Wie können uns informationstheoretische Maße helfen, die Produktion und die Verarbeitung gesprochener Sprache zu modellieren?

Text+ Plenary

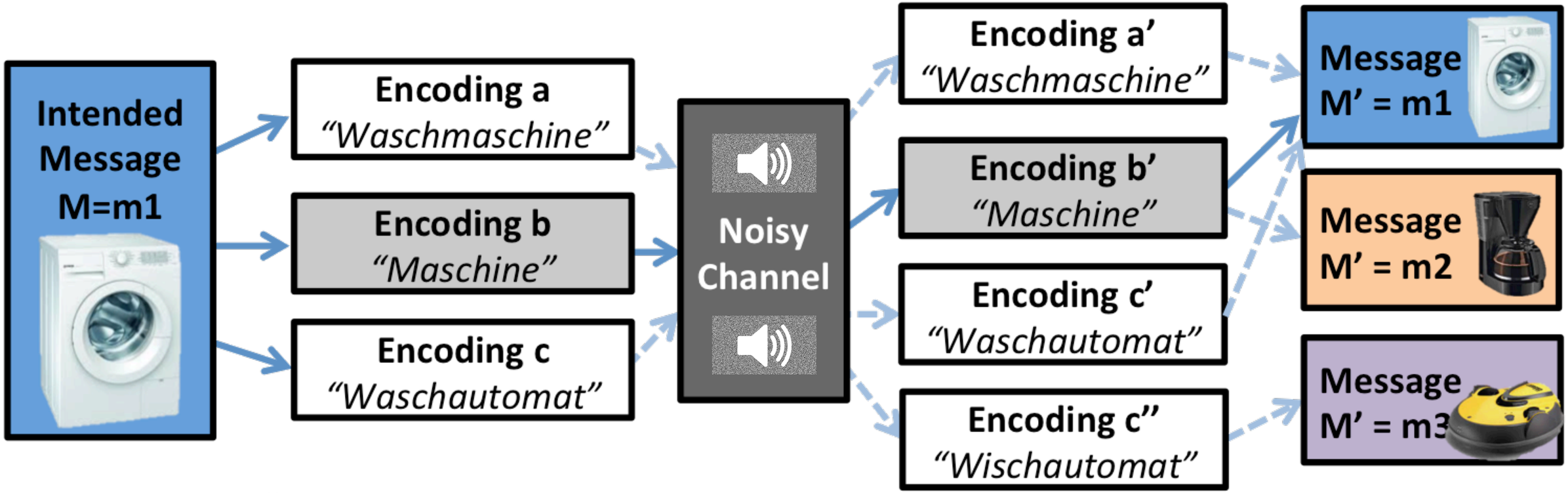
Vera Demberg, Sep 2023

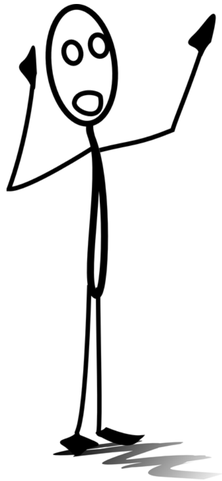
UNIVERSITÄT
DES
SAARLANDES

IDeaL
SFB 1102

European Research Council
Established by the European Commission

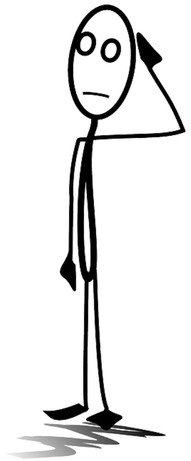
Noisy Channel Model als Modell für Kommunikation





Idee geht zurück auf Claude Shannon

- ursprünglich entwickelt für verrauschte Telefonleitung
- Ist es nützlich, diese Ideen auf menschliche Kommunikation zu übertragen?



Intended Message M=m1

Encoding a
“Waschmaschine”

Encoding b
“Maschine”

Encoding c
“Waschautomat”

Noisy Channel

Encoding a’
“Waschmaschine”

Encoding b’
“Maschine”

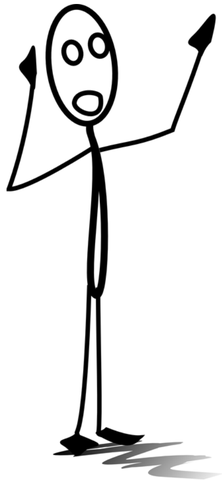
Encoding c’
“Waschautomat”

Encoding c”
“Wischautomat”

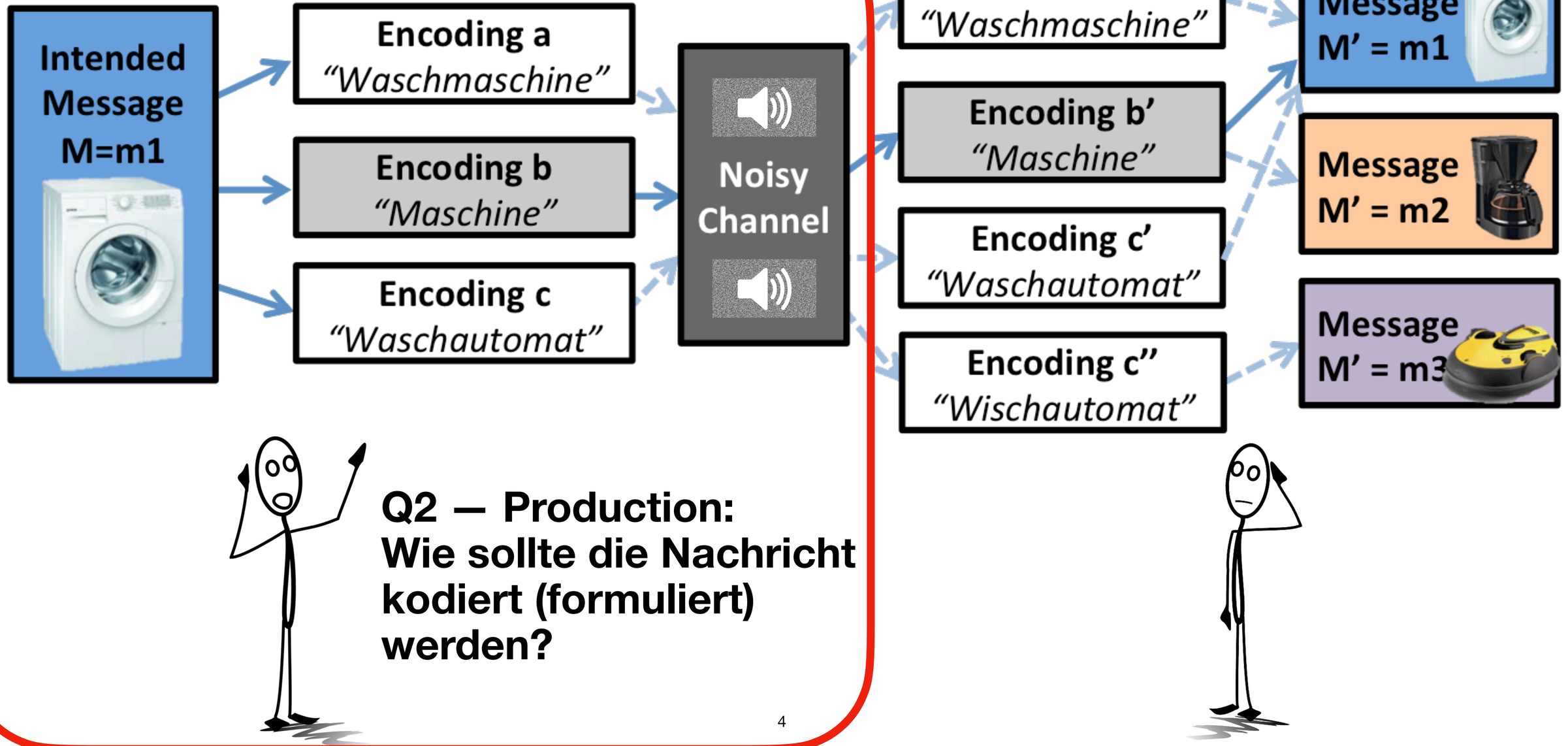
Message M’ = m1

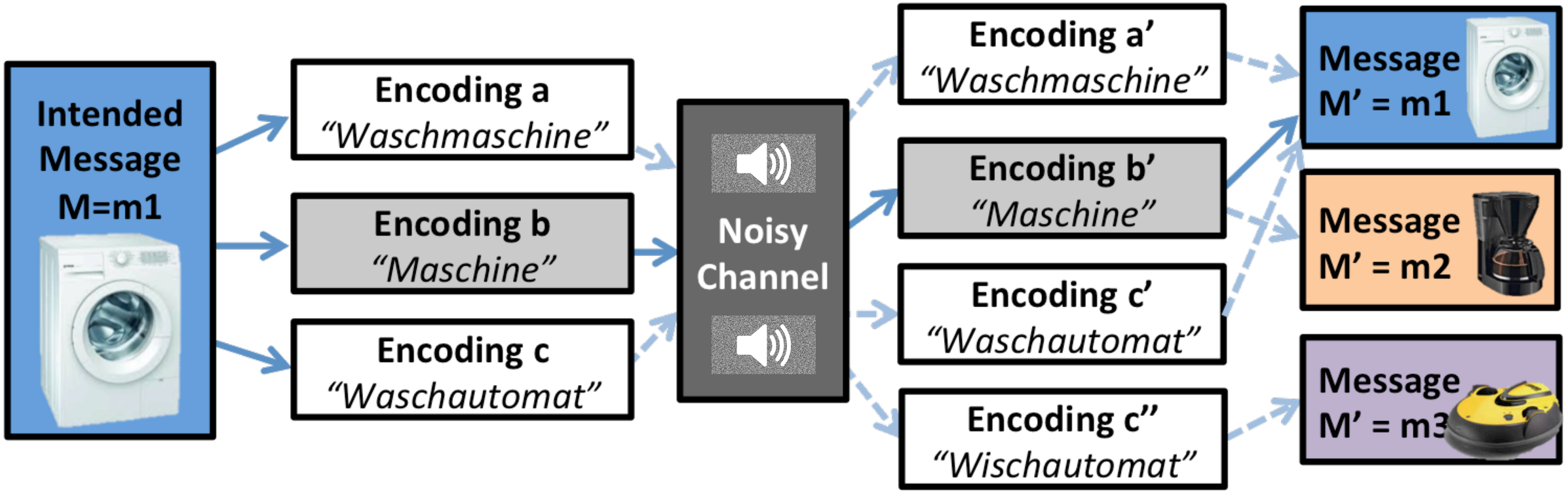
Message M’ = m2

Message M’ = m3



Q1 — Verstehen:
Wie wird ein verrauschtes Signal vom Hörer interpretiert?

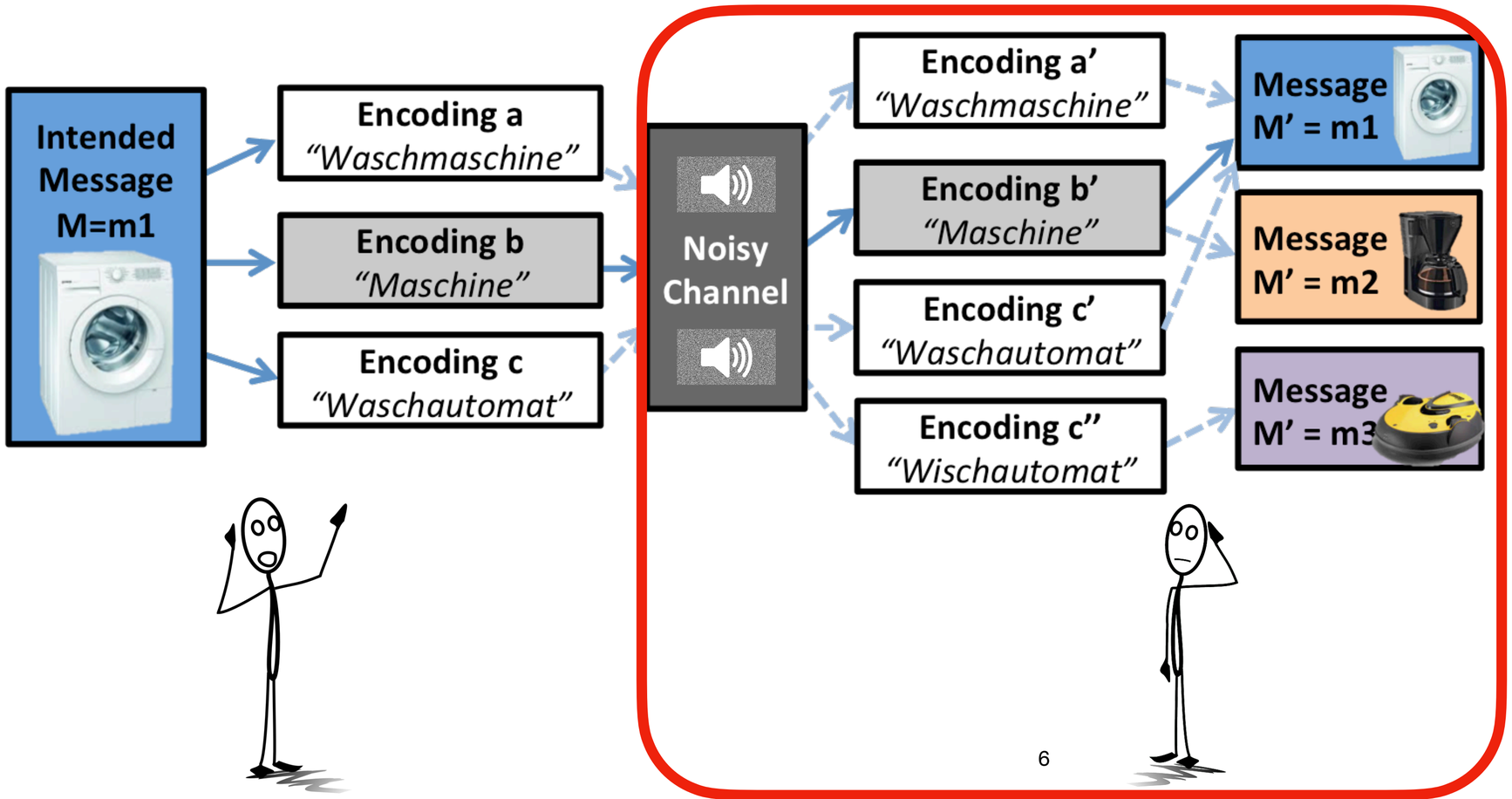

Intended Message M=m1
Encoding a "Waschmaschine"
Encoding b "Maschine"
Encoding c "Waschautomat"
Noisy Channel
Encoding a' "Waschmaschine"
Encoding b' "Maschine"
Encoding c' "Waschautomat"
Encoding c'' "Wischautomat"
Message M' = m1
Message M' = m2
Message M' = m3
Q2 — Production:
Wie sollte die Nachricht kodiert (formuliert) werden?



Q3 — Modellierung:
Das Noisy Channel model ist ein quantitatives Modell — für jedes Wort kann eine genaue Wahrscheinlichkeit der korrekten Übertragung berechnet werden.
Wie schätzen wir diese Wahrscheinlichkeiten am besten ab?

Verstehen

Q1 — Wie wird ein verrauschtes Signal vom Hörer interpretiert?



Model formalization:
understanding the meaning M' depends on the **(top-down) probability of M'** and **the (bottom-up) probability of the signal b'** given message M':

$$P(M'|b') \sim P(b'|M') * P(M')$$

acoustic model

language model

Sprachmodell

Wie berechnet man die Wahrscheinlichkeit einer Nachricht?

- abhängige Wahrscheinlichkeit eines Worts gegeben den Kontext:

  P(*Waschmaschine* | *Wir haben so viel dreckige Wäsche. Gibt es hier eine*)

- wird typischerweise durch Befragung von Menschen (cloze probability) oder durch Abschätzung aus einem Sprachmodell erhoben.

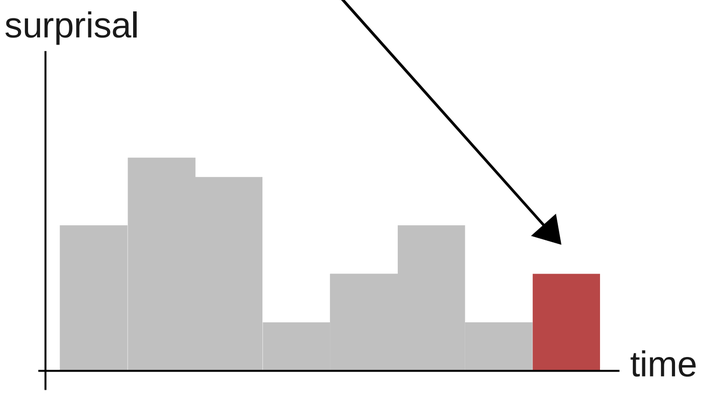
- diese Wahrscheinlichkeit hat eine hohe Bedeutung in der Psycholinguistik: Die negative log Wahrscheinlichkeit proportional ist zu messbaren Verarbeitungsschwierigkeit beim Sprachverstehen
  (e.g., Hale, 2001; Levy 2008; Demberg & Keller, 2008)

  surprisal (word) = - log P(w|context)

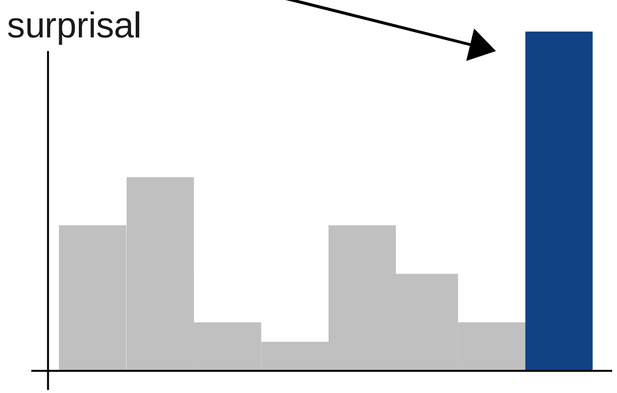
Surprisal

– log P(*stamp*| John accidentally mailed the letter without a)

– log P(*stamp*| John went to the shop to buy a)

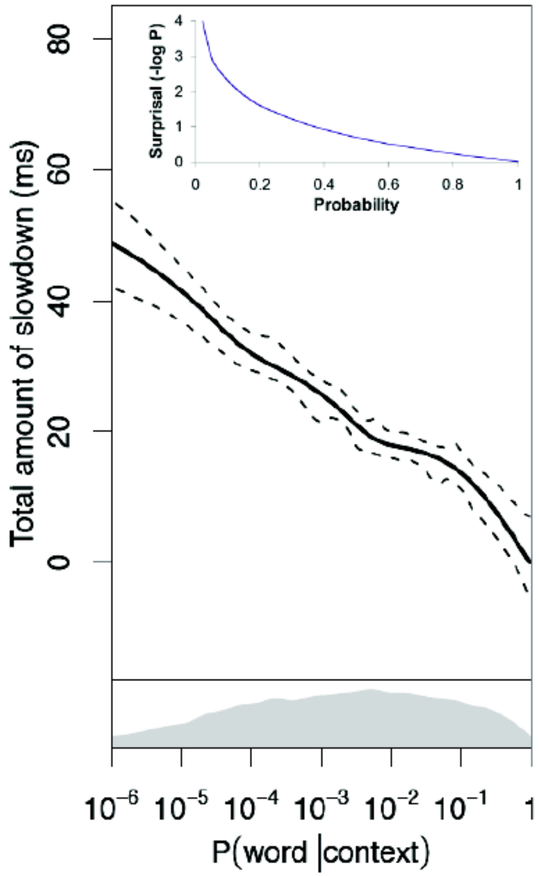


John mailed....without a stamp



John went... buy a stamp

Eye-tracking



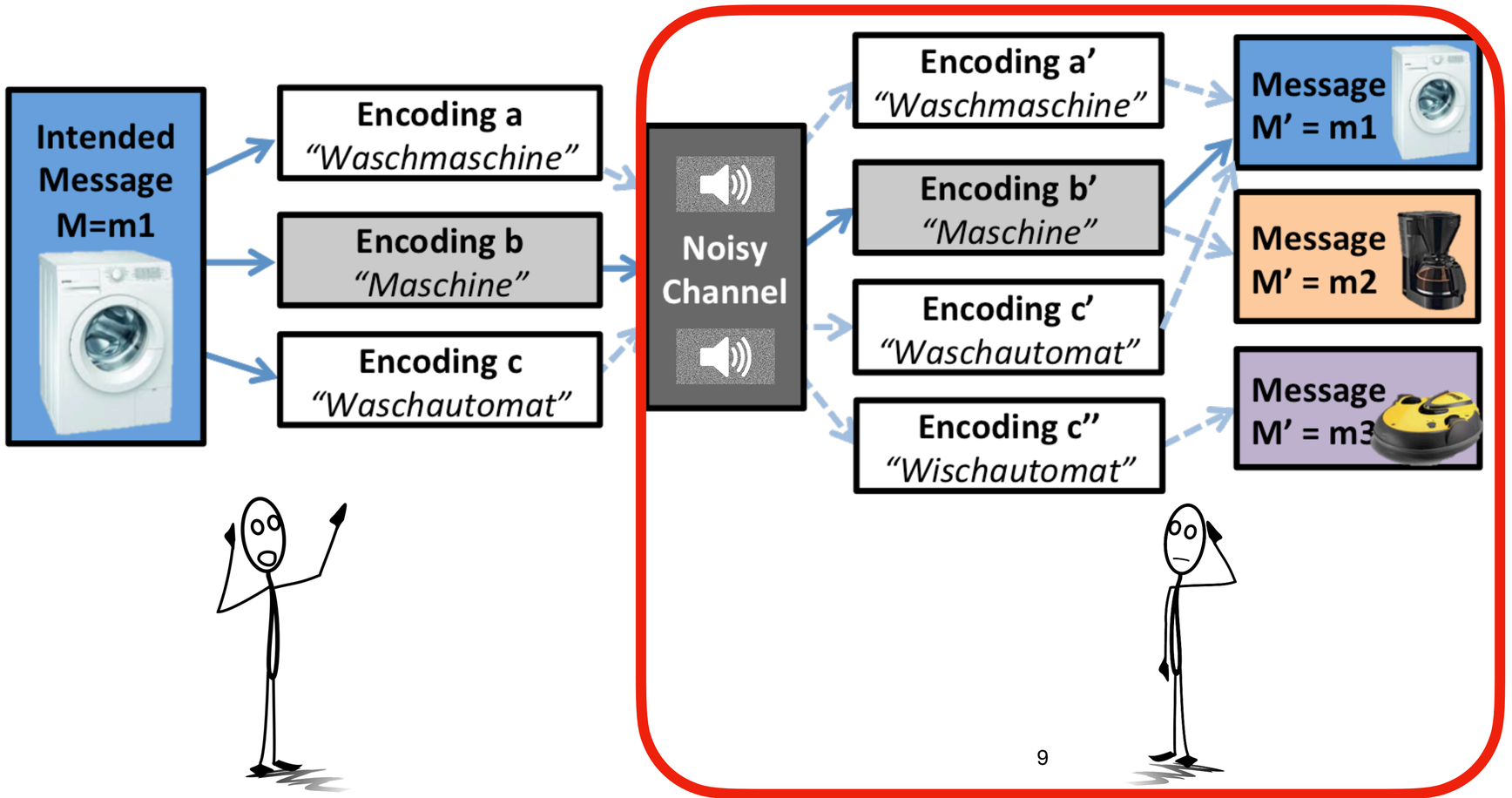
(from Smith & Levy 2013)

surprisal (word) = - log P(w|context)

- Surprisal erlaubt uns die Information, die durch ein Wort übertragen wird, zu quantifizieren.
- Surprisal korreliert gut mit Lesezeiten oder ERP Amplituden.

Verstehen

Q1 — Wie wird ein verrauschtes Signal vom Hörer interpretiert?



Heute möchte ich aber auf die Kombination der Informationen aus dem Sprachmodell und dem akustischen Modell eingehen.

$$P(M'|b') \sim P(b'|M') * P(M')$$

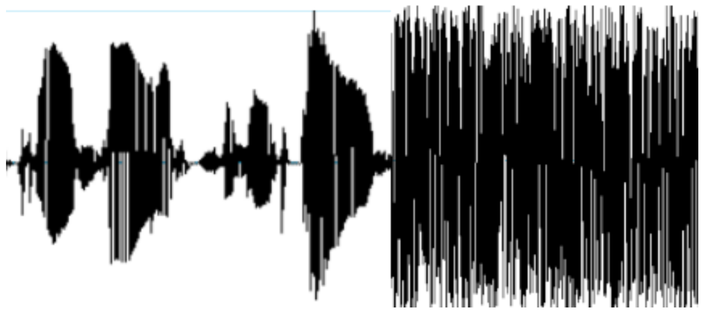
acoustic model

language model

Verstehen gesprochener Sprache

- Rauschen kann das akustische Sprachsignal stören (Brungard, 2001; Cooke, 2009)

- Energetic masking:



- verschiedene Laute werden unterschiedlich stark durch verschiedene Rauschquellen (z.B. Postersession vs. Flugzeuggeräusch) beeinträchtigt.

Experimentelles Design

- Sprachverstehen mit / ohne Störgeräusch
  - **Babble** noise and **white** noise at -5 dB SNR
- kritische Minimalpaare von Wörtern, z.B. “Liebe” / “Liege”
  - **Vowel** pairs, **plosive** pairs, **fricative** pairs
- Einbettung in Sätze, die das kritische Wort erwartbar machen (low surprisal) vs. wo es überraschend ist (high surprisal)

Low surprisal: Nach vier Jahren heiratete Paul seine große **Liebe**.
High surprisal: Nach vier Jahren heiratete Paul seine große **Liege.**

} minimal pair

counter-balancing:
Low surprisal: Am Hotelpool gab es nur noch eine freie **Liege**.
High surprisal: Am Hotelpool gab es nur noch eine freie **Liebe**.

Van Os, Kray, & Demberg, 2022
Frontiers in Psychology.

Analyse

High surprisal item: Am Hotelpool gab es nur noch eine freie **Liebe**.

- Antworten wurden annotiert als:
  - **Target**: *Liebe*
  - **Distractor:** *Liege*
  - **Wrong**: *anything else*

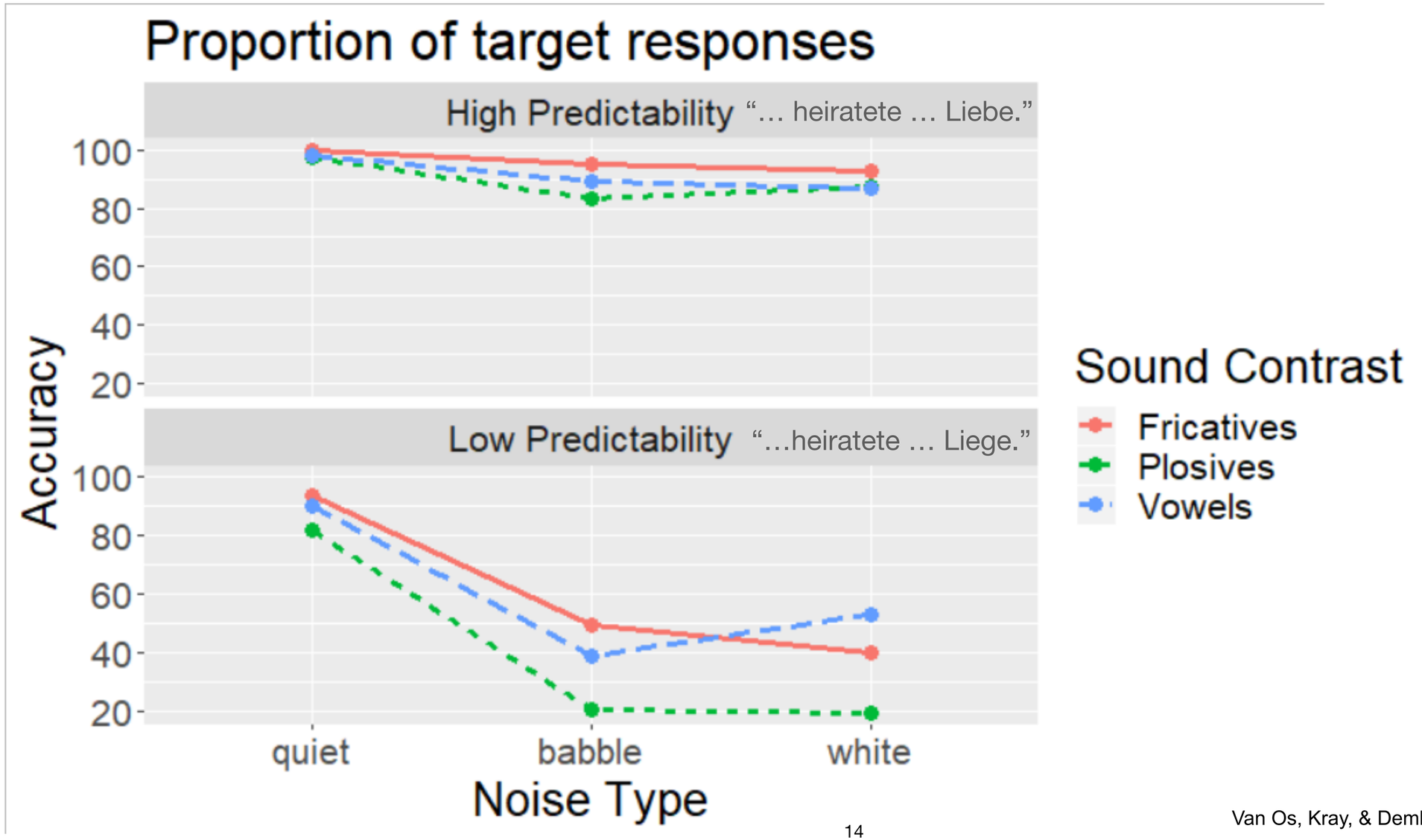
(Analyse hier beschränkt sich auf Akkuratheit in Bezug auf target.)

Van Os, Kray, & Demberg, 2022, Frontiers in Psychology.

Ergebnisse

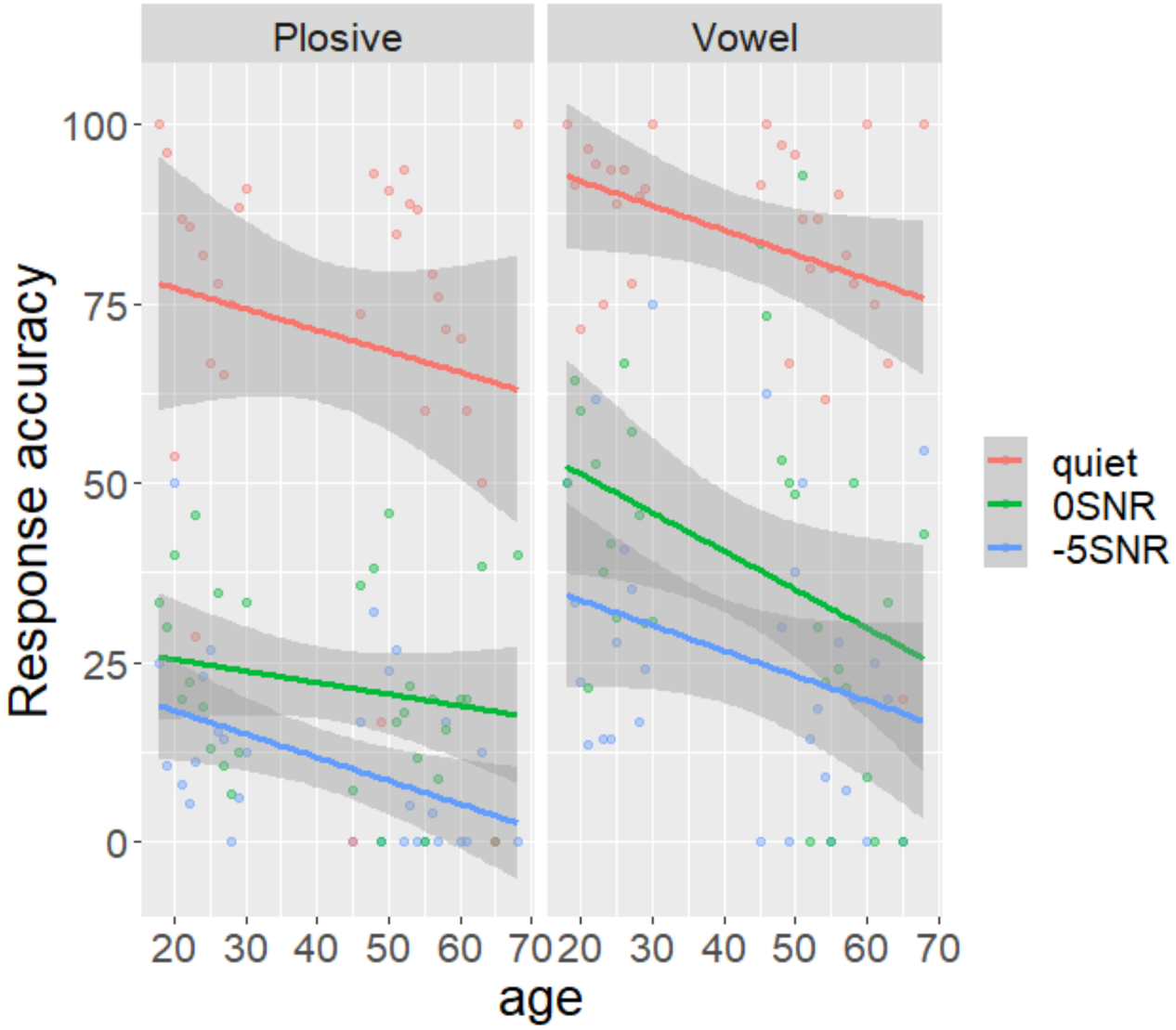
Van Os, Kray, & Demberg, 2022, Frontiers in Psychology.

Ergebnisse



Van Os, Kray, & Demberg, 2022, Frontiers in Psychology.

Ergebnisse (replication with wider age range)



- Im Alter verschlechtert sich das Hörvermögen. Daher ist das Sprachsignal für ältere Erwachsene weniger “verlässlich”.
- konsistente Ergebnisse:
  Plosive sind schwieriger als Vokale.
- **Ältere Erwachsene** geben häufiger an, das Wort, das inhaltlich passt, aber nicht im Signal vorhanden war, gehört zu haben.

=> konsistent mit noisy channel model

Van Os, Kray, & Demberg, 2022, Frontiers in Psychology.

Zusammenfassung Sprachverstehen

- Ergebnisse sind konsistent mit Noisy Channel Model:
  — das Rauschen beeinflusst, wie sehr sich Hörer auf das Signal verlassen.
  — Alter (Hörvermögen) beeinflusst, wie sehr sich Hörer auf das Signal verlassen

- Trade-off zwischen Kontexterwartung und Signal als rationale Kombination der beiden Wahrscheinlichkeiten.

Van Os, Kray, & Demberg, 2022, Frontiers in Psychology

praktische Anwendbarkeit

- Dialogsysteme könnten in lauten Umgebungen ihre Formulierungen so wählen, dass die Nachricht trotz Rauschen leichter verstanden werden kann.

- z.B. durch Wahl von Synonym oder alternativer syntaktischer Formulierung.

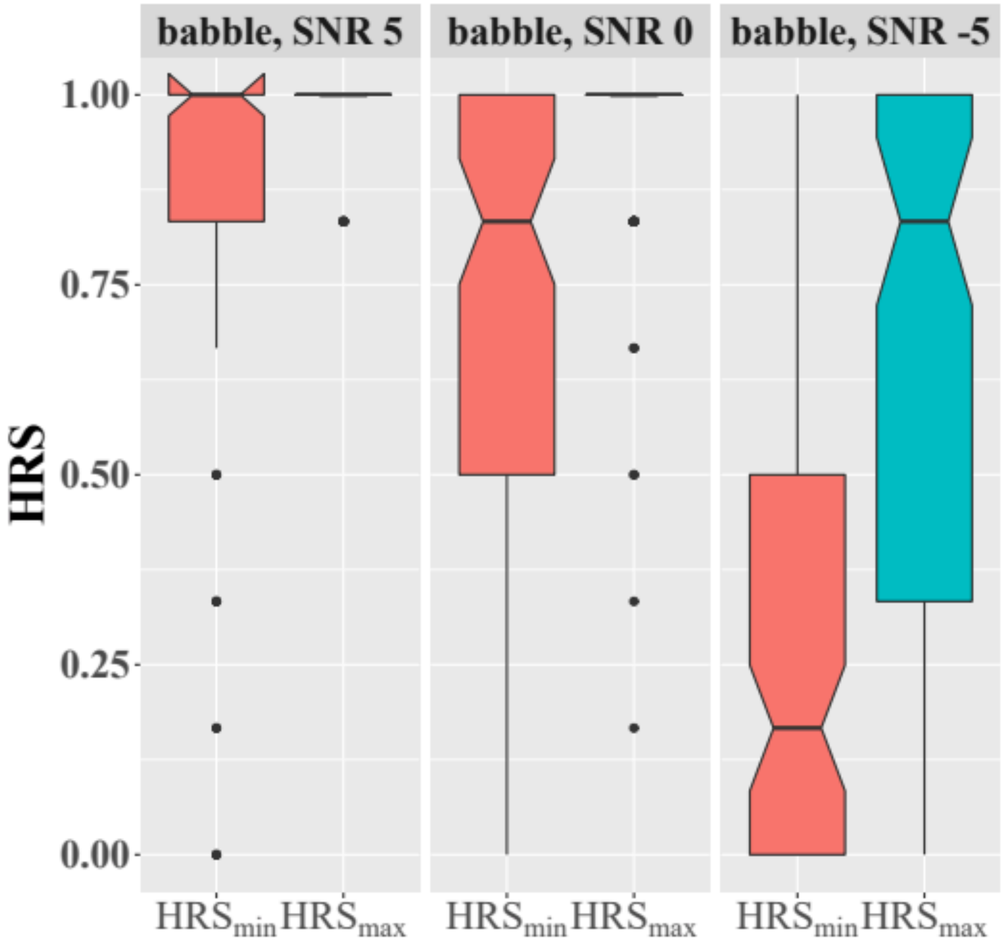
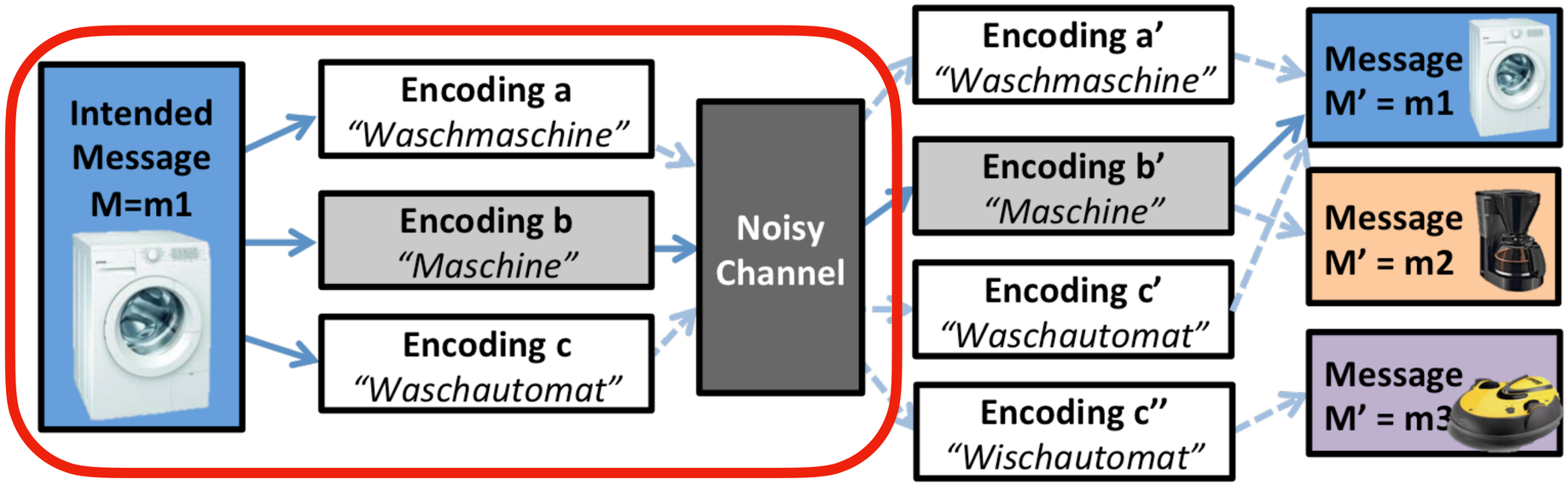


Figure 2: *Distinction between less intelligible and more intelligible synonyms, when they were presented* ***with context*** *under different listening environments.*

Chingacham, Demberg and Klakow, 2021
Interspeech best paper award

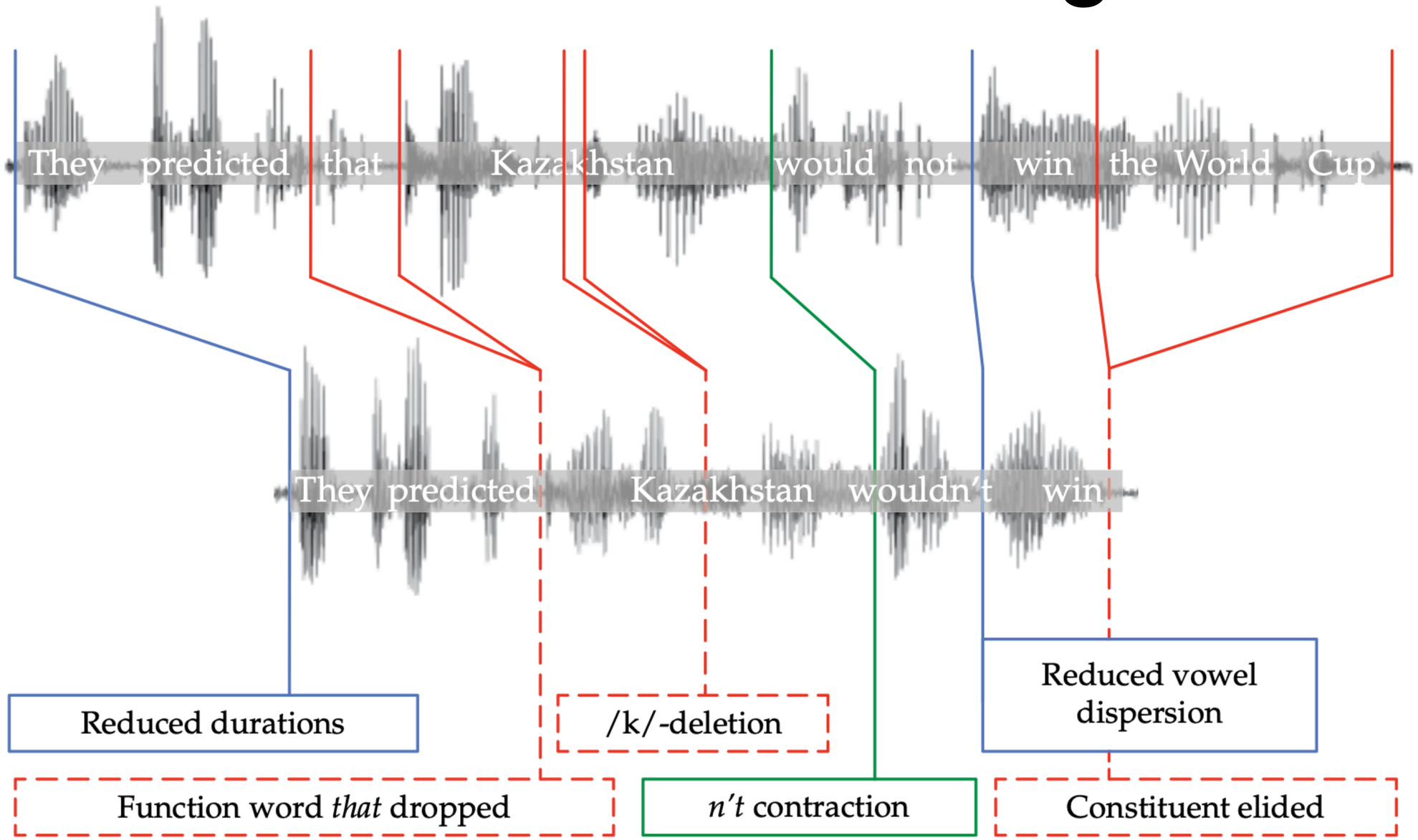
Sprachproduktion

optimale Auswahl der Formulierung



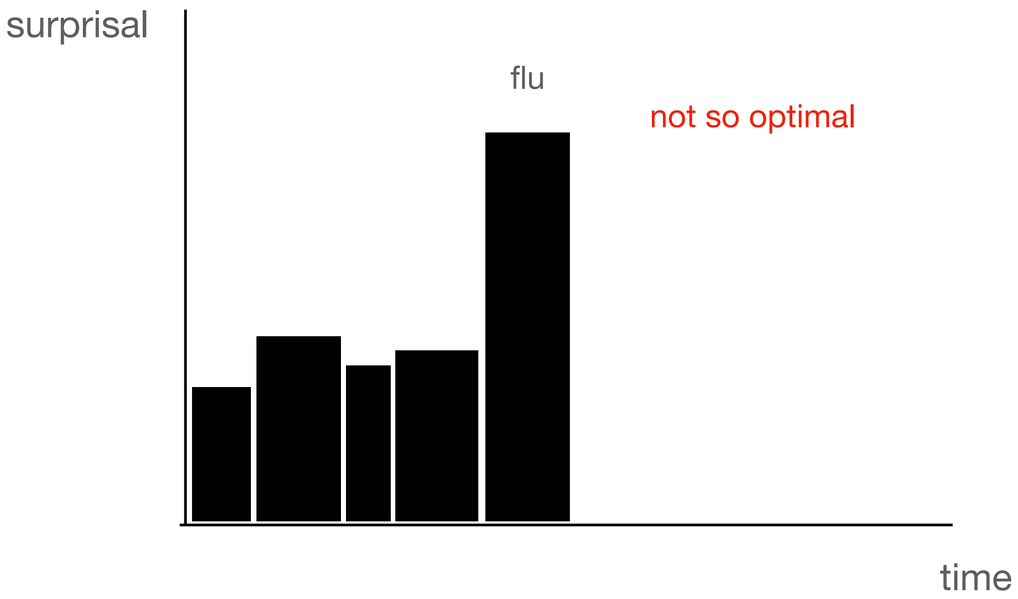
Konstanz im sprachlichen Informationsfluss (Fenk & Fenk-Oczlon, 1980)
Smooth signal redundancy hypothesis (Aylett and Turk, 2004)
Uniform information density hypothesis (Frank and Jaeger, 2008)
Expectancy hypothesis (Arnold 2001)

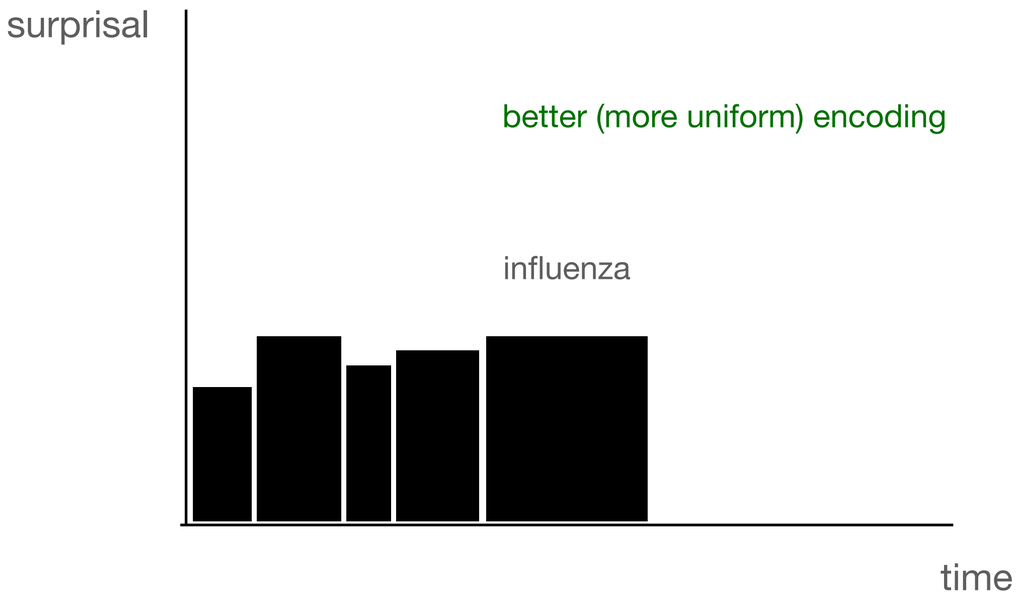
Sprecher haben viele Wahlmöglichkeiten



Wie kann Informationstheorie helfen, Wortwahl zu erklären?

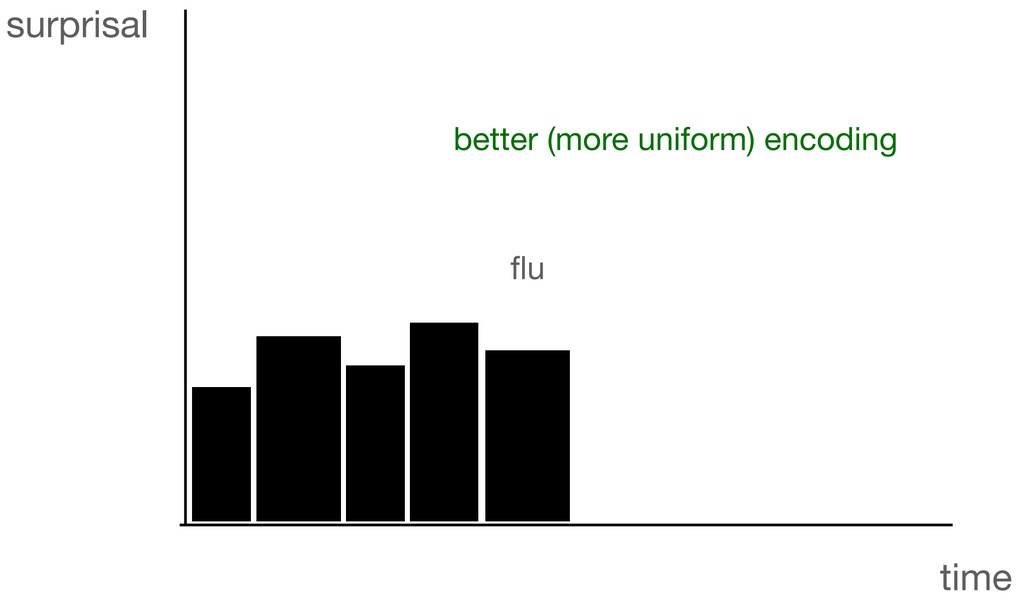
- Hängt Wortwahl vom Informationsgehalt (surprisal) ab?
- Beispiel: wenn ich “Grippe” ausdrücken möchte, Wahl: “flu” / “influenza”

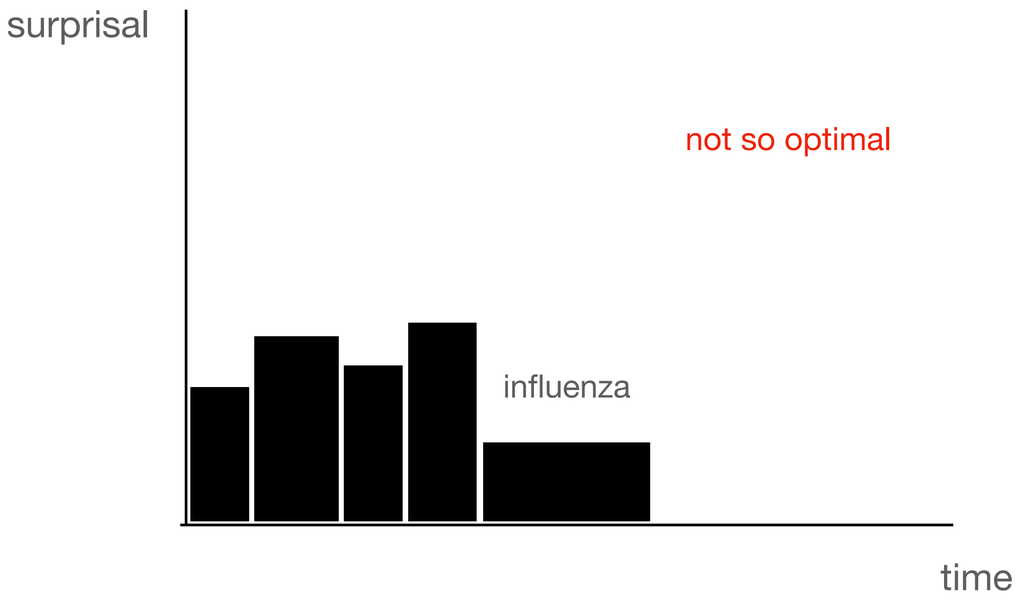




Wie kann Informationstheorie helfen, Wortwahl zu erklären?

- In einem anderen Kontext könnte “Grippe” sehr erwartbar sein.
- Optimale Wahl zwischen “flu” und “influenza” geht hier anders aus.





Empirische Überprüfung der Hypothese

- Nutzen Sprecher **längere Worte** für Bedeutungen, die **nicht vorhersagbar** sind, und **kürzere Worte** für **vorhersagbare** Bedeutungen?
- Methodologische Herausforderungen: Wortbedeutung konstant halten.

Mahowald et al. study (2013)

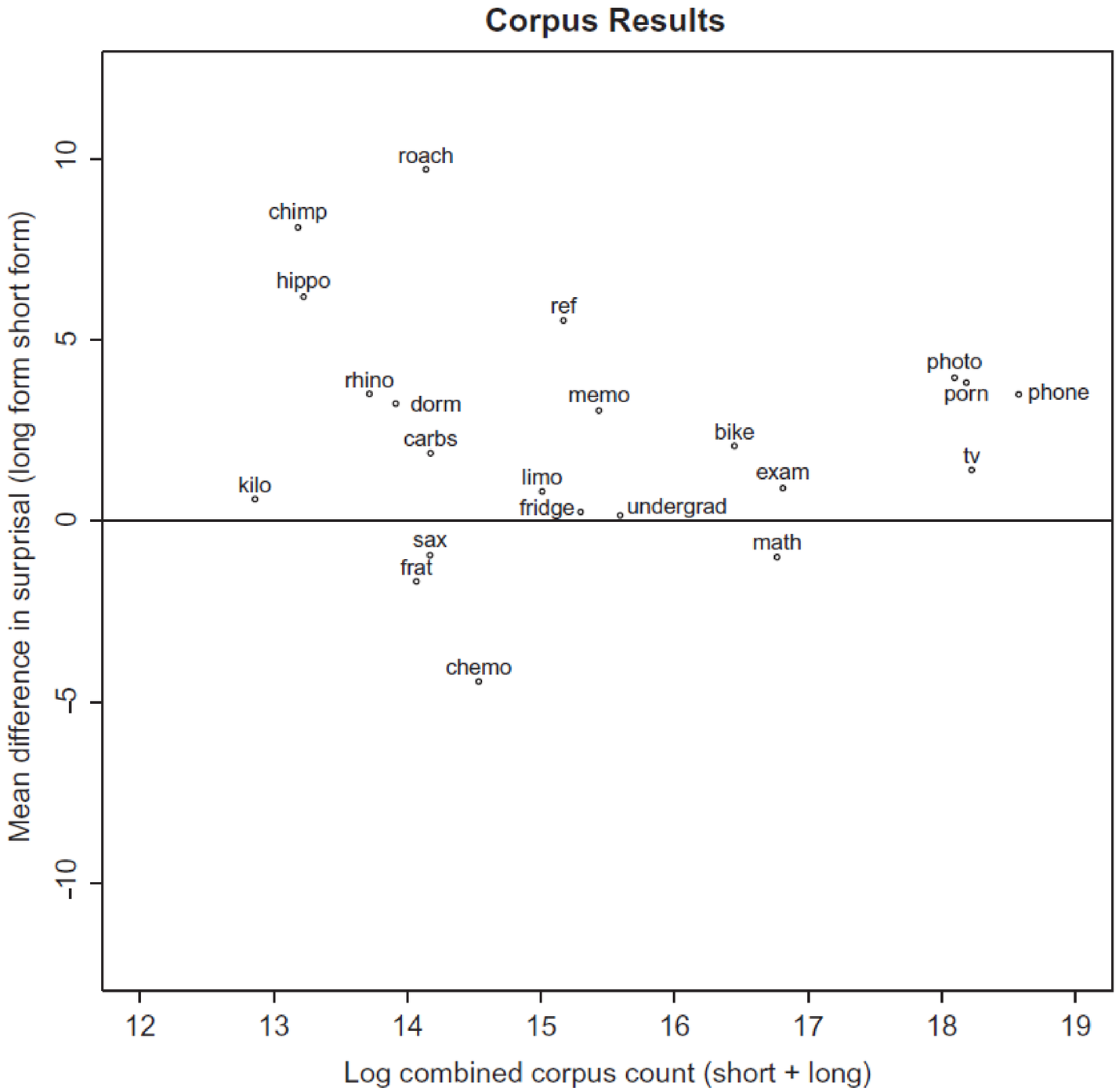
Test for correlation between length and information content while keeping the meaning constant.

chimp / *chimpanzee*
math / *mathematics*
chemo / *chemotherapy*

Resultate Korpusanalyse

Instanzen oberhalb der Linie sind konsistent mit Hypothese:

kurze Form ausgewählt in Kontexten, in denen das Konzept vorhersagbar war.



Experiment (forced choice)

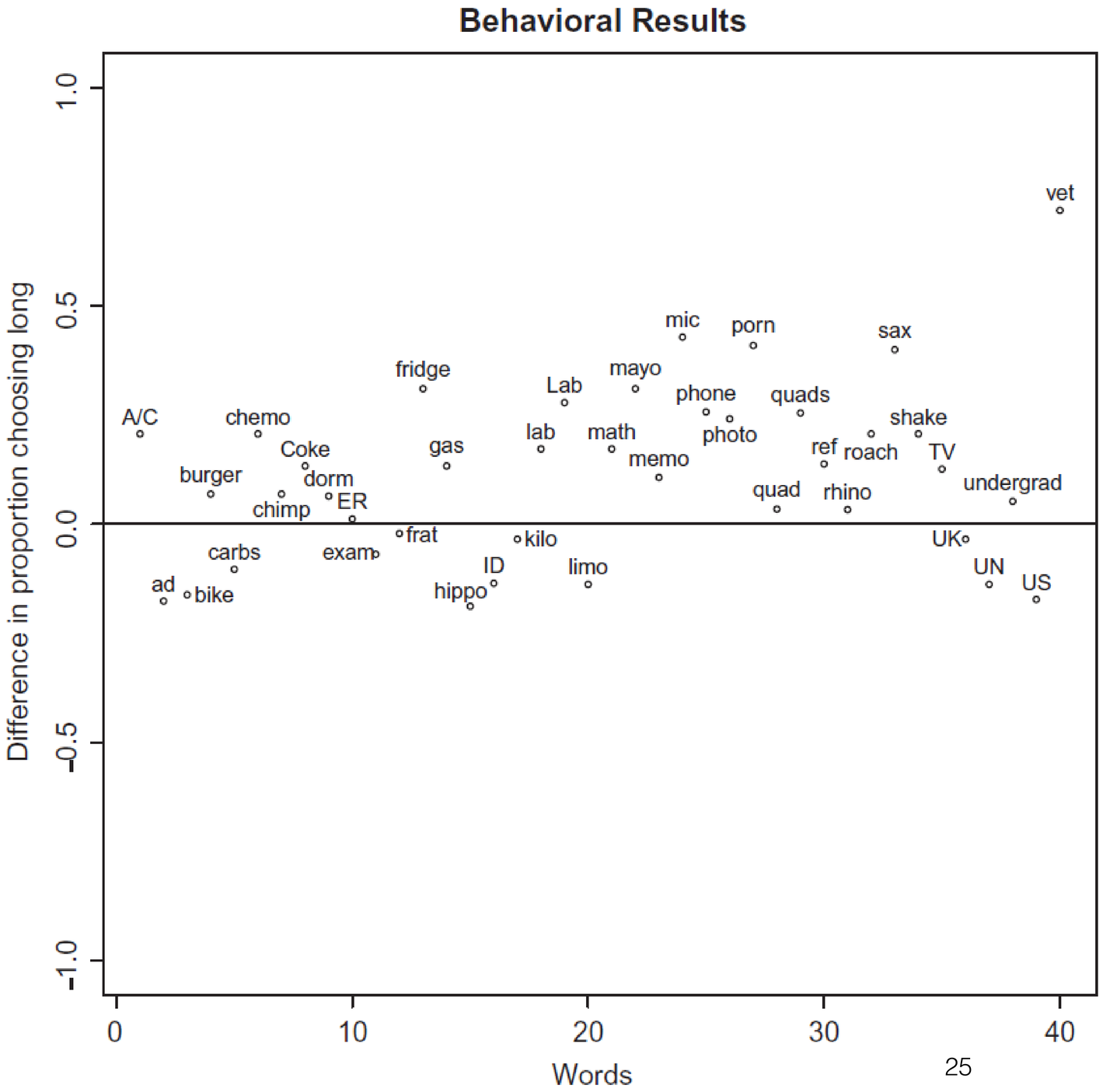
Forced choice completion test with neutral vs. predictive context:

Example item: math / mathematics

Supportive: Susan was very bad at algebra, so she hated ...

Neutral: Susan introduced herself to me as someone who loved ...

(Mahowald et al., 2013)



- Experimentelle Ergebnisse konsistent mit Korpusanalyse.
- Aber *Genre* (formal / informal) könnte Ergebnisse beeinflussen.
- Kontext / Genre war kein kontrollierter Faktor.

Experiment mit kontrollierteren Kontexten

- Kompositum vs. einfaches Wort (Weinglas vs. Glas)
  => sehr häufig und produktiveres Phänomen als Abkürzungen
- Kontext vor dem kritischen Wort soll gleich gehalten werden.
- Vorhersagbarkeit manipuliert durch Vorkontext.
- 36 items, 80 Teilnehmer.

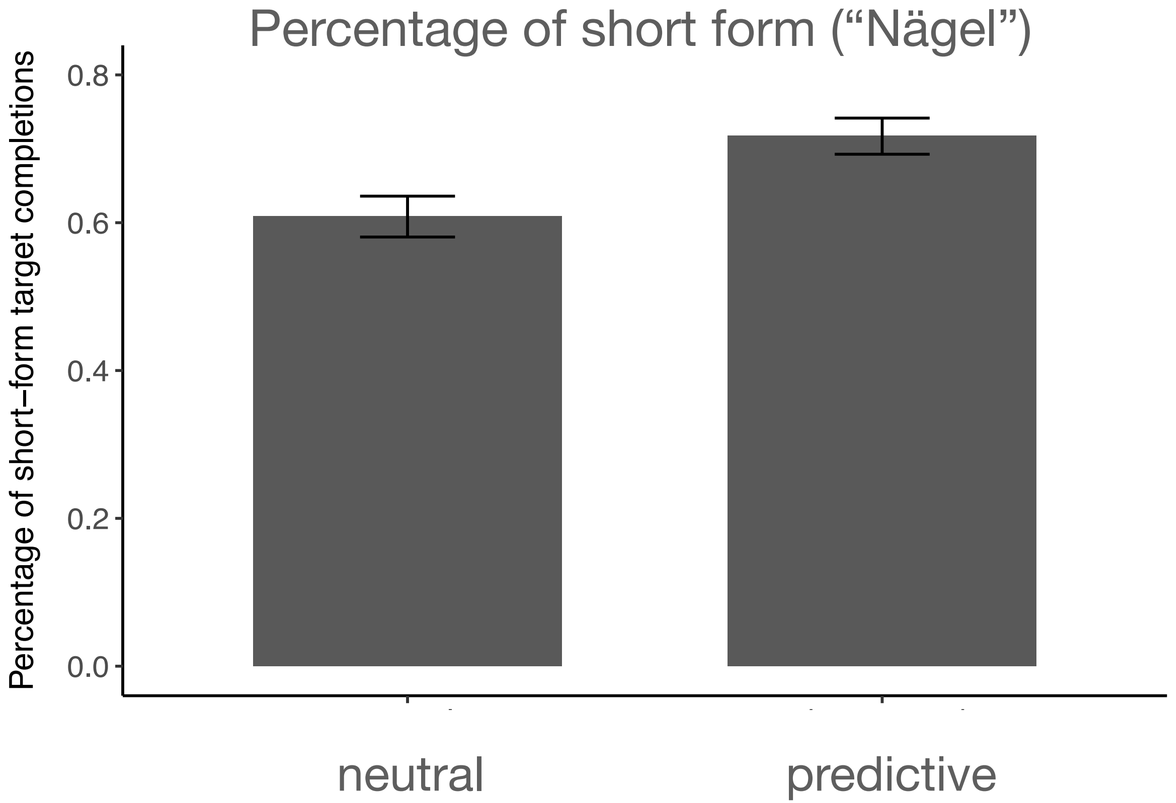
Experiment mit dt. Komposita

Predictive context:
Jeden zweiten Samstag ging Carola zur **Maniküre.**

Neutral context:
Carola war ihr Aussehen sehr wichtig.

Target:
Sie liebte es, wenn ihre **Nägel / Fingernägel** farblich auf ihr Outfit abgestimmt waren.

Ending:
Diesmal hatte sie sich für einen grellgrünen Lack entschieden.



- Kurze Formen wurden im Kontext mit höherer Erwartbarkeit präferiert ($p<0.01$).

Zarcone and Demberg, 2021

Zusammenfassung

- Wortwahl hängt von kontextueller Erwartbarkeit des Konzepts ab.
- Experimentelle Evidenz für große Breite an Phänomenen.
- kontroverser Fall in der Literatur: Pronominalisierung

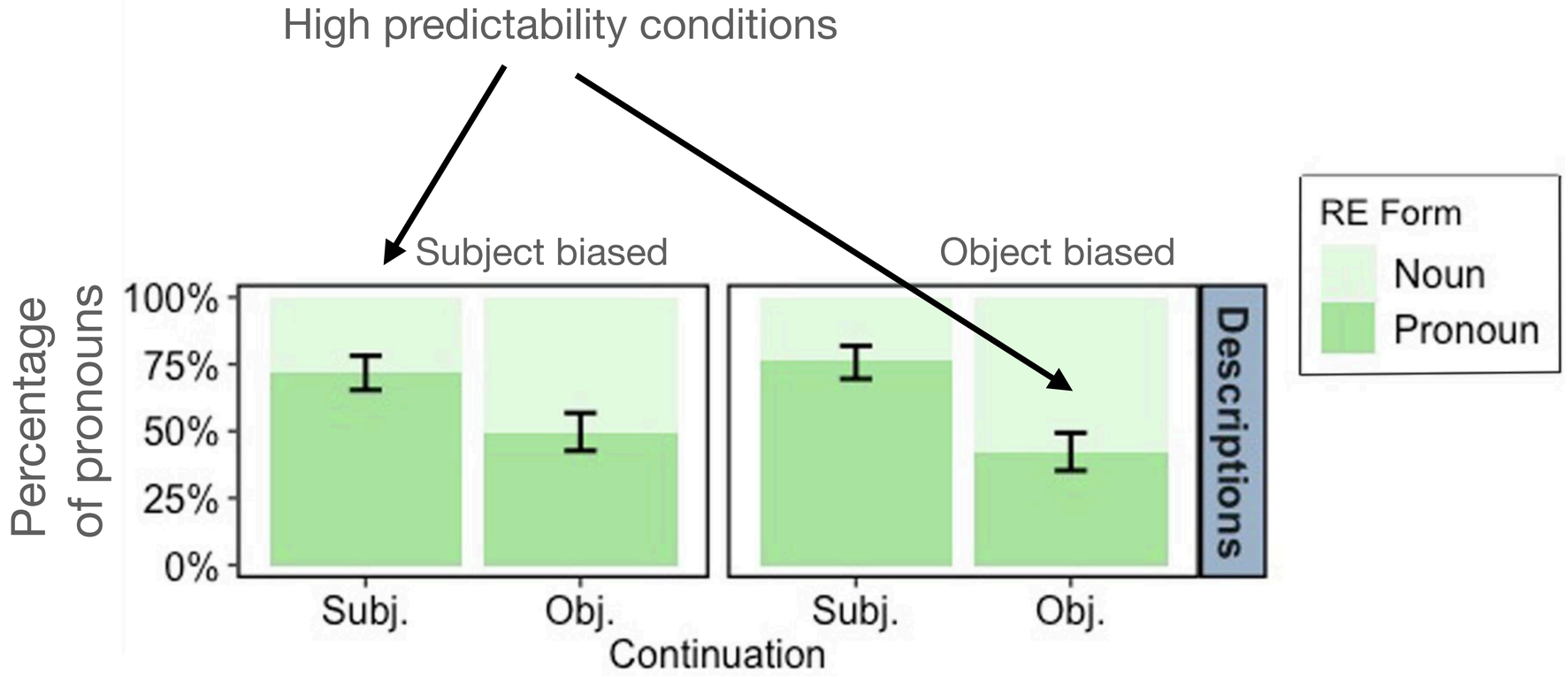
Pronominalization

- Do speakers choose a pronoun (as opposed to a name of full noun phrase) when the referent is predictable?
- Evidence on whether pronominalization follows UID predictions is highly controversial.

	reference task	antec. gender	verb type	interact/ context	effect found
Arnold (2001)	free	different	ToP	Yes	X
Fukumura and van Gompel (2010)	constr.	different	IC	No	–
Rohde and Kehler (2014)	free	same	IC	No	–
Holler and Suckow (2016)	constr.	same	IC	No	–
Rosa and Arnold (2017)-e1,e2	constr.	both	ToP	Yes	X
Rosa and Arnold (2017)-e3	constr.	both	ToP	No	(X)
Vogels (2019)	constr.	same	ToP,IC	No	X
Weatherford and Arnold (2021)	constr.	both	IC	Yes	X
Bott and Solstad (2023)-e1	constr.	both	ToP	No	X
Bott and Solstad (2023)-e2	constr.	both	IC	No	X
Bott and Solstad (2023)-e3	constr.	both	IC,AE	No	–
Bott and Solstad (2023)-e4	constr.	both	IC	No	X

Large-scale study

- Using a traditional paradigm
  ("*Peter admires / fascinates Paul because…*")
  we only found null-effects, even in highly powered studies.



=> predictability doesn't seem to make a difference!

Demberg, Kravtchenko and Loy, 2022

Paradigm with contextualized stimuli

Fantasy context

This is the sorcerer. He is skilled in ancient magic.

This is the psychic. She sees the future in her crystal ball.

This is the archer. He is never without his bow and arrow.

This is the witch. She specialises in glamour spells.

This is the thief. He steals when your back is turned.

This is the healer. She knows all about potions and antidotes.

Materials:

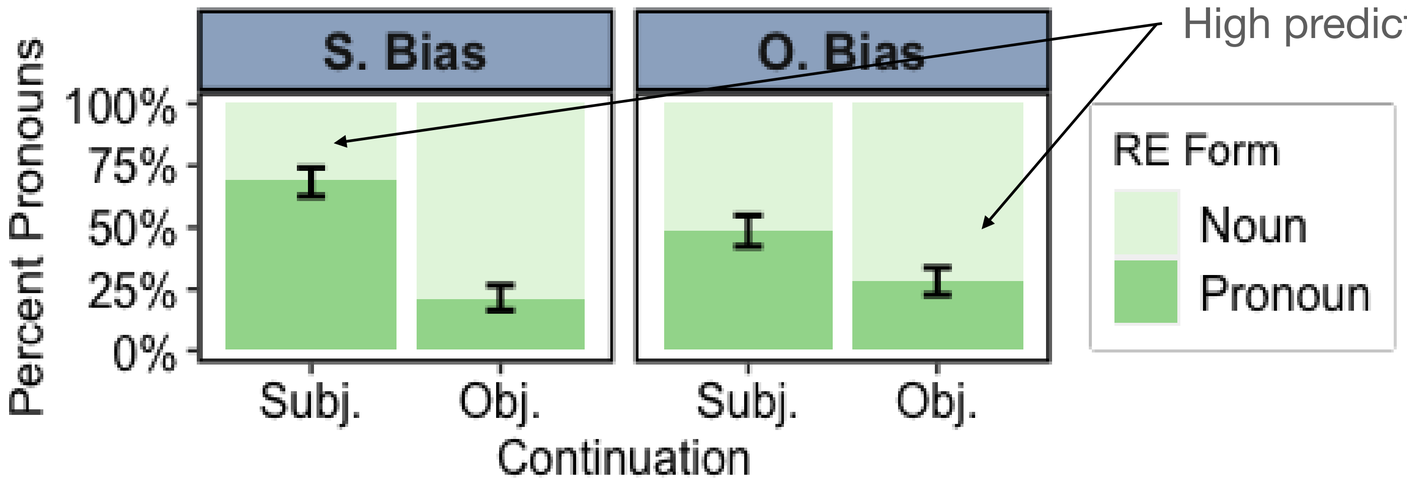
The psychic had been poisoned by venom bleeding ivy. The healer rushed over with an assortment of antidotes.

The healer sold a vial of antidote to the psychic./The psychic bought a vial of antidote from the healer.

Next, ...

Demberg, Kravtchenko and Loy, 2022

Paradigm with contextualized stimuli



High predictability conditions

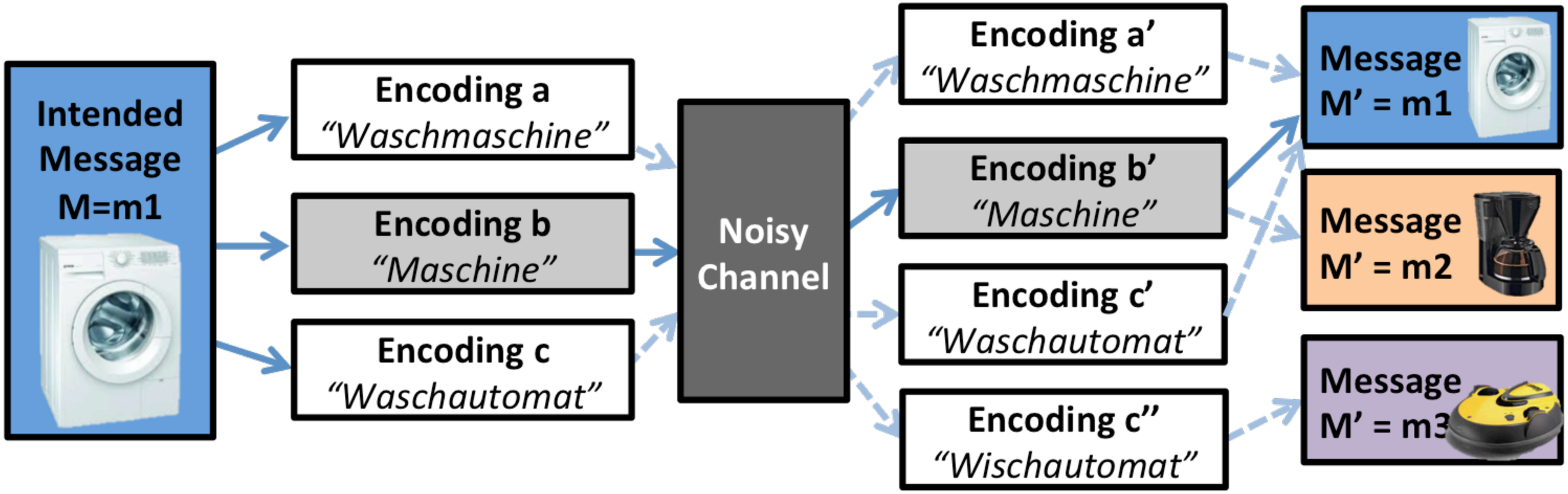
- Contextualization <u>does</u> make a difference:
  in this setting, the pronominalization rate of predictable referents
  is larger than the pronominalization rate of unpredictable ones ($p<0.001$)
  => naturalistic tasks should be taken more seriously in experimental design.

- Nevertheless, the largest effect on pronoun choice is whether the pronoun refers to the subject or the object of the previous sentence.

Demberg, Kravtchenko and Loy, 2023

Speakers select encodings rationally.

Listeners rationally combine bottom-up and top-down evidence



Q3 — Wie können wir die nötigen Wahrscheinlichkeiten abschätzen?

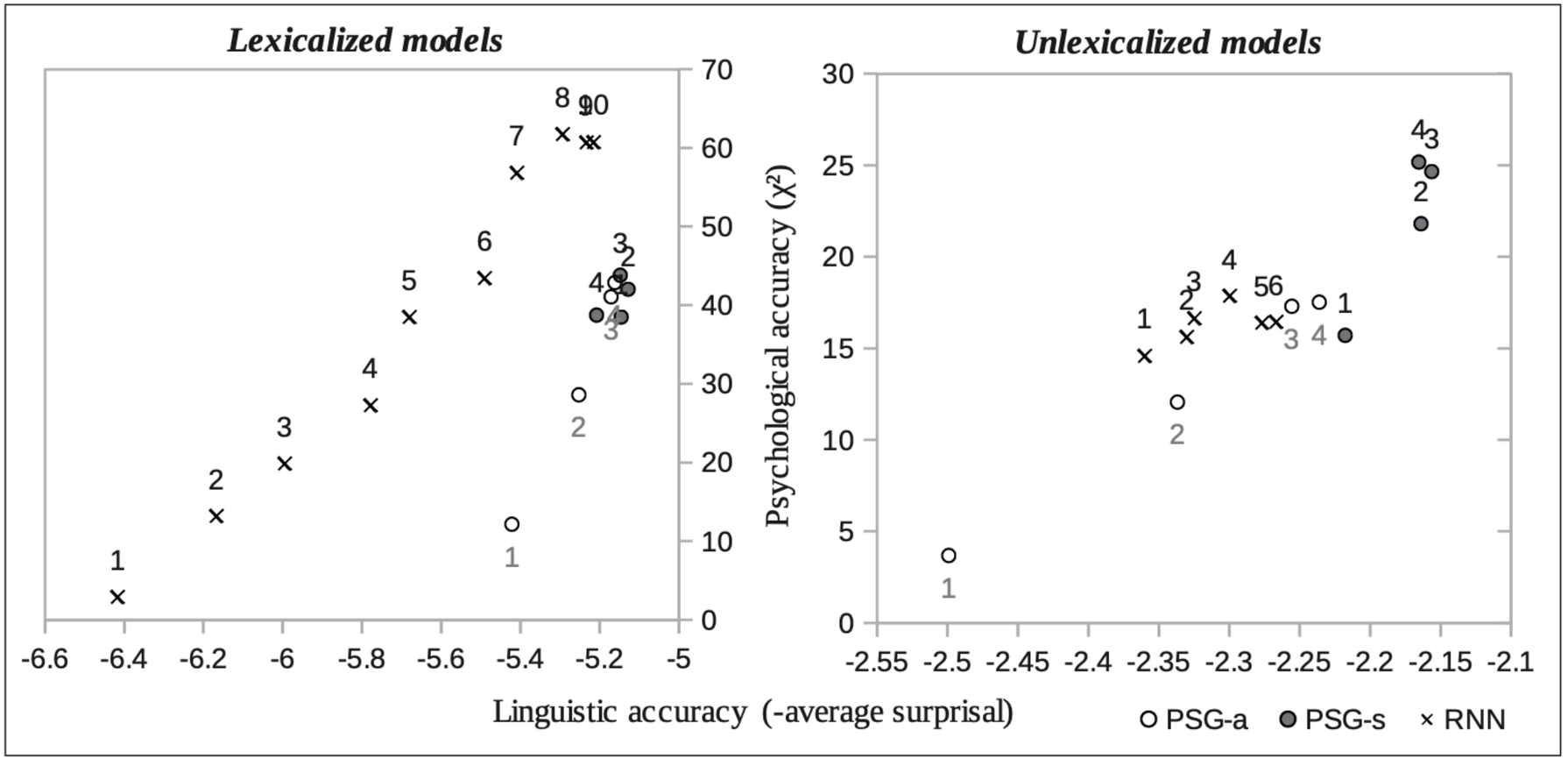
Estimating surprisal

Extensive literature on how to estimate surprisal:

- from humans using cloze probabilities
- from language models
    - n-gram models
    - syntactic parsers
    - neural models (feedforward)
    - newest large language models
- recent studies find that estimation from language models may be more accurate than estimation from cloze probabilities (Michaelov et al., 2023).

Findings 2008 – 2020

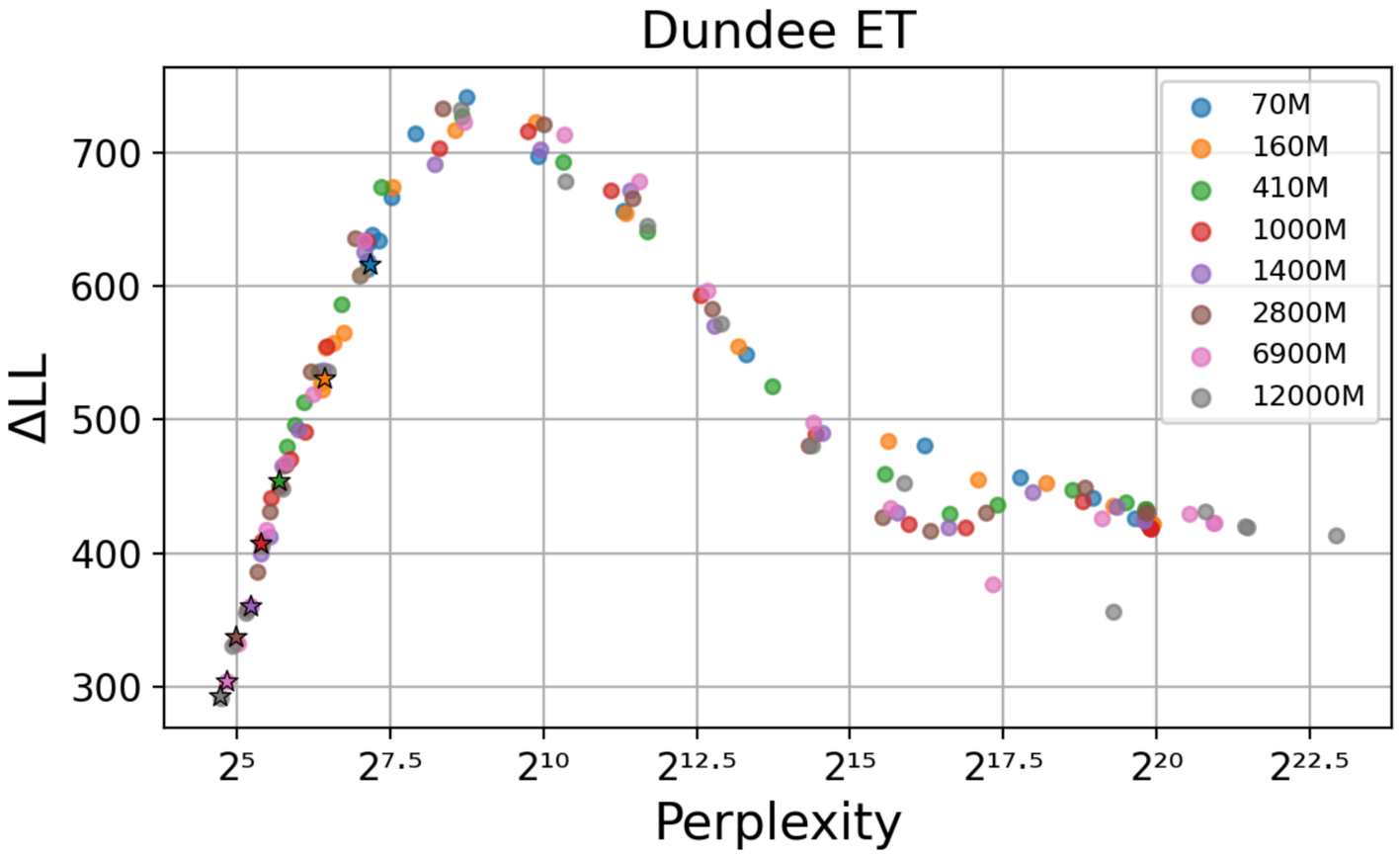
Better language models are typically also better at predicting psychometric measures.



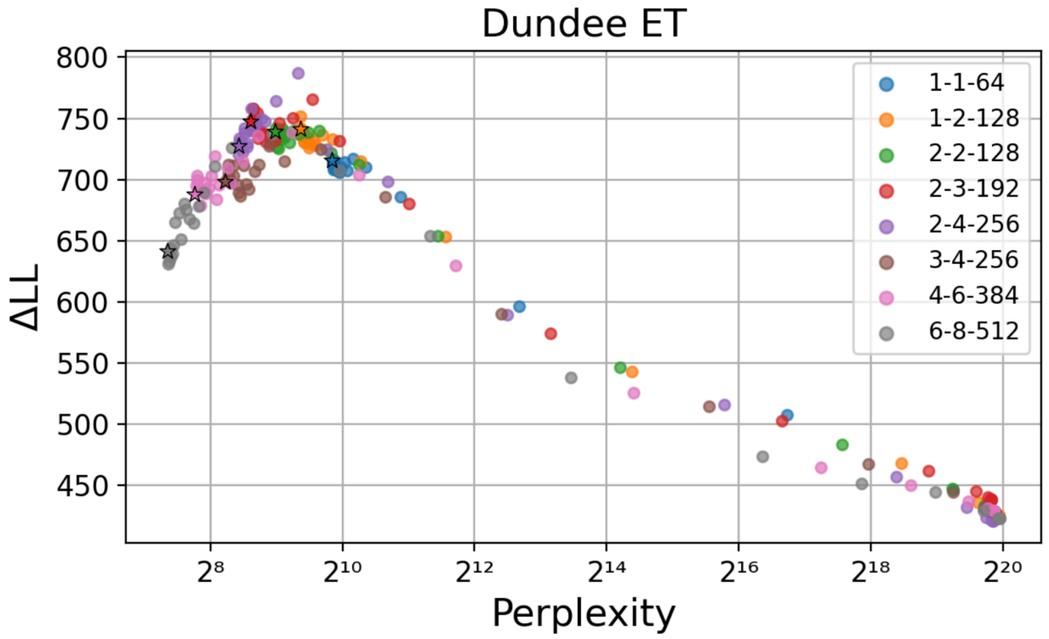
Graph from Monsalve et al., 2012

Recent Findings on LLMs

Better language models are NOT necessarily better at predicting psychometric measures — there seems to be a sweet spot (Kuribayashi et al, 2021; Oh and Schuler, 2023).



- number of parameters;
- stars indicate fully trained models;
- 70M model from left plot equals 6-8-512 model from right plot.



What does this mean?

Transformers work differently from humans in language processing

- they may miss some world-knowledge related predictability effects
- they are better at storing prior context verbatim
- have read A LOT more texts than humans can ever read in their lifetime, which may lead to underestimates of reading times on e.g., named entities (Oh and Schuler, 2023)
- or, maybe, our assumptions of a linear relationship between surprisal and human reading times are wrong.

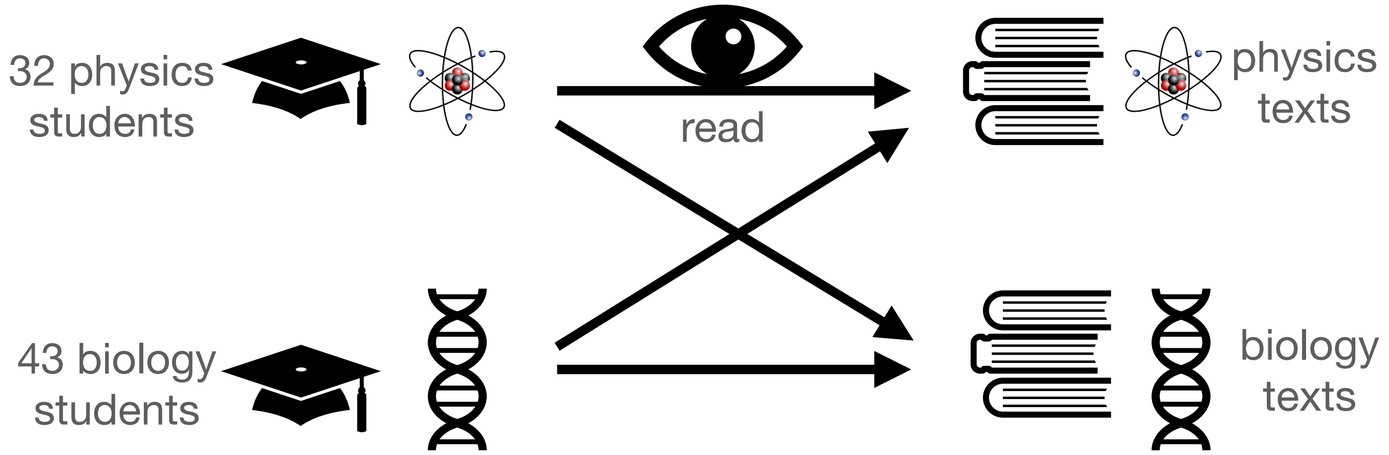
- We should rather estimate models that fit humans / human experience better.
  - To what extent does the language model match the readers and their background knowledge? → **Unexplored as of yet**

Reading data: POTEC corpus

POTEC corpus (Jäger et al., 2021): Scientific text with technical terms

Um das Vorhandensein der Polymerasekettenreaktion-Produkte feststellen zu können, verwendet man die Gelelektrophorese mit einer anschließenden Behandlung des Gels durch Ethidiumbromid. …

Eye-tracking dataset: POTEC



- Terminology annotation: common and technical terms

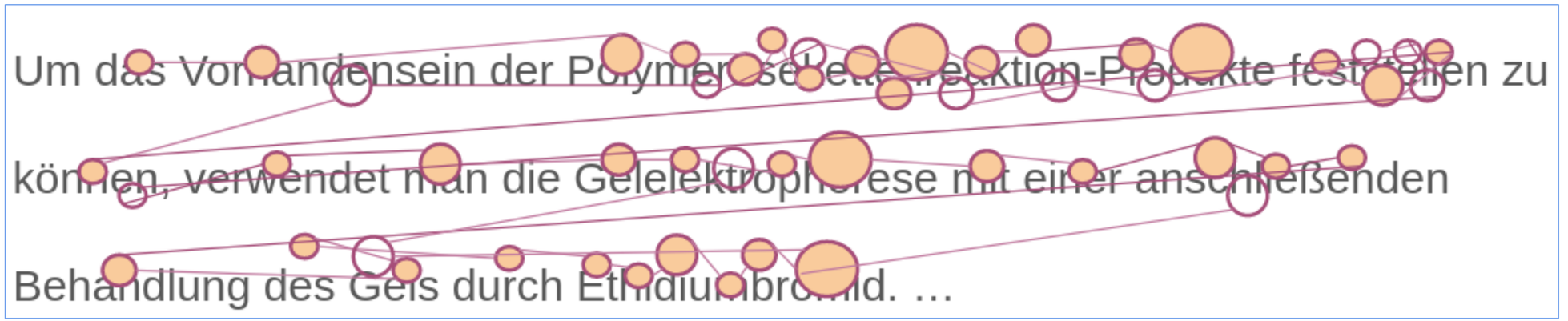
  Um das Vorhandensein der Polymerasekettenreaktion-Produkte feststellen zu können, verwendet man die Gelelektrophorese

Skrjanec, Broy and Demberg, 2023

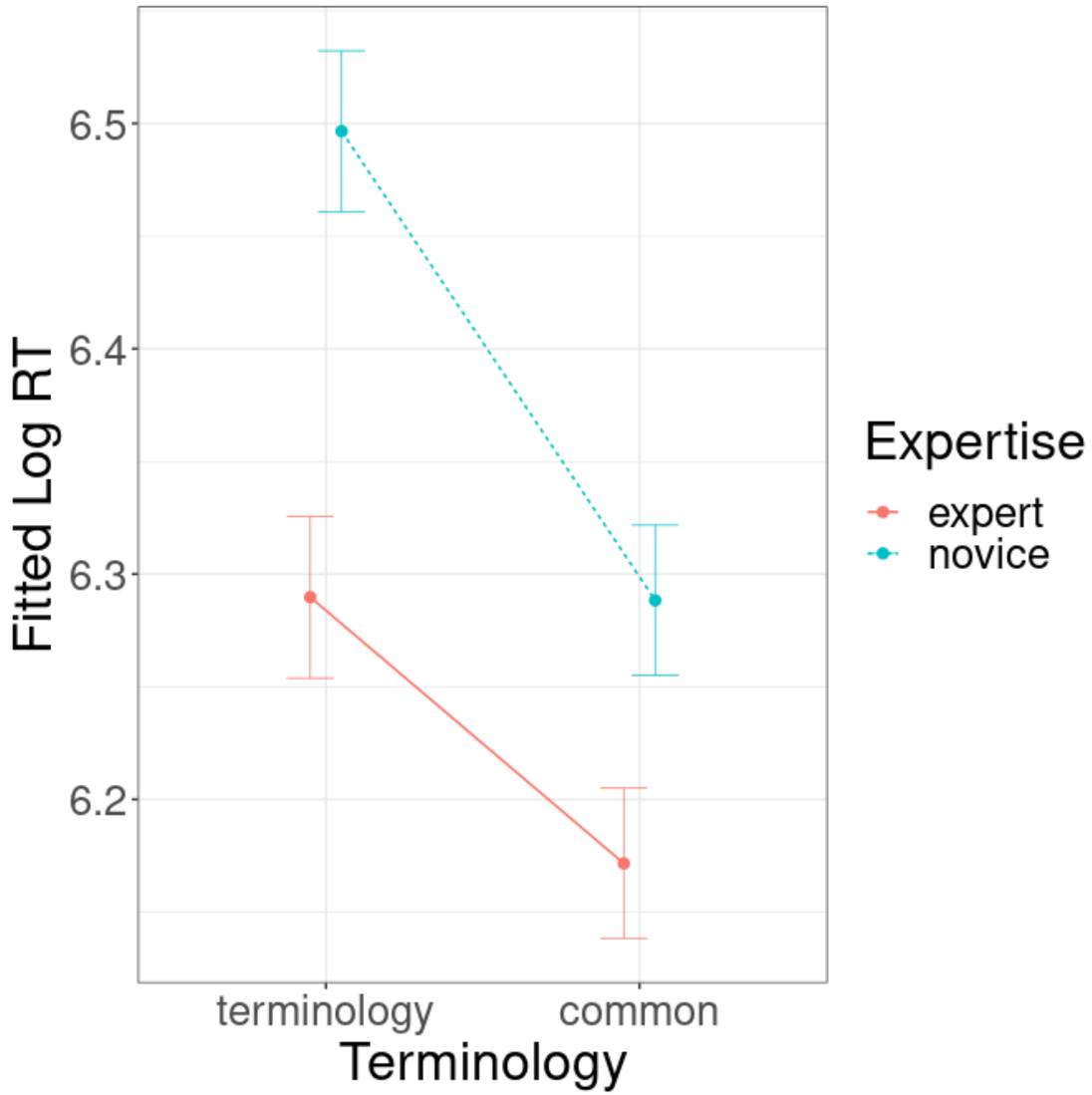
Same text, different readers

Expert in biology

Novice in biology



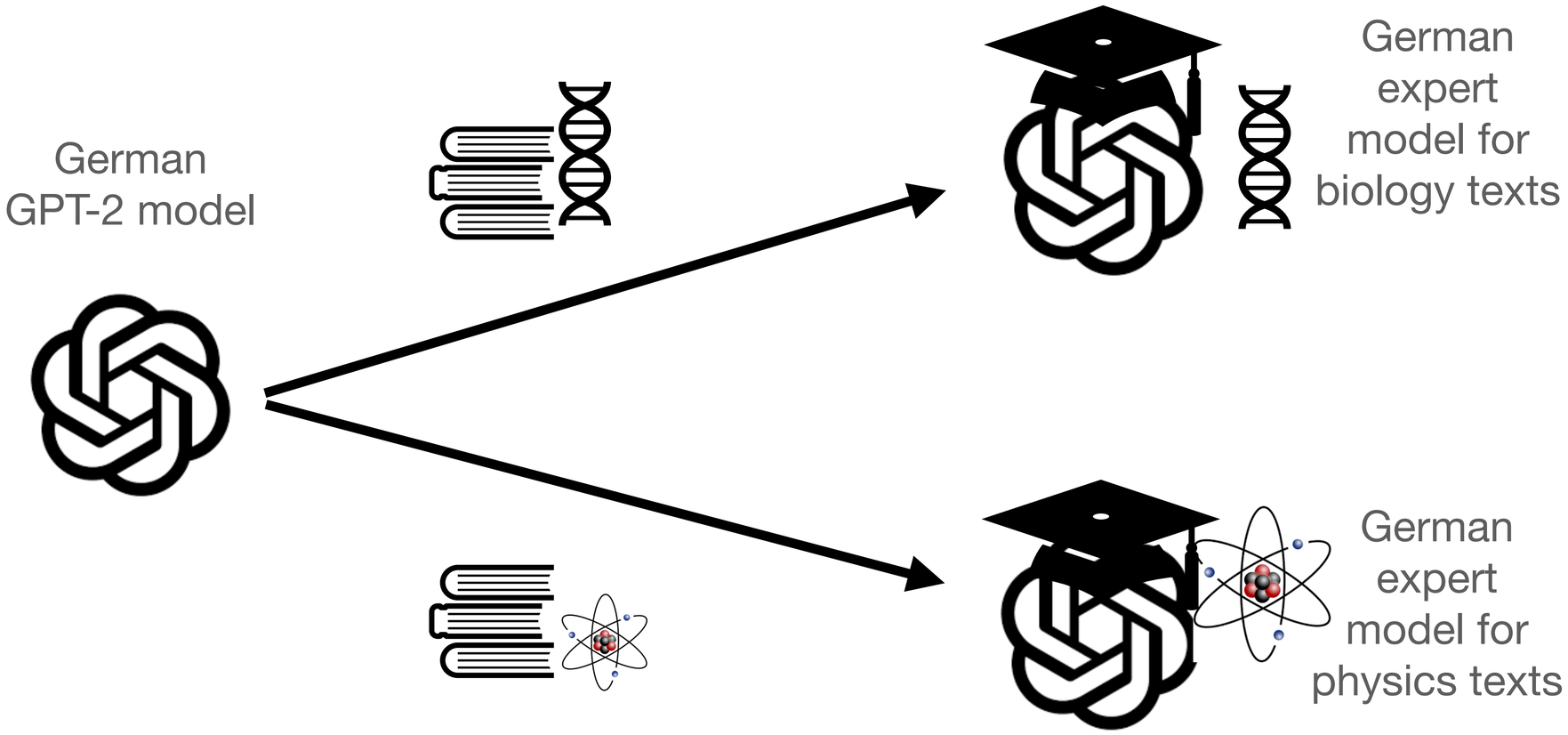
Effect of expertise



- Experts read faster than novices
- Terminology is read more slowly than common words
- There is a significant interaction such that novices have less of a terminology “penalty”.

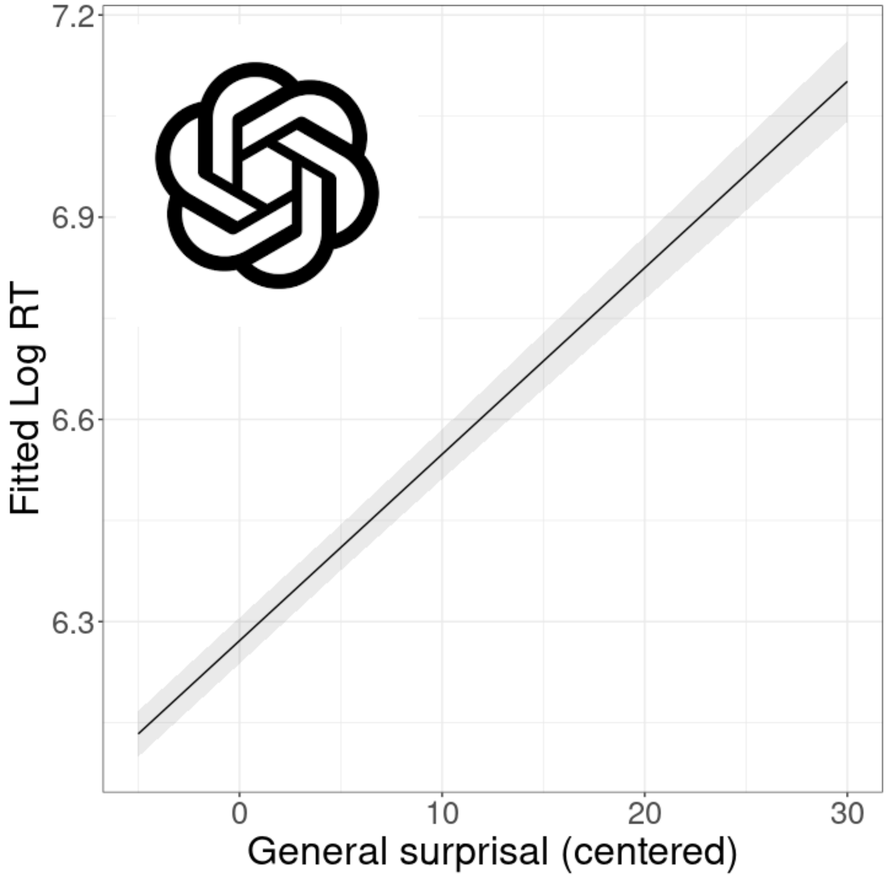
Skrjanec, Broy and Demberg, 2023

General & expert models



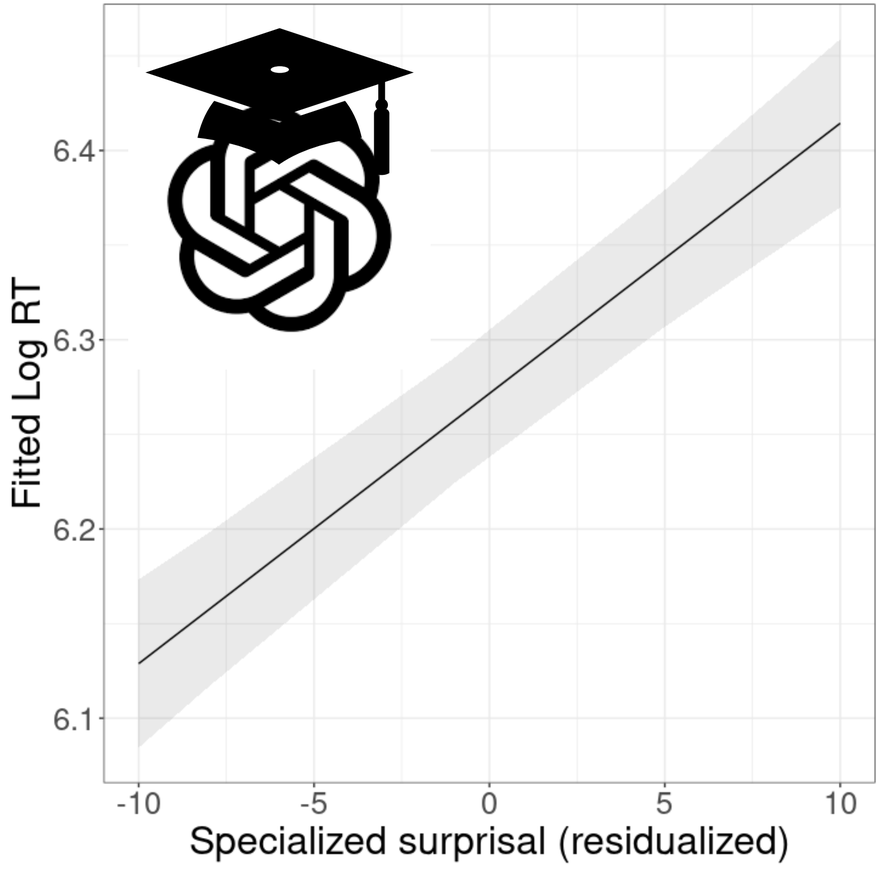
Skrjanec, Broy and Demberg, 2023

Performance of surprisal models



Surprisal estimates from a LM are predictive of reading times.

An “expert” model further trained on in-domain texts predicts expert reading times more accruately than a non-expert model.



Skrjanec, Broy and Demberg, 2023

Zusammenfassung

- Noisy channel model besagt, dass Hörer bottom-up Information aus Signal mit top-down Kontexterwartungen rational kombinieren.
- Diese Vorhersagen werden durch experimentelle Daten bestätigt.
- Sprecher richten Wortwahl nach kontextueller Erwartbarkeit aus.
  => Sprachwandel
- Sprachmodelle zur Abschätzung der Wahrscheinlichkeiten sollten mit Bedacht ausgewählt werden; evtl. an den Leser angepasst werden.

vielen Dank für Ihre Aufmerksamkeit!

vielen Dank an die beteiligten Mitarbeiter / Koautoren:
Marjolein van Os, Jia Loy, Katja Kravtchenko, Iza Skrjanec, Anupama Chingacham, Alessandra Zarcone.

UNIVERSITÄT
DES
SAARLANDES

IDeaL
SFB 1102

erc

European Research Council
Established by the European Commission