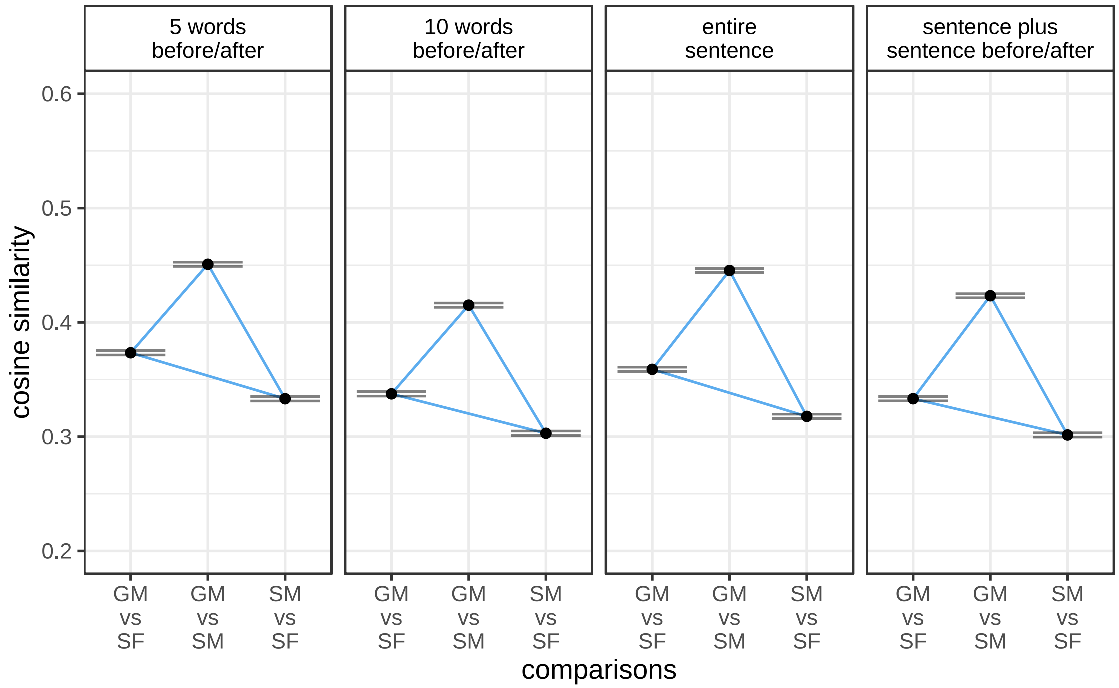
LLMs as a solution

- Large language models (e.g. GPT2, BERT, etc.) provide contextualised token embeddings
- Each token receives a vector that already integrates
  - its local morphosyntax,
  - its sentence semantics,
  - broader discourse context
- LLMs do not require us to specify genericity for the model
  - Those distinctions (if present) must emerge from the context itself
  - This eliminates the asymmetry: all tokens receive equally rich contextual representations

Schmitz et al. (submitted)

LLMs: using BERT to model generic masculines

- Even with rich discourse context, generic masculines remain semantically closest to specific masculines
- Context might enlarge semantic distinction overall, but does not selectively pull generic masculine toward a gender-neutral meaning



Schmitz et al. (submitted)

LDL beyond the model: /ɐ/ duration

- NDL gave us semantic embeddings
- LDL added processing mappings (form → meaning; meaning → form)
- But so far, all results were abstract: activations, similarities, neighbourhoods
- The crucial question:
  Do these lexical-semantic differences have consequences for actual, measurable linguistic behaviour?
- LDL predicts that differences in meaning structure should modulate phonetic realisation in production (cf. Schmitz et al. 2021)

Schmitz (in prep.)

LDL beyond the model: /ɐ/ duration

- Two tasks: **reading** and **recall**
  - 20 masculine role nouns ending in -er (/ɐ/)
  - Sense disambiguated by short context (generic vs. specific)
- Result: generic /ɐ/ **longer** than specific /ɐ/
- The crucial question: why?

Schmitz (in prep.)

LDL beyond the model: /ɐ/ duration

C matrix

- Phonological trigrams used as form cues
- Each word type corresponds to a row; each trigram cue is a column
- Sparse, binary coding: cue present = 1, absent = 0

S matrix

- Context-informed vectors from BERT

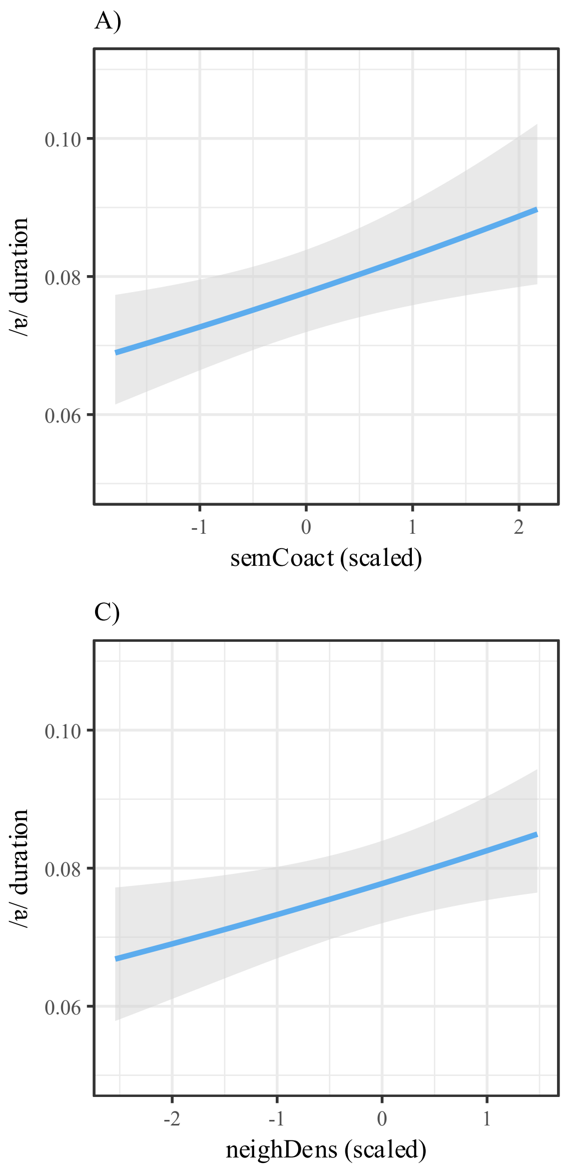
Mappings

- Comprehension: form → meaning
- Production: meaning → form

Schmitz (in prep.)

LDL beyond the model: /ɐ/ duration

- In both tasks, measures derived from the LDL model explain the duration of /ɐ/
- For example, in the reading data, generic masculines come with higher levels of semantic co-activation and denser neighbourhoods
- Higher values of semantic co-activation and denser neighbourhoods, in turn, come with longer /ɐ/ durations
- This is in line with the analysis independent of LDL, in which generic masculines show longer /ɐ/ durations



Schmitz (in prep.)

LDL beyond the model: /ɐ/ duration

- Generic and gender-specific masculines differ systematically in word-final /ɐ/ duration across tasks
- LDL measures show why
  - Generic senses produce broader semantic activation and denser semantic neighbourhoods → longer duration

Schmitz (in prep.)

Computational methods in gender linguistic research

distributional semantics

pre-processing

context

model selection

count-based

predictive

word2vec

subword

fastText

LLMs

BERT

instance vectors

discriminative learning

NDL

LDL

C

S

F

G

masculines' conflated meaning

LDL beyond the model

male bias confirmed

male bias explained

duration explained

Computational methods in gender linguistic research

- Distributional semantics approaches confirm previous findings: generic masculines come with a male bias
- NDL lets us reconstruct these meaning differences from word-to-word associations alone, confirming that semantics emerges from discriminative mappings
- LDL shows how these discriminative mappings shape comprehension, production, and subphonemic detail: differences in semantic co-activation and neighbourhood density predict duration asymmetries
- Context-sensitive embeddings demonstrate how grammatical information and contextual cues can be integrated, addressing long-standing problems in modelling genericity
- Across all methods, the result is robust: generic masculines come with a male bias

Schmitz et al. (2023); Schmitz (2024); Schmitz et al. (submitted); Schmitz (in prep.)

THANK YOU!

References 1

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, *12*(4), e0174623.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, *2019*, 4895891.

Baayen, R. H., & Smolka, E. (2020). Modelling morphological priming in German with naive discriminative learning. *Frontiers in Communication, Section Language Sciences*, *5*(17), 1–23.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*(1), 213–234.

Chuang, Y.-Y., & Baayen, R. H. (2021). Discriminative learning and the lexicon: NDL and LDL. *Oxford Research Encyclopedia of Linguistics*.

Chuang, Y. Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2020). The processing of nonword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*, *53*(3), 945–976.

Firth, J.R. (1957). *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis: 1–32. London: Longman.

Gabriel, U., Gygax, P., Sarrasin, O., Garnham, A., & Oakhill, J. (2008). Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German. *Behavior Research Methods*, *40*(1), 206–212.

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2–3), 146–162.

Boleda (2020)

References 2

Lapesa, G., Kawaletz, L., Plag, I., Andreou, M., Kisselew, M., & Padó, S. (2018). Disambiguation of newly derived nominalizations in context: A Distributional Semantics approach. *Word Structure*, *11*(3), 277–312.

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13) - Volume 2*.

Müller-Spitzer, C., Ochs, S., Rüdiger, J. O., & Wolfer, S. (2025). *Geschlechtsübergreifende Maskulina im Sprachgebrauch: Eine korpusbasierte Untersuchung zu lexemspezifischen Unterschieden*.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Schmitz, D. (2024). Instances of bias: The gendered semantics of generic masculines in German revealed by instance vectors. *Zeitschrift für Sprachwissenschaft*, *43*(2).

Schmitz, D. (in prep.). Autohyponymy and the phonetic signal: The duration of word-final -er in German.

Schmitz, D., Ochs, S., Rüdiger, J. O., & Müller-Spitzer, C. (submitted). Context fails to neutralise the male bias of generic masculines.

Schmitz, D., Plag, I., Baer-Henney, D., & Stein, S. D. (2021). Durational differences of word-final /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning. *Frontiers in Psychology*, *12*.

Schmitz, D., Schneider, V., & Esser, J. (2023). No genericity in sight: An exploration of the semantics of masculine generics in German. *Glossa Psycholinguistics*, *2*(1).

Thater, S. (2011). Distributionelle Semantik [PowerPoint slides]. Seminar Allgemeine Linguistik (Computerlinguistik), Universität des Saarlandes. https://www.coli.uni-saarland.de/courses/ds-11/material/01-Intro.pdf

Boleda (2020)

Subword models: *fastText* - technical detail

- *fastText* uses the Skip-Gram architecture (like *word2vec*), expanded to include subwords
- The model tries to maximise the probability of context words c given the target word w

$$\log p(c|w)$$

implemented via Negative Sampling, so the model learns to

- increase similarity between the word/subwords and real context words
- decrease similarity between the word/subwords and random “negative” words

Subword models: *fastText* - technical detail

- Training loop (simplified)

1. Input word $w \rightarrow$ get its subword set $G(w)$
2. Compute $v(w)$ as the sum of subword vectors
3. Predict context words
4. Update vectors of
   - the word
   - all its subwords
   - the context words
   - negative samples

Subword models: *fastText* - technical detail

- Prediction of context words

  1. The model takes a target word, e.g. *Lehrer*
  2. It computes the word's vector (sum of subword vectors + word vector)
  3. It compares this vector to the vectors of many other words in the vocabulary
  4. Words that are good neighbours in the corpus (e.g. *unterrichten*, *Schüler*, *Schule*) should be scored as "likely context words"
  5. The model adjusts its vectors until the real neighbours score higher than random words

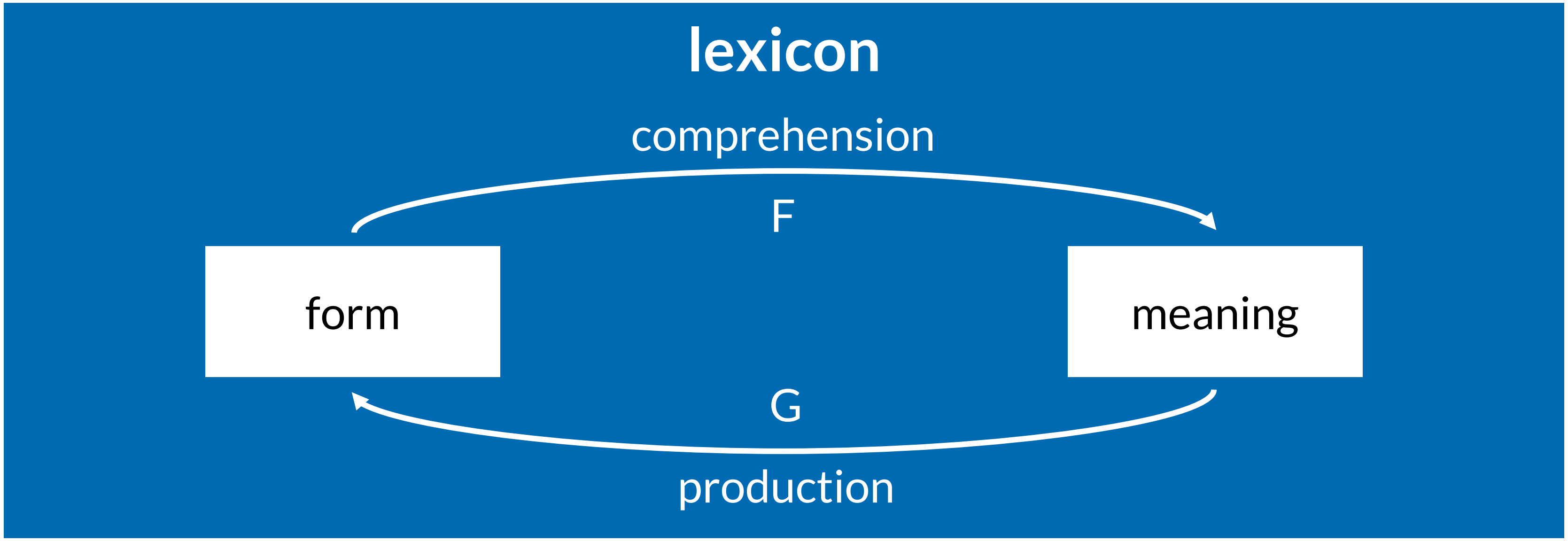
Applying NDL: modelling generic and specific meanings

- The semantic space derived from actual usage shows male-biased structure
- Generic masculines do not form a gender-neutral semantic category
- Instead, their contextual distribution mirrors the male-specific meaning
- This matches previous findings

Schmitz et al. (2023)

What is LDL?

A linear learning model in which **form vectors** and **meaning vectors** are linked through **linear mappings** that are learned from experience



Baayen et al. (2019); Chuang & Baayen (2021)

Applying LDL: generic masculines in the mental lexicon

- Across all comprehension-based measures, generic masculine forms pattern almost identically with specific masculine forms
- Specific feminine forms diverge clearly
- This reproduces the pattern established in experimental work and distributional-semantic modelling: generic masculines align with male-specific meanings rather than forming a gender-neutral category
- The male bias persists, and is in fact most visible, when form and semantics are treated as jointly learned and mutually constraining, as assumed in the discriminative mental lexicon

Schmitz et al. (2023)

LLMs: using BERT to model generic masculines

- Previous models used limited context (*fastText*: narrow windows; NDL/LDL: sentence-level cues)
- Human annotators rely heavily on wider context to disambiguate generic masculines (Müller-Spitzer et al. 2025)
- With contextualised embeddings, we can test whether richer context helps neutralise the male bias

Schmitz et al. (submitted)

LLMs: using BERT to model generic masculines

- 6,114 annotated tokens of 17 role nouns from DeReKo full texts (2006–2020)
- Each token hand-coded as
  - generic masculine
  - specific masculine
  - specific feminine
- Covariates: number, definiteness, stereotypicality

Schmitz et al. (submitted)

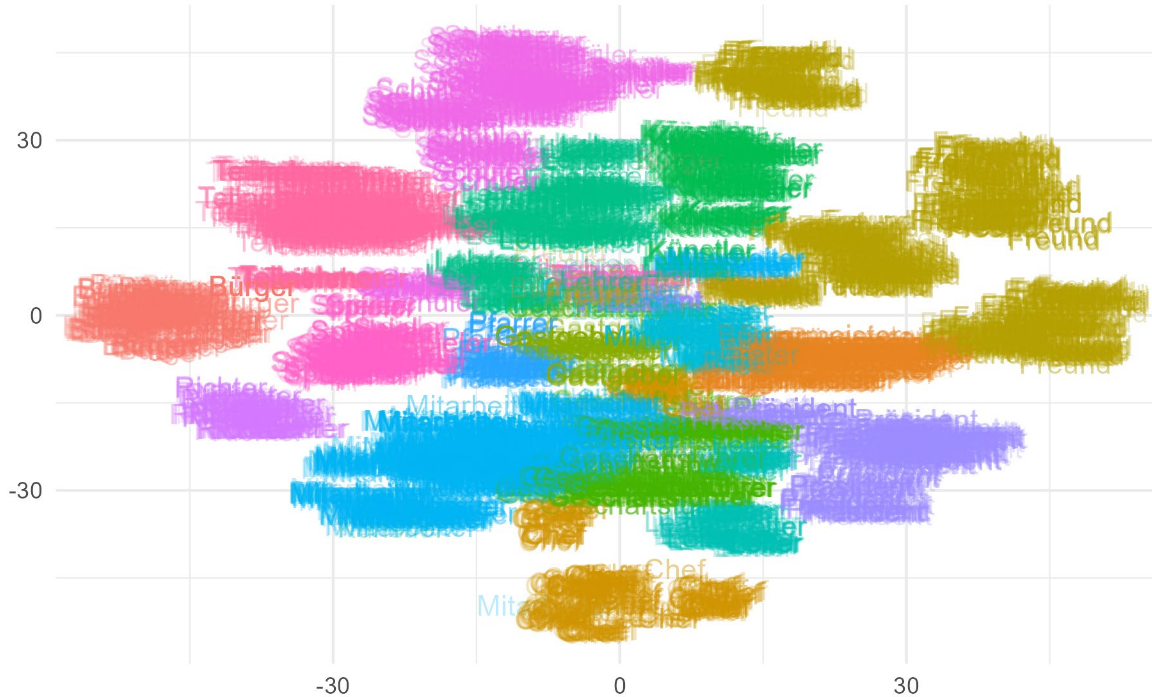
LLMs: using BERT to model generic masculines

- Model: *bert-base-german-cased*
- Input tokens processed with four context window sizes
  - ±5 words
  - ±10 words
  - full sentence
  - sentence ± one adjacent sentence on both sides
- Each token → one vector
- Token clustering (t-SNE) confirms lemma coherence and form coherence

https://huggingface.co/google-bert/bert-base-german-cased; Schmitz et al. (submitted)

Excursion: t-SNE

- t-Distributed Stochastic Neighbor Embedding (van der Maaten & Hinton, 2008)
- Aim: reduce high-dimensional vectors to 2 or 3 dimensions without loosing local patterns or structure between the vectors



LLMs: using BERT to model generic masculines

- For each target, compute cosine similarities between
  - generic masculine vs specific masculines,
  - generic masculine vs specific feminines,
  - specific masculines vs specific feminines
- Separate analyses for singular and plural
- 100-fold cross-validation to control for uneven token numbers
- Beta-regression to predict similarity patterns per context window

Schmitz et al. (submitted)