Trade-offs between information and speech rate in naturalistic speech from 49 non-WEIRD languages

Christophe Coupé, Yoon Mi Oh, Dan Dediu, Frank Seifart & François Pellegrino

57th Annual Meeting of the Societas Linguistica Europaea

21/08/2024

Overview

I. Introduction: Information regulation in language

II. Methodology: Corpus, variables and models

III. Results

IV. Discussion

Overview

I. Introduction: Information regulation in language

II. Methodology: Corpus, variables and models

III. Results

IV. Discussion

Information regulation in language

Information transmission as a landmark of all communication systems

- (human) language among others

Information is very likely not distributed at random throughout linguistic communication for speaker and/or listener's sake

- Least-effort principle

Information can be distributed 'strategically' over linguistic units

➔ Information regulation should be visible as statistical trade-offs

Informational trade-offs

Within a language and across speakers / texts

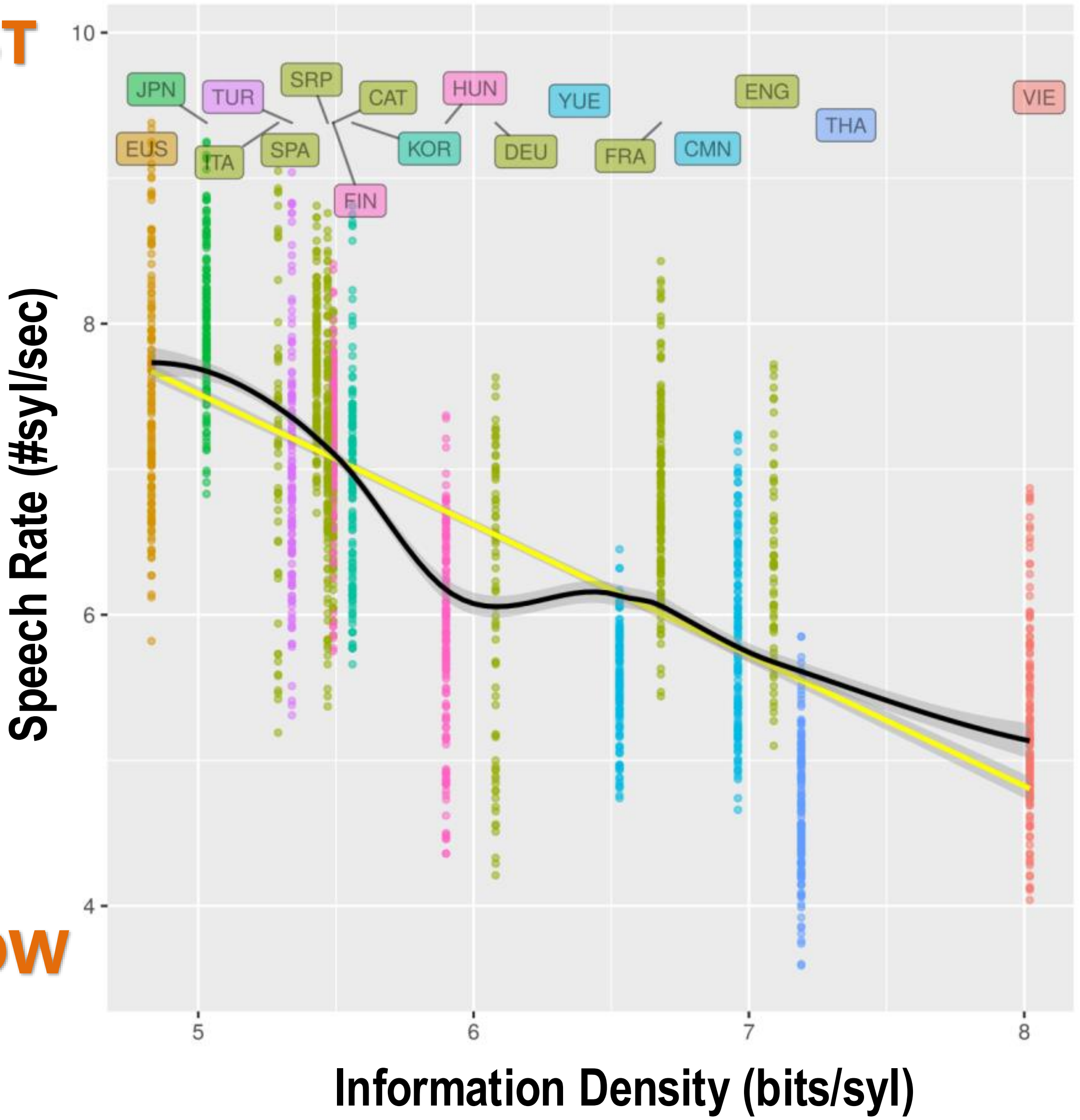
- Information borne by phones and their acoustic durations are related (Pimentel et al., 2021) [600 languages]
- Regulation between syntactic complexity, lexical information, and speech rate in English (Cohen Priva, 2017)
- Uniform Information Density Hypothesis towards a language level attractor (Meister et al., 2021)

Across languages

- Potential attractor in terms of average information rate at the syllabic level (Coupé et al., 2019; Pellegrino et al., 2011) [17 languages]

Cohen Priva, U. (2017). Not so fast : Fast speech correlates with lower lexical and structural information. *Cognition*
Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density Hypothesis. *Proc. EMNLP 2021*
Pimentel, T., Meister, C., Salesky, E., Teufel, S., Blasi, D., & Cotterell, R. (2021). A surprisal—Duration trade-off across and within the world's languages. *Proc. EMNLP 2021*

Density vs Speed trade-off



FAST

SLOW

LIGHT

DENSE

Coupé, C., Oh, Y. M., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*
Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*

Current limitations in most studies

Proposals often based on read speech produced by few speakers and/or data from only a few “big” languages

- But considering sociolinguistic and typological diversity is essential

Primary sources mostly *textual* and polished by several authors, often without any temporal information

- But *speech* is the backbone of human language and is by nature, temporal

Current limitations in most studies

Regulation mostly investigated at phonological (phone, syllable) and word level (word surprisal and syntactic structure)

- But morphology should also be considered, as it is the gateway to semantics

Cross-language comparability is ensured through the use of parallel corpora (similar semantic content shared across languages)

- Strong limitation and incompatibility with the concept of natural speech…

Do the previous hypotheses ‘survive’ in natural speech and also generalize to non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic; see Henrich et al. 2010) languages?

→ This study…

- is based on a large dataset of natural speech in (mostly) non-WEIRD languages
- proposes a methodology adequate to deal with non-parallel datasets
- explores the existence of a trade-off between information density and speech rate at both syllabic and morphological levels

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences

Overview

I. Introduction: Information regulation in language

II. Methodology: Corpus, variables and models

III. Results

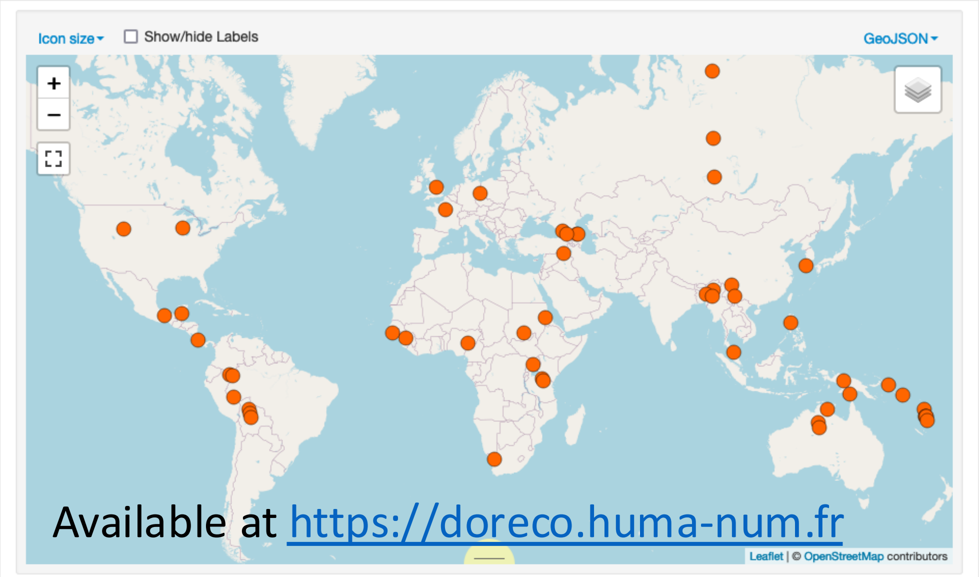
IV. Discussion

Data source: DoReCo

Language Documentation Reference Corpora

Natural speech in 53 languages

- 6 continents (31 families + 3 isolates)
- Non-WEIRD and mostly endangered languages (+ English and French)
- Transcribed, translated, and authored by experts in the languages
- Interlinear glosses provided for most languages



Available at https://doreco.huma-num.fr

Time alignment of the transcription at the phone, morph and word levels

All corpora ca. ~10,000 words

Primarily monological – personal narratives, traditional narratives and descriptions

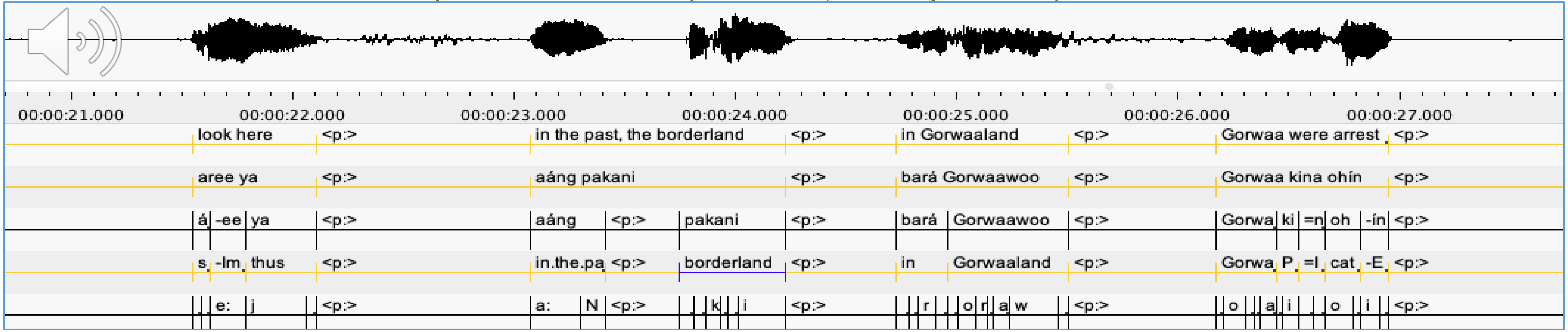
v1.2 (2022): 50 lgs; v1.3 (forthcoming): 53 lgs

Seifart, F., Paschen, L., & Stave, M. (Eds.). (2022). Language Documentation Reference Corpus (DoReCo) 1.2. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Univ. Lyon 2). https://doi.org/10.34847/nkl.7cbfq779

Seifart, F., Paschen L., & Stave, M. (Eds.). (Forthc.) Language Documentation Reference Corpus (DoReCo) 1.3. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Univ. Lyon 2).

Example informativeness vs. speed trade-off

informative and **slow** speech in Gorwaa (Tanzania; Harvey Forthc.):



look here <p:> in the past, the borderland <p:> in Gorwaaland <p:> Gorwaa were arrest <p:>

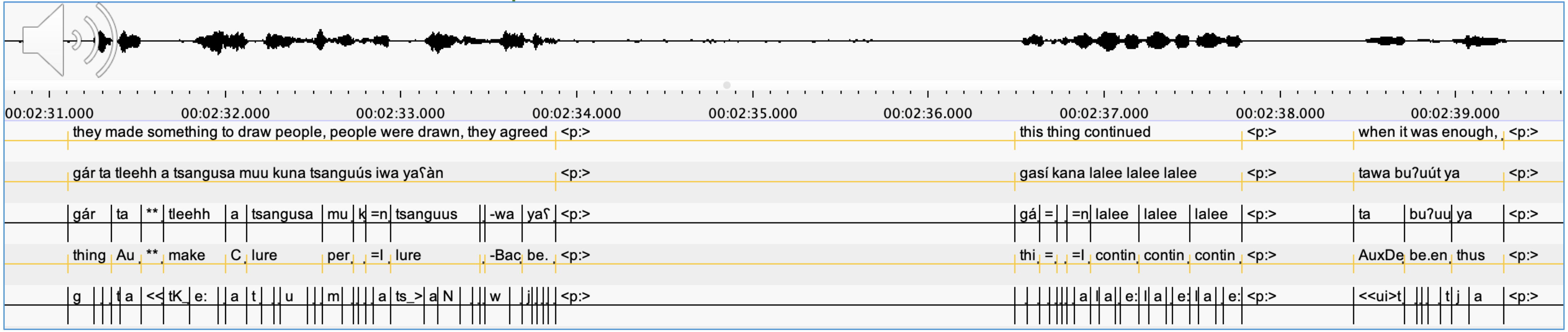
aree ya <p:> aáng pakani <p:> bará Gorwaawoo <p:> Gorwaa kina ohín <p:>

á -ee ya <p:> aáng <p:> pakani <p:> bará Gorwaawoo <p:> Gorwa ki =n oh -ín <p:>

s -Im thus <p:> in.the.pa <p:> borderland <p:> in Gorwaaland <p:> Gorwa P =I cat -E <p:>

e: j <p:> a: N <p:> k i <p:> r o r a w <p:> o a i o i <p:>

vs. **uninformative** and **fast** speech in Gorwaa



they made something to draw people, people were drawn, they agreed <p:> this thing continued <p:> when it was enough, <p:>

gár ta tleehh a tsangusa muu kuna tsanguús iwa yaʕàn <p:> gasí kana lalee lalee lalee <p:> tawa buʔuút ya <p:>

gár ta ** tleehh a tsangusa mu k =n tsanguus -wa yaʕ <p:> gá = =n lalee lalee lalee <p:> ta buʔuu ya <p:>

thing Au ** make C lure per =I lure -Bac be. <p:> thi = =I contin contin contin <p:> AuxDe be.en thus <p:>

g t a << tK_ e: a t u m a ts_> a N w j <p:> a l a e: l a e: l a e: <p:> <<ui>t t j a <p:>

The dataset of this study

Subset of 36 languages

- with interlinear glosses (for the analysis at the morphological level)
- with English translations (some languages only have Spanish or French translations)

apah1238 (Yali); bain1259 (Gubeeher); beja1238 (Beja); bora1263 (Bora); cabe1245 (Cabecar); cash1254 (Cashinahua); dolg1241 (Dolgan); even1259 (Evenki); goem1240 (Goemai); goro1270 (Gorwaa); hoch1243 (Hoocak); jeha1242 (Jahai); jeju1234 (Jejuan); kaka1265 (Kakabe); kama1351 (Kamas); movi1243 (Movima); ngal1292 (Dalabon); nngg1234 (Nuu); nort2641 (Northern Kurdish); nort2875 (Northern Alta); orko1234 (Fanbyak); pnar1238 (Pnar); port1286 (Daakie); ruul1235 (Ruuli); sanz1248 (Sanzhi); savo1255 (Savosavo); sout2856 (Nafsan); sout3282 (English); sumi1235 (Sumi); taba1259 (Tabasaran); teop1238 (Teop); texi1237 (Popoluca); toto1304 (Totoli); trin1278 (Mojeno); urum1249 (Urum); vera1241 (Veraa).

Dataset: Curation & Preparation

Pauses removed

- Speech rate in terms of articulation rate

Units of analysis: sections of ~10 seconds

- Interested in average rates and not local variation
- Shorter annotation units concatenated when possible
- Longer annotation units kept unchanged
- Short residual sections (< 5 seconds) removed

Dataset: Main figures

Number of languages: 36

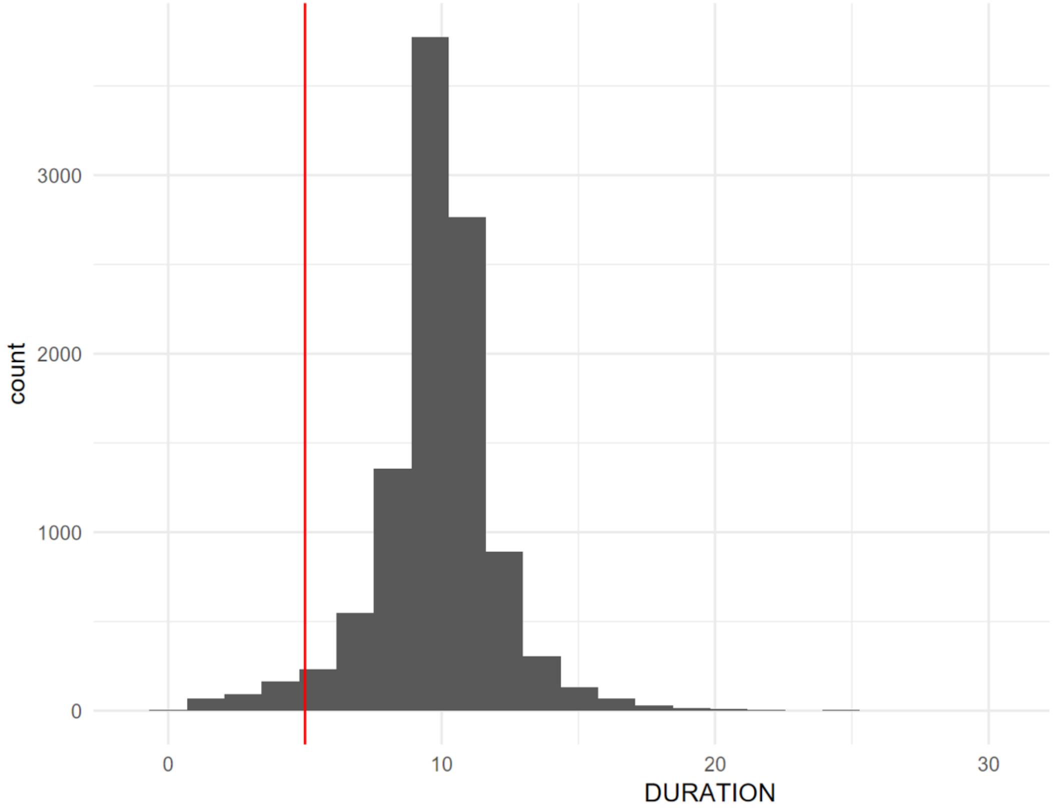
Total duration: 27.05 hours

Total number of speakers: 273

Total number of sections: 9,668

Speech type:

- Monological narratives: 84%
- Conversation: 9%
- Stimulus retelling: 7%



Main variables of interest

(all defined at the level of sections)

Duration(DURATION): #seconds

Syllabic rate (SR): #syllables/second

Morph rate (MR) : #morphs/second

Amount of information (SURPRISAL)

- Approximated by the Shannonian surprisal of the English translations, estimated with a GPT-2 model

N.B. All numerical variables were standardized without subtraction of the mean

Information density

- Adapted from Pellegrino, Marsico & Coupé (2011)
- **Unitless ratio** with English as the reference
- Based on the respective number of "units" in the source language and its English translation
- **Syllabic density** (SD)

  = Syllable Count in English * / Syllable Count in Source Language
- **Morphological density** (MD)

  = Morph Count in English* / Morph Count in Source Language

** Estimated with a SpaCy analyzer*

Pellegrino, F., Marsico, E., & Coupé, C. (2011) A cross-language perspective on speech information rate. *Language*

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density Hypothesis. *Proc. EMNLP 2021*

Wilcox, E.G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *Proc. 42nd CogSci*

Additional variables

Speaker sex (SPK_SEX)

Speaker age (SPK_AGE)

Type of speech (conversation, monological narrative, stimulus retelling) (TYPE)

Language area, as extracted from Glottocode (AREA)

Language (LG_CODE)

Speaker (SPK)

File (FILE)

+ (not commented here):

SR, MR, and Surprisal of the previous section

Cumulative duration since the beginning of the production

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2024). *Glottolog 5.0*

Methodology

1. **Estimate statistical models (mixed-effects linear regressions) to assess the key impact of syllabic/morphological density on syllabic/morphological rate**
   - Do we confirm the existence of a trade-off in natural speech even when controlling for other factors?
   - First, with main effects only

2. **Investigate other interesting effects, in particular the effect of surprisal on syllabic/morphological rate**

3. **Consider interactions to get a more detailed understanding of speech production**

4. **Shift from regression models to causal models**
   - Path analysis

Overview

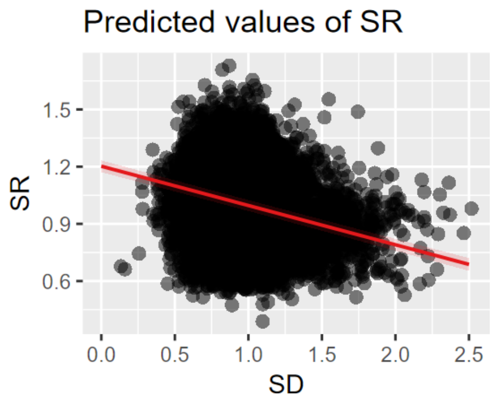
I. Introduction: Information regulation in language

II. Methodology: Corpus, variables and models

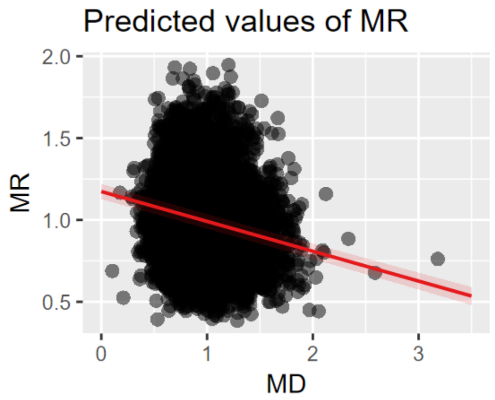
III. Results

IV. Discussion

Main effects without interactions: Existence of a trade-off between density and speech



Predicted values of SR

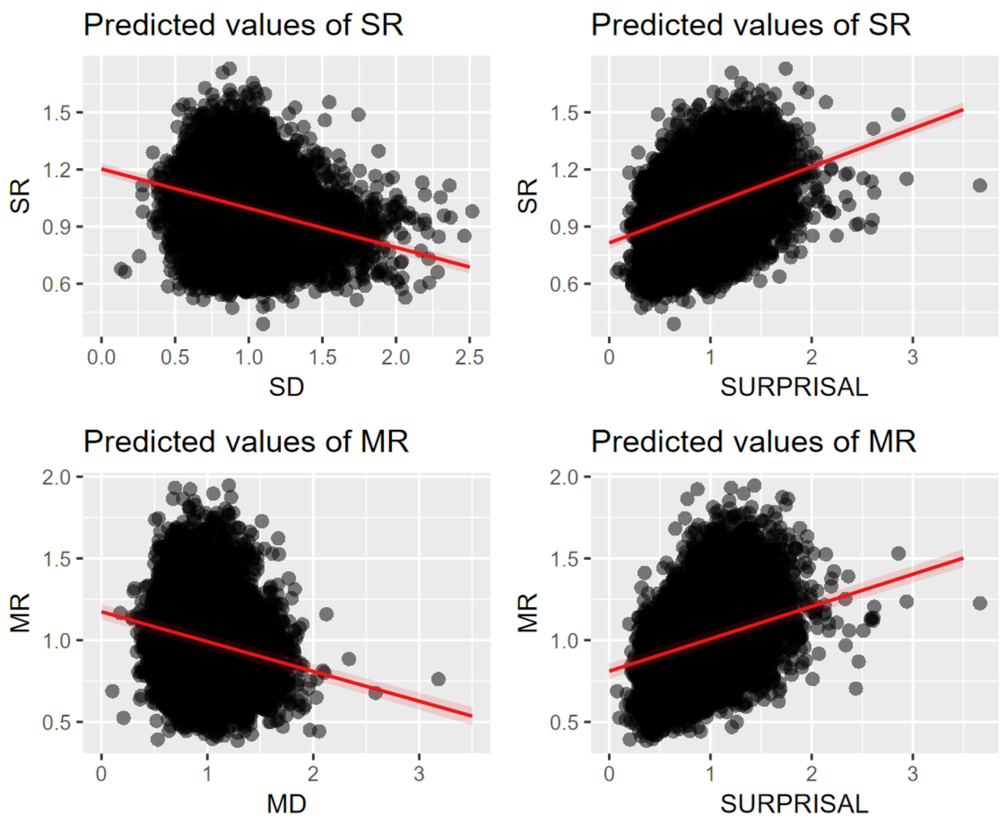


Predicted values of MR

(Figures report *estimated marginal trends / means*)

❶ The density-speed trade-off is clear at both syllabic and morphological levels (the higher the density, the slower the speech)

Main effects without interactions: Existence of a trade-off between density and speech

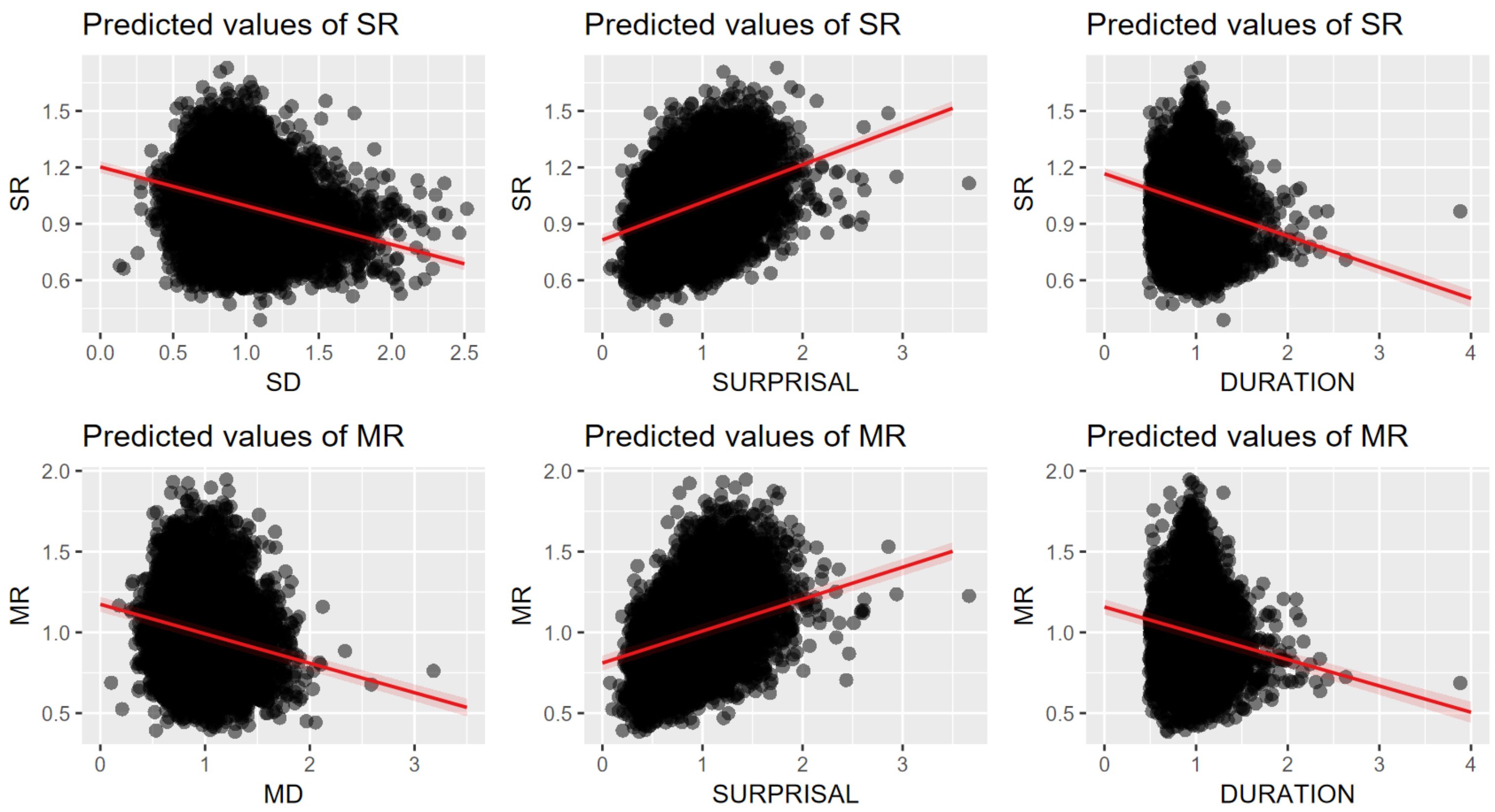


(Figures report *estimated marginal trends / means*)

➊ The density-speed trade-off is clear at both syllabic and morphological levels (the higher the density, the slower the speech)

➋ SURPRISAL positively impacts both SR and MR (the more informative the section, the faster the speech)

Main effects without interactions: Existence of a trade-off between density and speech



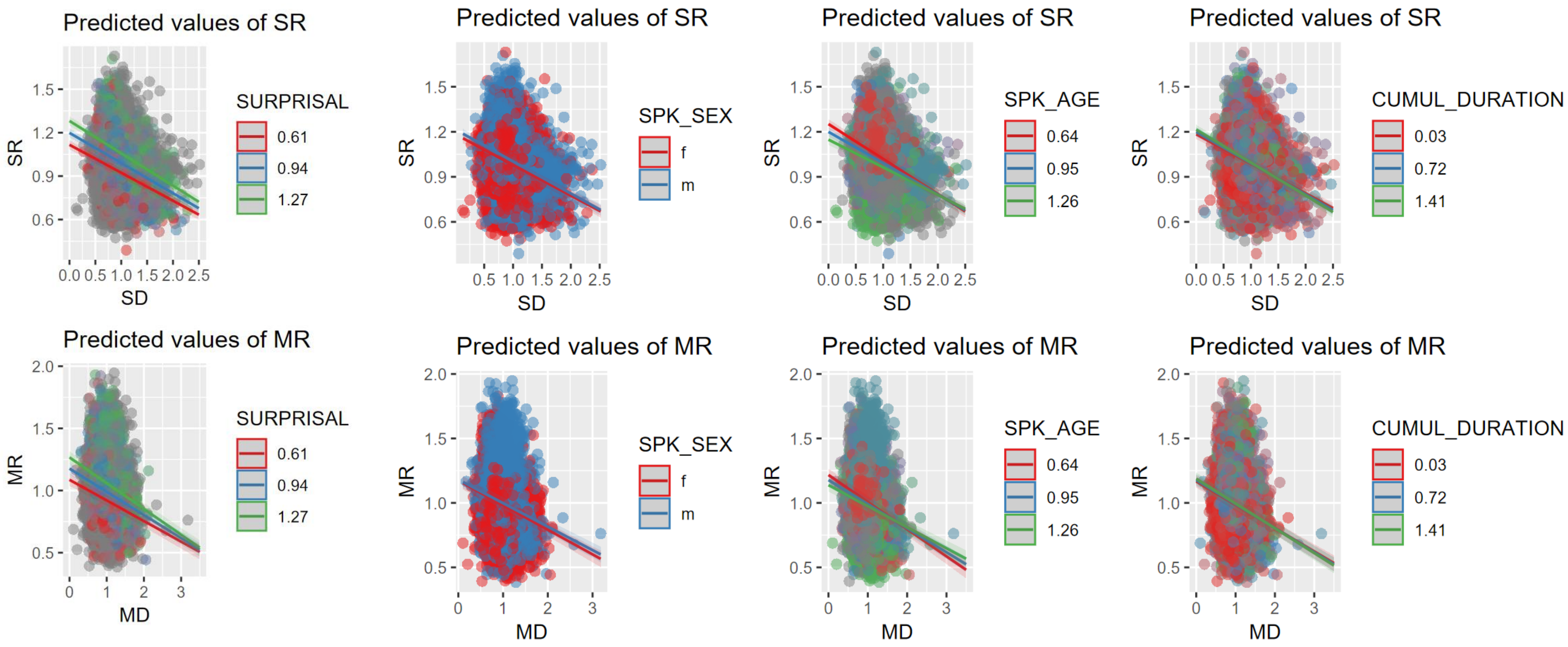
(Figures report *estimated marginal trends / means*)

❶ The density-speed trade-off is clear at both syllabic and morphological levels (the higher the density, the slower the speech)

❷ SURPRISAL positively impacts both SR and MR (the more informative the section, the faster the speech)

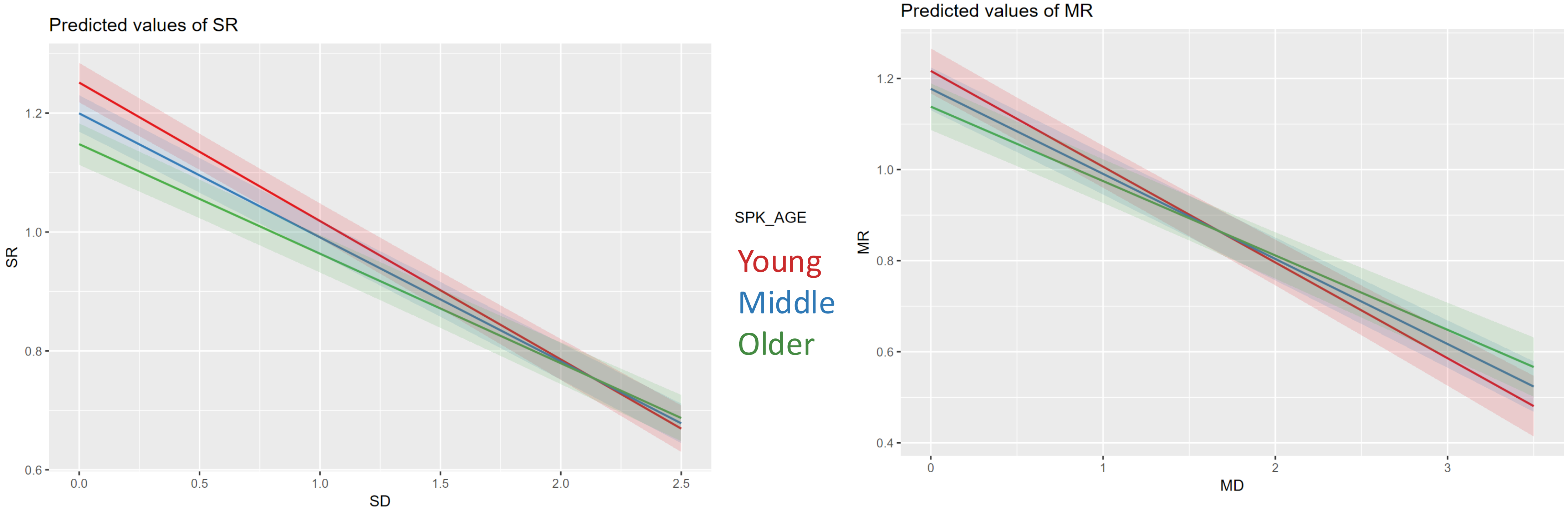
❸ DURATION negatively impacts both SR and MR (the longer the section, the slower the speech)

Accounting for interactions: density-rate trade-off



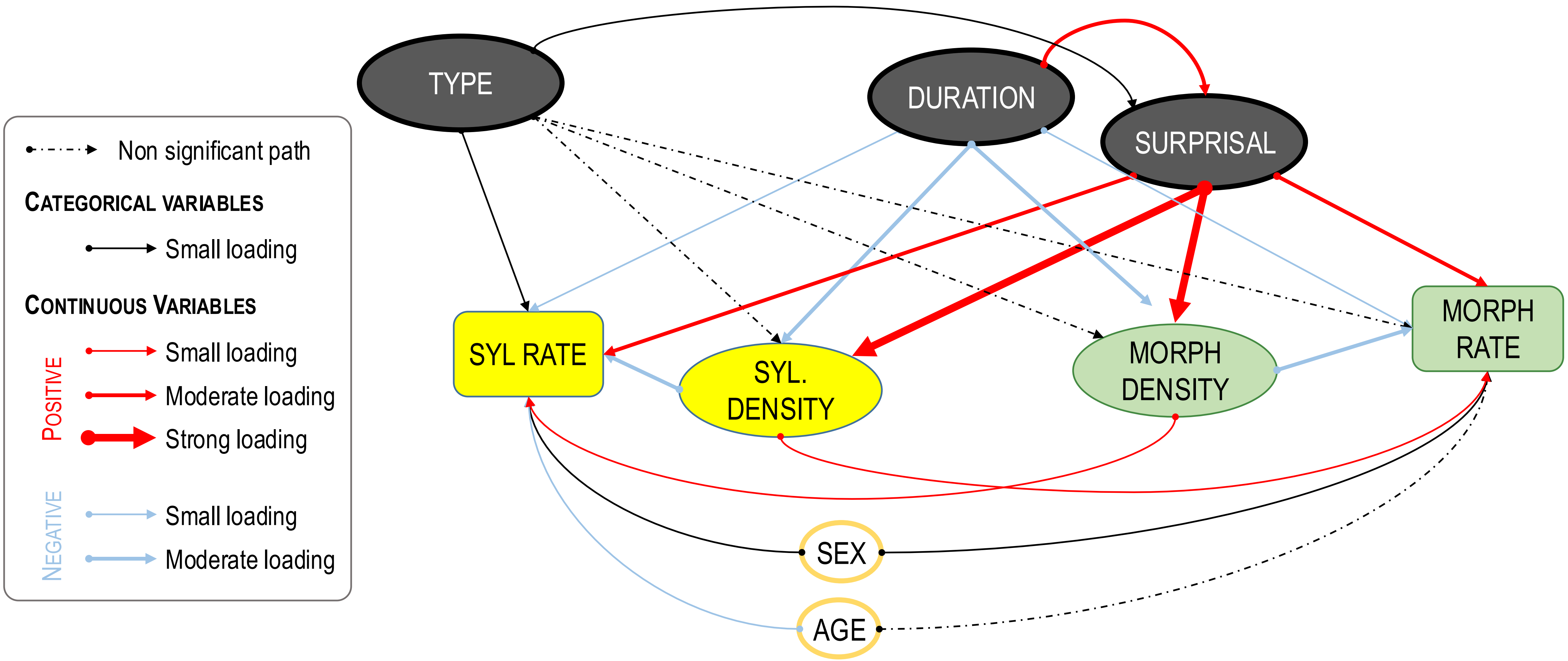
The density-rate trade-off exists for males and females, and regardless of age, surprisal or cumulative duration

Accounting for interactions: AGE



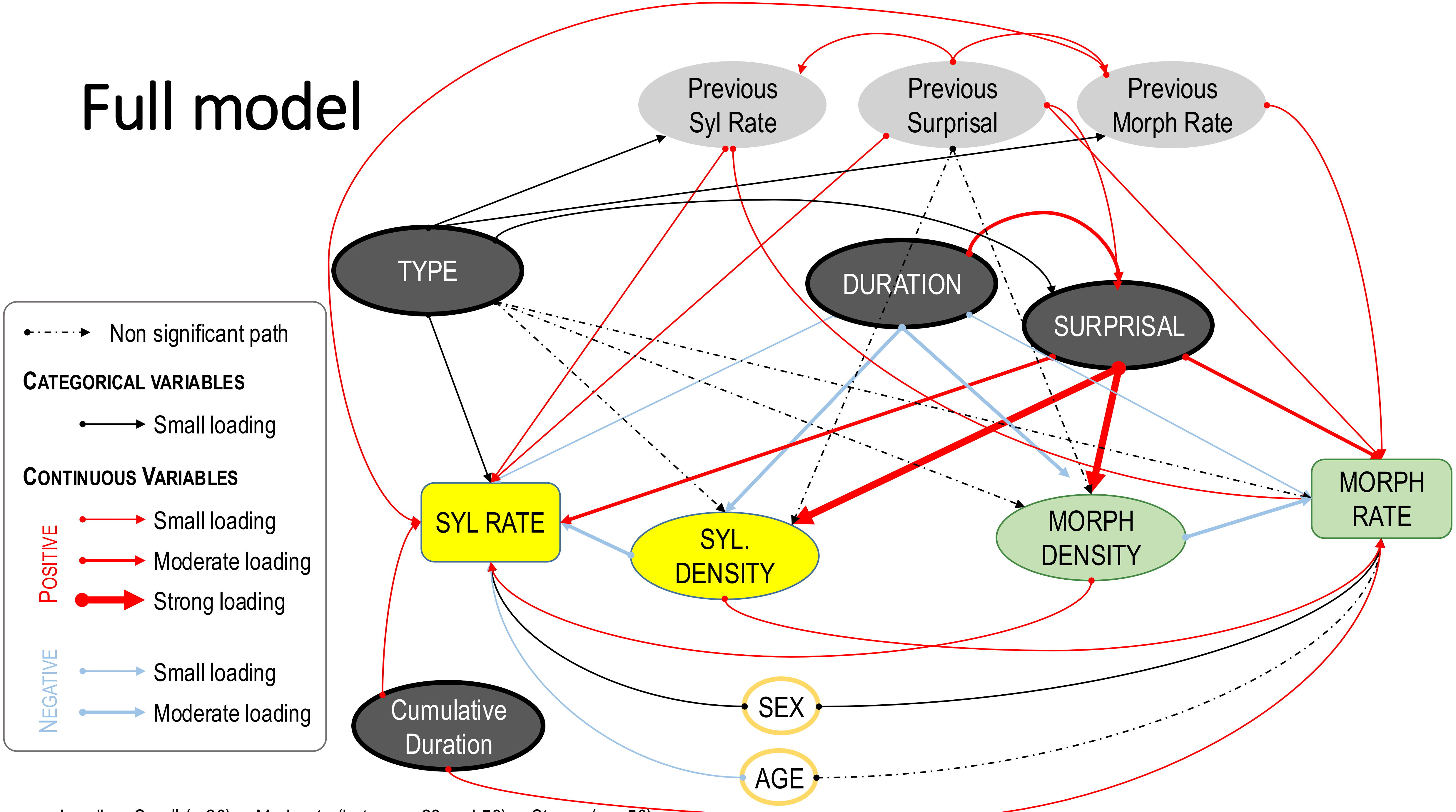
The younger the speakers, the stronger the modulation between density and speed (steeper slope)

Path Modeling



Loading: Small (<.20) ● Moderate (between .20 and.50) ● Strong (>= .50)

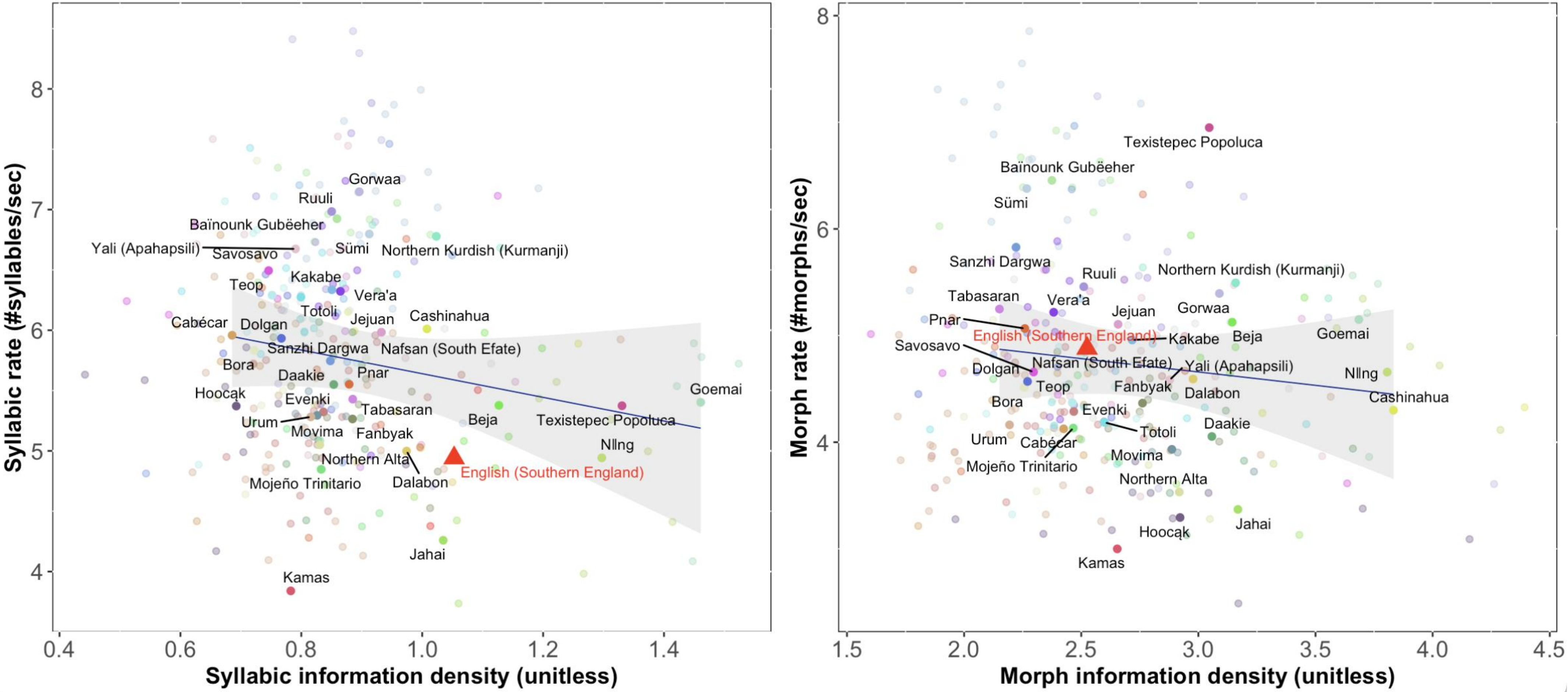
Full model



Loading: Small (<.20) ● Moderate (between .20 and.50) ● Strong (>= .50)

Where does English stand in this picture?

Transparent points: Individual speakers; Blue line: regression line for the 36 languages



At the syllabic level, English stands outside the central cluster of languages, without being an outlier.
At the morphological level, English is quite unremarkable.

Overview

I. Introduction: Information regulation in language

II. Methodology: Corpus, variables and models

III. Results

IV. Discussion

Discussion

A robust trade-off between density and speed in natural speech across 36 languages

- (Not shown here: also found at syllabic level for 50 languages)

This trade-off exists at both syllabic (encoding) and morphological (semantic) levels

Expected sociolinguistic effects are found (in most situations, males speak faster than females and older speakers speak slower than younger speaker)

Beyond, the speed/density modulation seems to attenuate with aging

- Speculatively, older speakers may converge on slower and more uniform speech rates, regardless of their language

Limits and ongoing/future work

Further exploration needed to understand the causal relationship between the syllabic and morphological levels

- According to path models, MD as a small influence of SR and vice-versa with SD and MR
- Could/should be done in light of languages' typological profile

Several limits given how information is estimated and modelling is done

- But the exact same methodology is applied to all languages in the dataset
- Non-linear phenomena can be modelled

Regulations at the (local) level of individual syllables and morphemes should be investigated, as well as pauses and tones. Other indices of informational density / amount can be considered.

Conclusions

A strong trade-off regulating the flow of information can be found across typologically and sociolinguistically diverse languages

The time has come to study natural speech in non-WEIRD languages

- Thanks to corpora like DoReCo

This study breaks the twofold barrier of read speech and parallel corpus and paves the way for studying the interaction between cognition and both typology and sociolinguistics

Thank you for your attention

DoReCo corpus references (1/4)

All contributions below to appear in:

F. Seifart, L. Paschen, & M. Stave (Eds.), *Language Documentation Reference Corpus (DoReCo) 1.3. Laboratoire* Dynamique Du Langage (UMR5596, CNRS & Univ. Lyon 2).

Bardají, M., Bracks, C., Leto, C., Hasan, D., Riesberg, S., Alamudi, W. S., & Himmelmann, N. P. (Forthc.). **Totoli** DoReCo dataset.

Bartels, H., & Szczepański, M. (Forthc.). **Lower Sorbian** DoReCo dataset.

Bogomolova, N., Ganenkov, D., & Schiborr, N. N. (Forthc.). **Tabasaran** DoReCo dataset.

Burenhult, N. (Forthc.). **Jahai** DoReCo dataset.

Cobbinah, A. Y. (Forthc.). **Baïnounk Gubëeher** DoReCo dataset.

Cowell, A. (Forthc.). **Arapaho** DoReCo dataset.

Däbritz, C. L., Kudryakova, N., Stapert, E., & Arkhipov, A. (Forthc.). **Dolgan** DoReCo dataset.

Döhler, C. (Forthc.). **Komnzo** DoReCo dataset.

Forker, D., & Schiborr, N. N. (Forthc.). **Sanzhi Dargwa** DoReCo dataset.

Franjieh, M. (Forthc.). **Fanbyak** DoReCo dataset.

Garcia-Laguia, A. (Forthc.). **Northern Alta** DoReCo dataset.

Gipper, S., & Ballivián Torrico, J. (Forthc.). **Yurakaré** DoReCo dataset.

DoReCo corpus references (2/4)

Gippert, J. (Forthc.). Svan DoReCo dataset.

Griscom, R. (Forthc.). Asimjeeg Datooga DoReCo dataset.

Güldemann, T., Ernszt, M., Siegmund, S., & Witzlack-Makarevich, A. (Forthc.). N||ng DoReCo dataset.

Gippert, J. (Forthc.). Svan DoReCo dataset.

Griscom, R. (Forthc.). Asimjeeg Datooga DoReCo dataset.

Güldemann, T., Ernszt, M., Siegmund, S., & Witzlack-Makarevich, A. (Forthc.). N||ng DoReCo dataset.

Gusev, V., Klooster, T., Wagner-Nagy, B., & Arkhipov, A. (Forthc.). Kamas DoReCo dataset.

Haig, G., Vollmer, M., & Thiele, H. (Forthc.). Northern Kurdish (Kurmanji) DoReCo dataset.

Hartmann, I. (Forthc.). Hoocąk DoReCo dataset.

Harvey, A. (Forthc.). Gorwaa DoReCo dataset.

Haude, K. (Forthc.). Movima DoReCo dataset.

Hellwig, B. (Forthc.). Goemai DoReCo dataset.

Hellwig, B., Schneider-Blum, G., & Ismail, K. B. K. (Forthc.). Tabaq (Karko) DoReCo dataset.

DoReCo corpus references (3/4)

Kazakevich, O., & Klyachko, E. (Forthc.). **Evenki** DoReCo dataset.

Kim, S.-U. (Forthc.). **Jejuan** DoReCo dataset.

Krifka, M. (Forthc.). **Daakie** DoReCo dataset.

Meakins, F. (Forthc.). **Gurindji** DoReCo dataset.

Michaud, A. (Forthc.). **Yongning Na** DoReCo dataset.

Mosel, U. (Forthc.). **Teop** DoReCo dataset.

O'Shannessy, C. (Forthc.). **Light Warlpiri** DoReCo dataset.

O'Shannessy, C. (Forthc.). **Warlpiri** DoReCo dataset.

Ozerov, P. (Forthc.). **Anal** DoReCo dataset.

Ponsonnet, M. (Forthc.). **Dalabon** DoReCo dataset.

Quesada, J. D., Skopeteas, S., Pasamonik, C., Brokmann, C., & Fischer, F. (Forthc.). **Cabécar** DoReCo dataset.

Reiter, S. (Forthc.). **Cashinahua** DoReCo dataset.

Riesberg, S. (Forthc.). **Yali (Apahapsili)** DoReCo dataset.

DoReCo corpus references (4/4)

Ring, H. (Forthc.). Pnar DoReCo dataset.

Rose, F. (Forthc.). Mojeño Trinitario DoReCo dataset.

Schiborr, N. N. (Forthc.). English (Southern England) DoReCo dataset.

Schnell, S. (Forthc.). Vera'a DoReCo dataset.

Seifart, F. (Forthc.). Bora DoReCo dataset.

Skopeteas, S. (Forthc.). Yucatec Maya DoReCo dataset.

Skopeteas, S., Moisidi, V., Tsetereli, N., Lorenz, J., & Schröter, S. (Forthc.). Urum DoReCo dataset.

Teo, A. (Forthc.). Sümi DoReCo dataset.

Thieberger, N. (Forthc.). Nafsan (South Efate) DoReCo dataset.

Vanhove, M. (Forthc.). Beja DoReCo dataset.

Vydrina, A. (Forthc.). Kakabe DoReCo dataset.

Wegener, C. (Forthc.). Savosavo DoReCo dataset.

Wichmann, S. (Forthc.). Texistepec Popoluca DoReCo dataset.

Witzlack-Makarevich, A., Namyalo, S., Kiriggwajjo, A., & Molochieva, Z. (Forthc.). Ruuli DoReCo dataset.

Xu, X., & Bai, B. (Forthc.). Sadu DoReCo dataset.

Curation of sections (1/2)

Rationale

- Study focused on average rates and not on local variation
  ➔ a sufficiently long temporal window is necessary
- No “natural” window duration
  - Initial corpus segmentation is provided by annotation units
  - Heterogeneity across languages related to each author’s habits or interests
  - Punctuation not always present
  - Annotation units range from a few seconds to a few tens of seconds

  ➔ Need for a more consistent analysis window

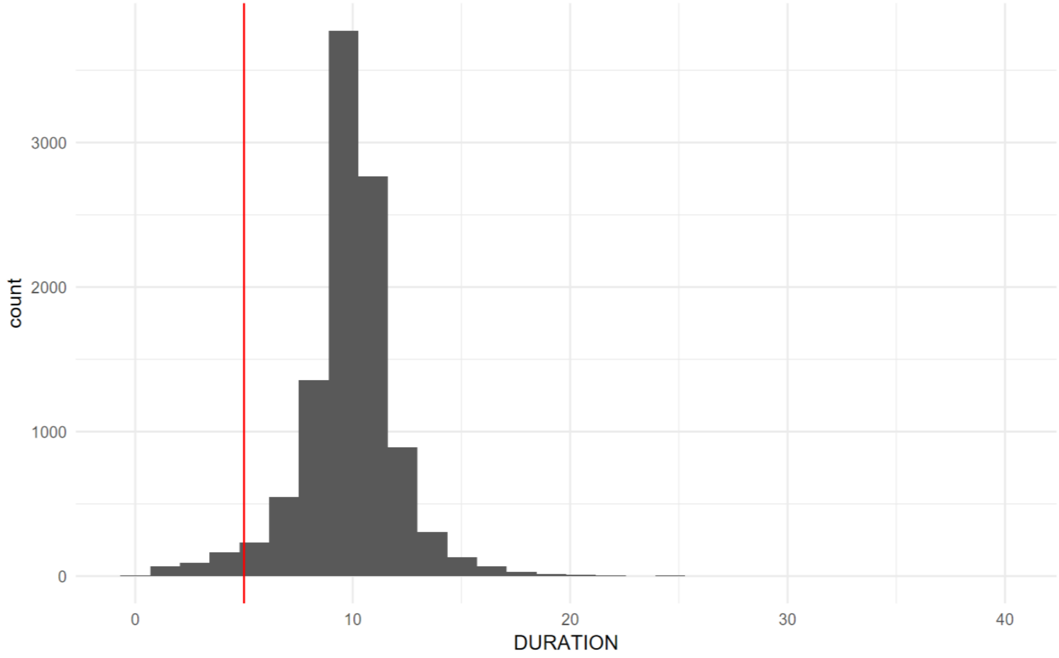
Procedure

- Short annotations units (same speaker, same file) are concatenated until their cumulative duration exceeds 10 seconds
- Long annotation units are left unchanged

Curation of sections (2/2)

Results

- Unimodal distribution of durations around 10 seconds
- Sufficient variation to study the influence of duration on rates
- Units shorter than 5 seconds (usually the last “residual” portion of a file) are removed to limit noise in estimating the indices



Estimating Information with Surprisal (1/2)

Rationale

- By definition, the surprisal of a section is the sum of the surprisal of its words. For each word, the surprisal measures the amount of information it conveys given the context consisting of the previous words in the section. It is estimated using an autoregressive language model (here a GPT-2 Large Language Model)
- Surprisal is related to human reading rate in several languages (e.g. Wilcox et al., 2023)
- It arguably provides a proxy for the amount of semantic information

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *TACL*

Estimating Information with Surprisal (2/2)

Procedure

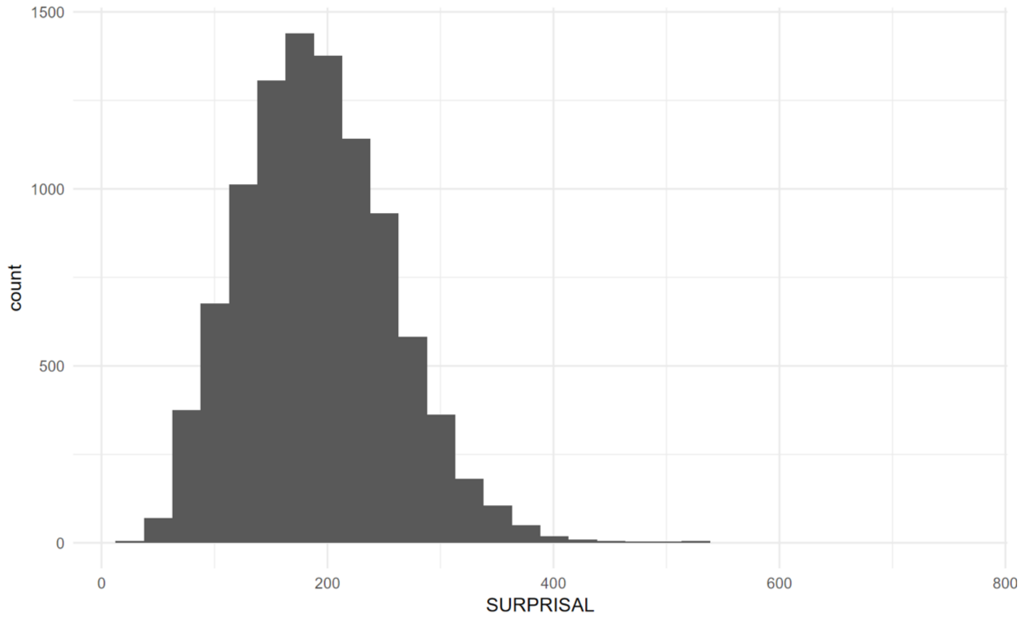
- All English translations are preprocessed to make punctuation consistent across languages
- A GPT-2 model is fine-tuned (with a next-word prediction task) on a DoReCo subset unused in the main analysis (Extended datasets without temporal alignment)
- A normalization is performed for the surprisals of Named Entities (locations and persons mainly) to correct for artefactual surprisal inflation
  - Example: GPT-2 is “less surprised” by *Alice lives in Chicago and Bob lives in Houston* than by *Firaon lives in Angguruk and Awanon lives in Hukalopunu* although the information may be very similar

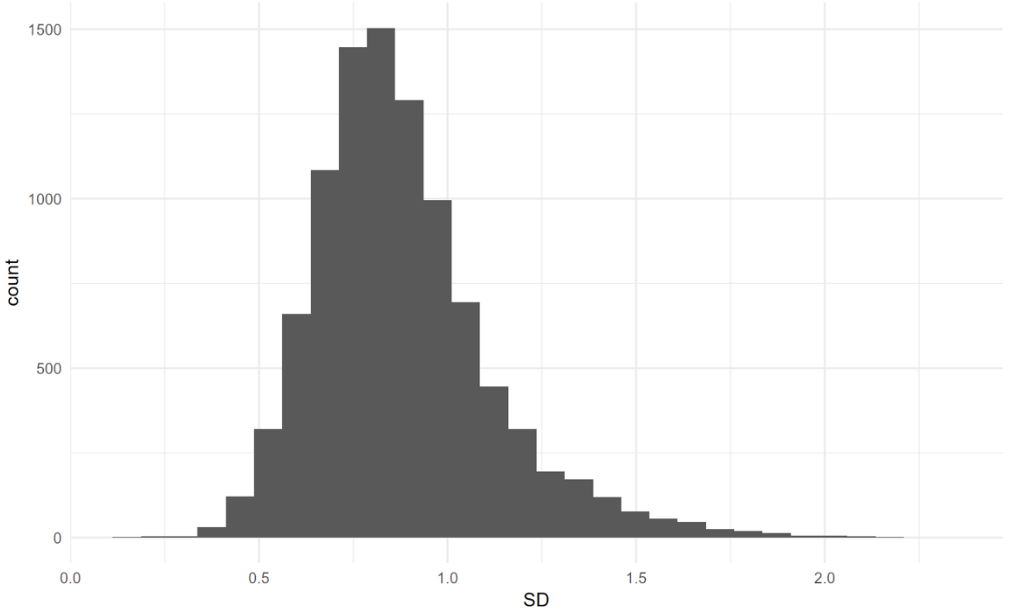
Drawback: procedure based on the translation

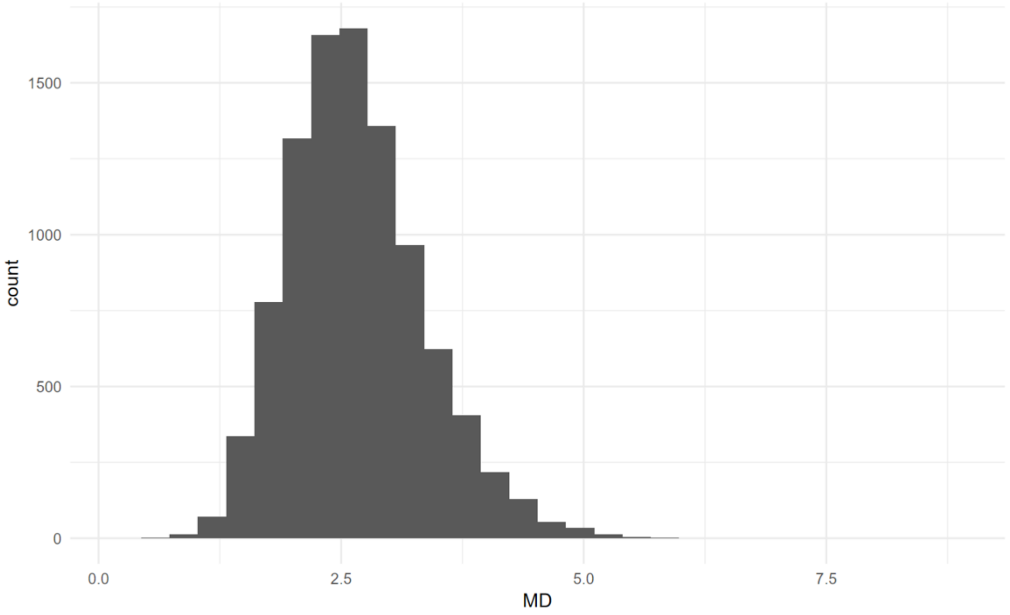
- Potential lack of precision (arguably compensated by sample size)
- No way to disentangle the contribution from the source language from the linguist’s idiosyncratic translation bias

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *TACL*

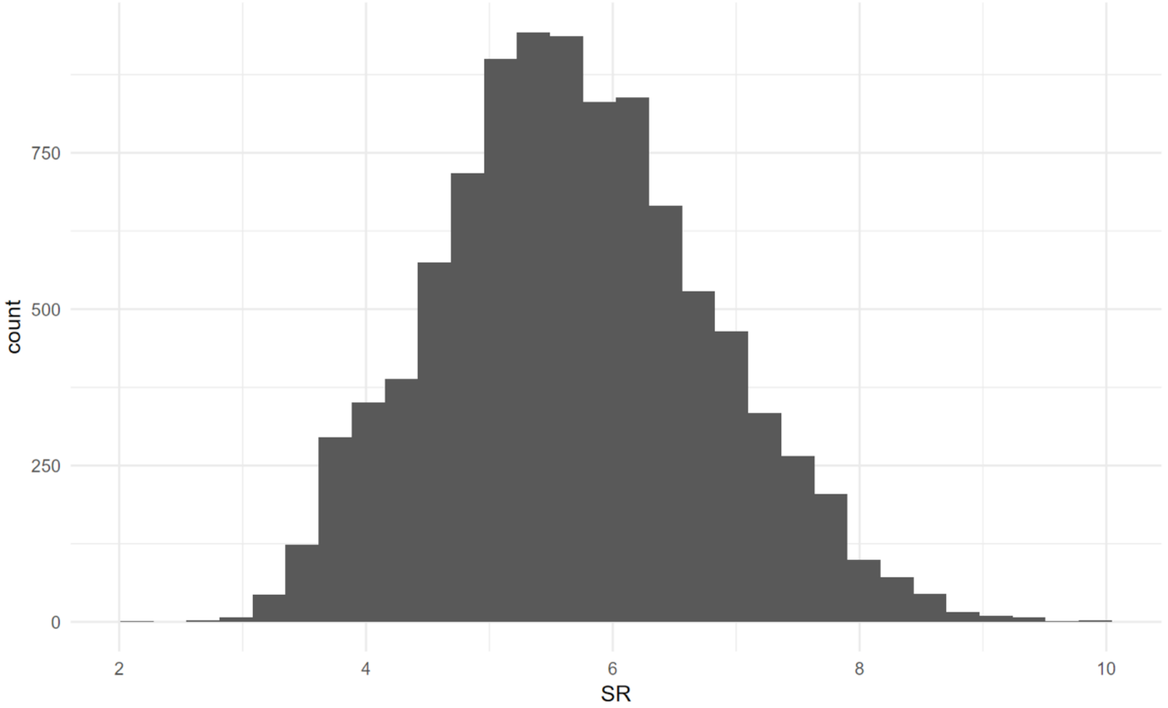
SP, SD, MD

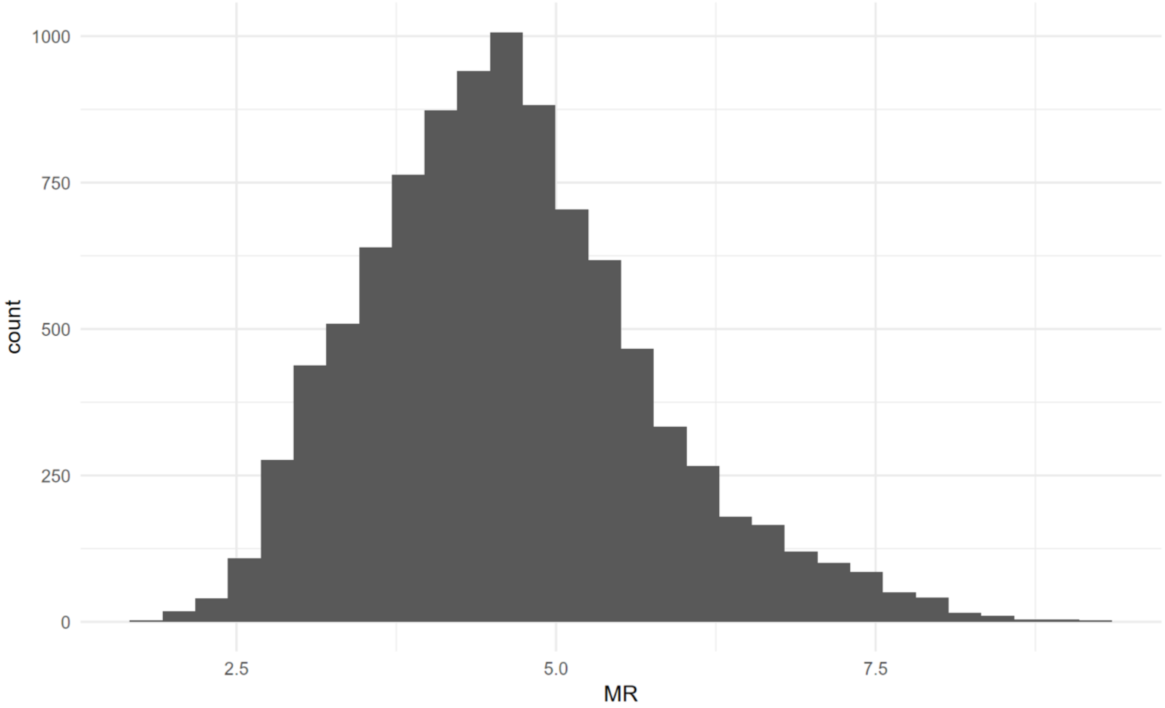

1500
1000
500
0
count
0
200
400
600
800
SURPRISAL

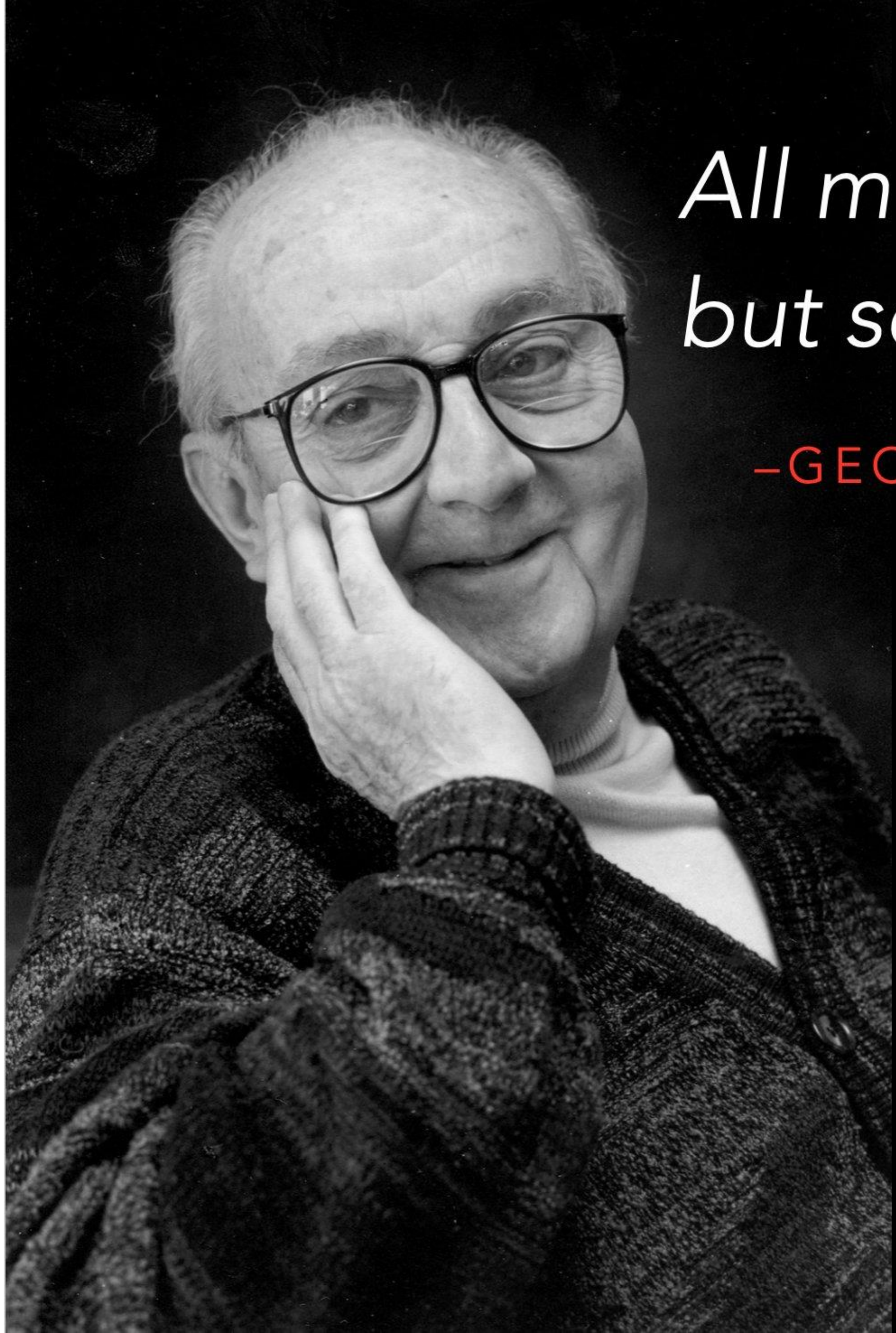

1500
1000
500
0
count
0.0
0.5
1.0
1.5
2.0
SD


1500
1000
500
0
count
0.0
2.5
5.0
7.5
MD

Distribution of SR and MR







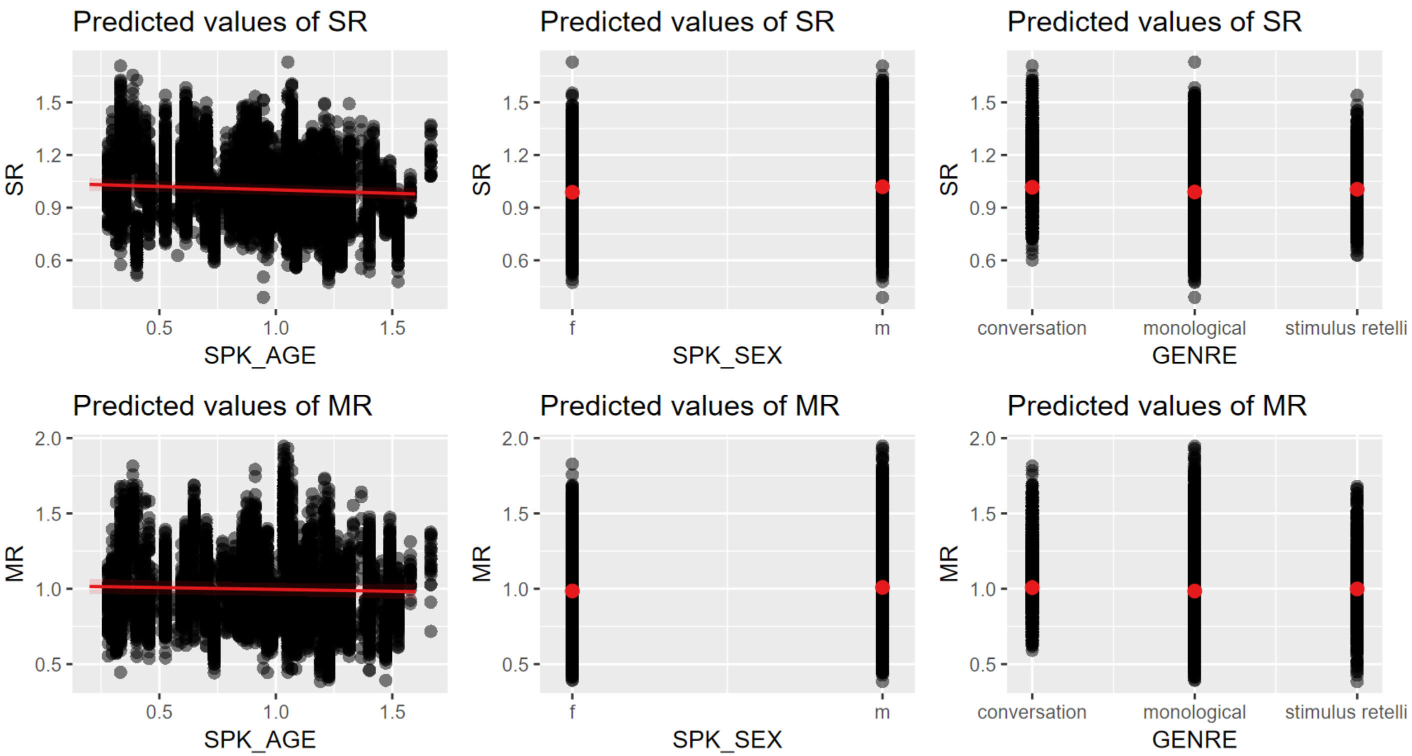
All models are wrong, but some are useful.

–GEORGE BOX, UW-MADISON

Models without interactions (1/2)

The older the speaker, the lower the syllabic rate (tendency only), but not the morphological rate.

Male speakers are faster than female speakers.



Predicted values of SR

Predicted values of SR

Predicted values of SR

Predicted values of MR

Predicted values of MR

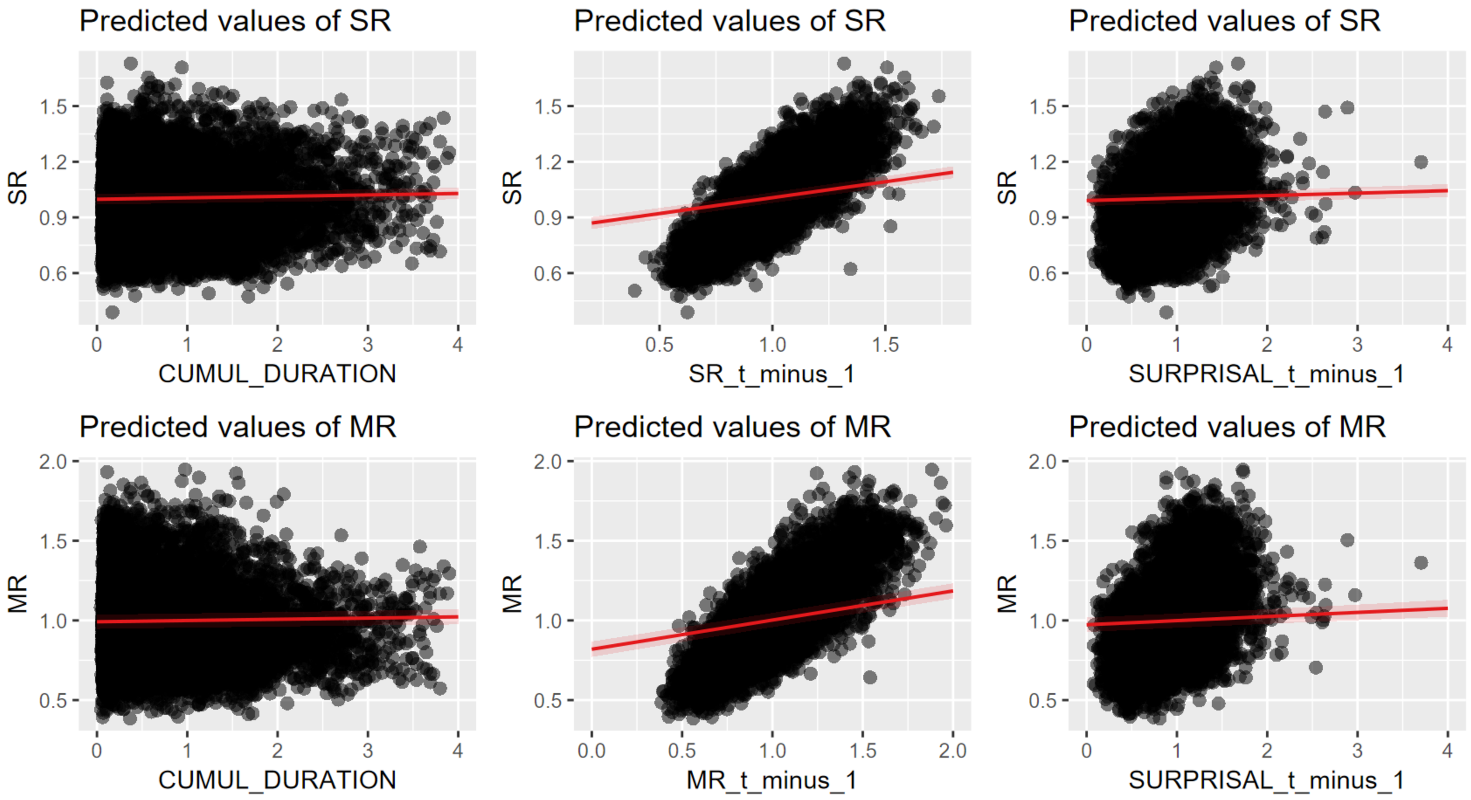
Predicted values of MR

The syllabic rate is higher in conversations than in monological

Models without interactions (2/2)

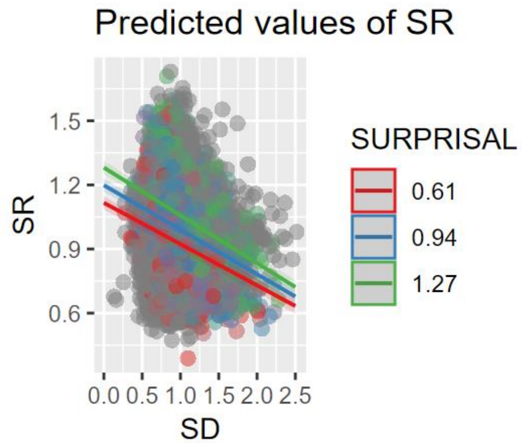
The higher the cumulative duration, the higher the speed.

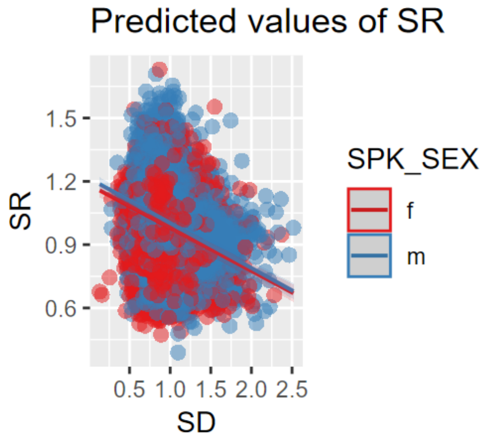
The higher the surprisal at t-1, the higher the speed (at t)

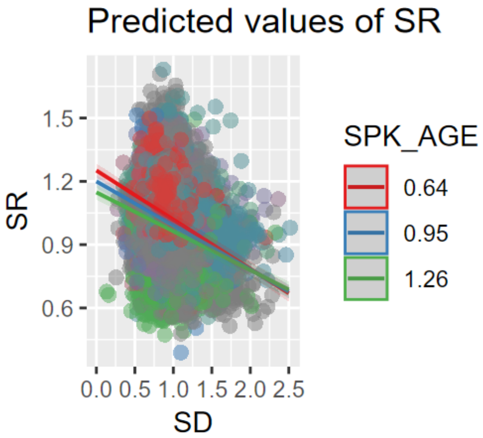


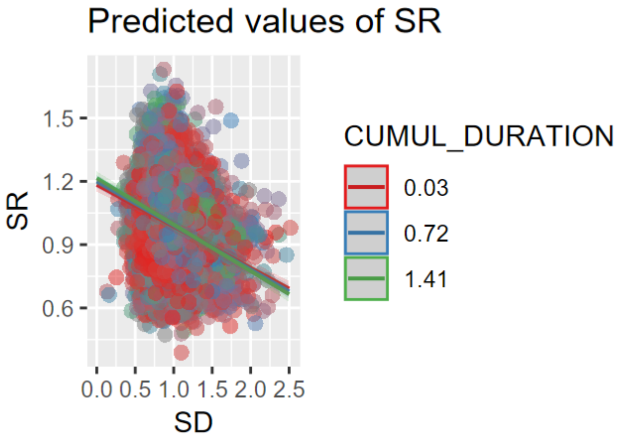
The higher the speed in the previous section, the higher the speed

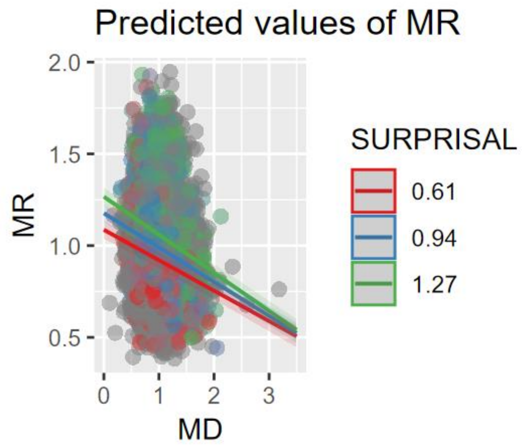
Accounting for interactions: SURPRISAL

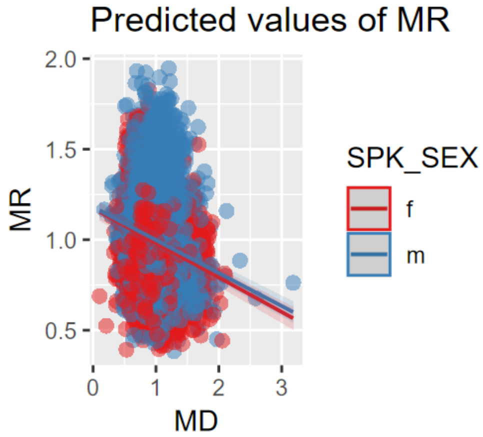

Predicted values of SR

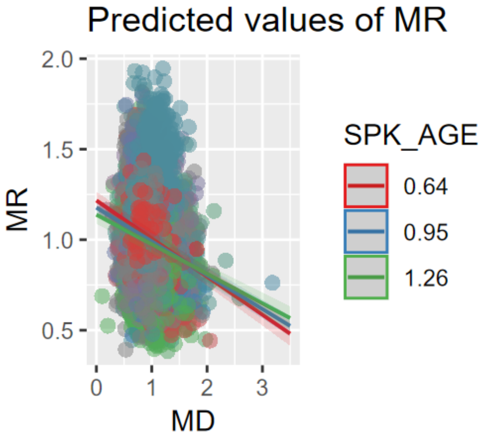

Predicted values of SR

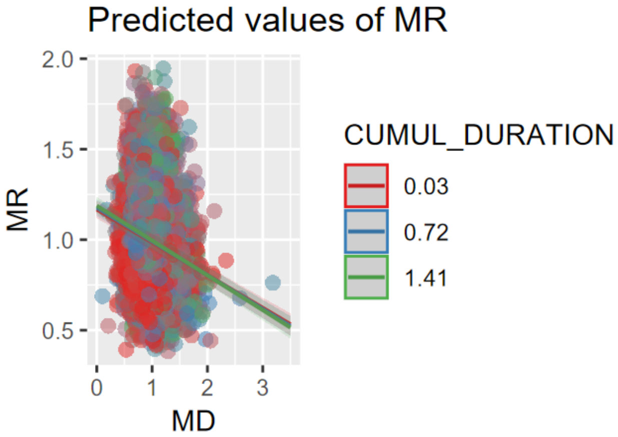

Predicted values of SR


Predicted values of SR


Predicted values of MR

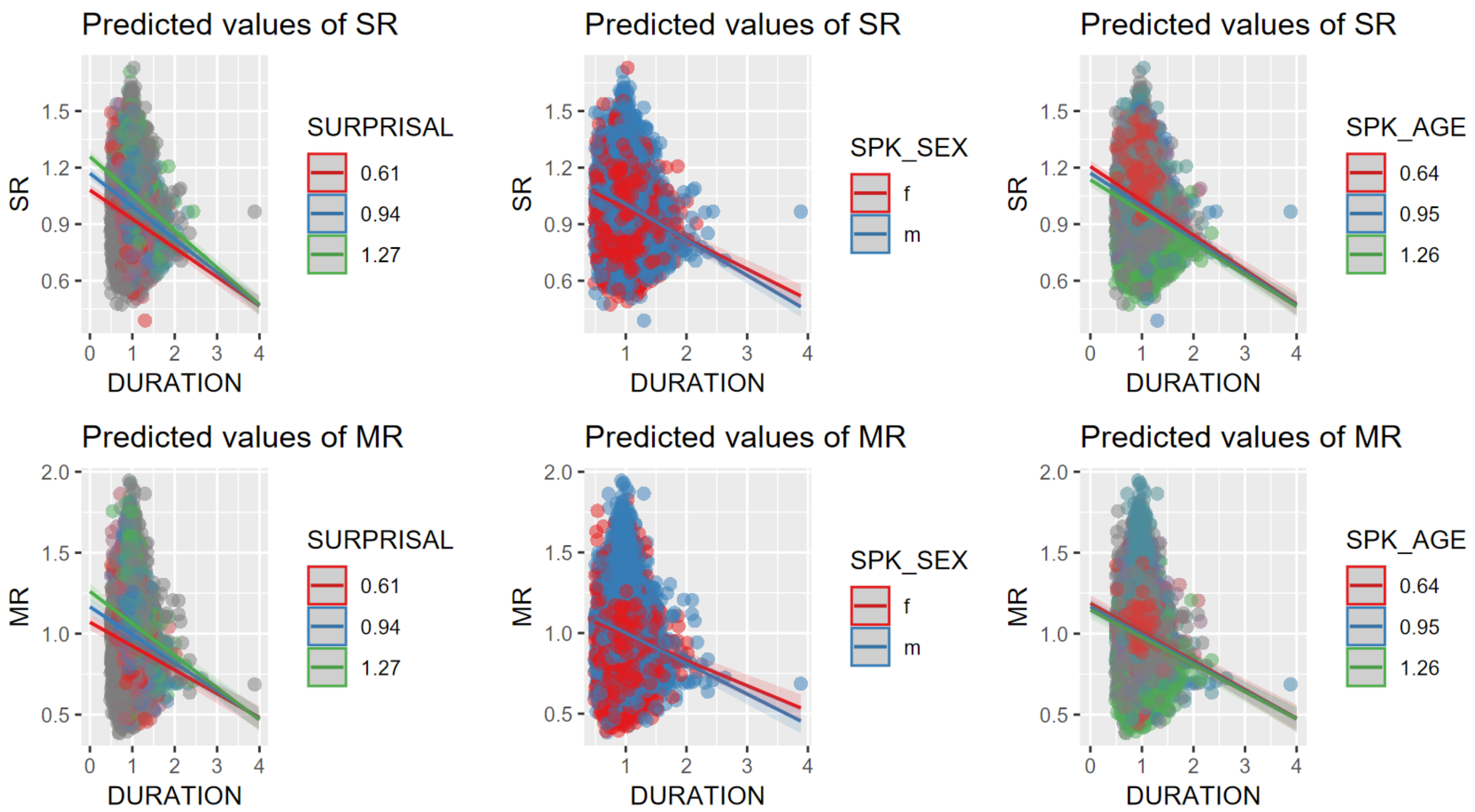

Predicted values of MR


Predicted values of MR


Predicted values of MR

Speed increases with SURPRISAL in all conditions except for large values of DURATION

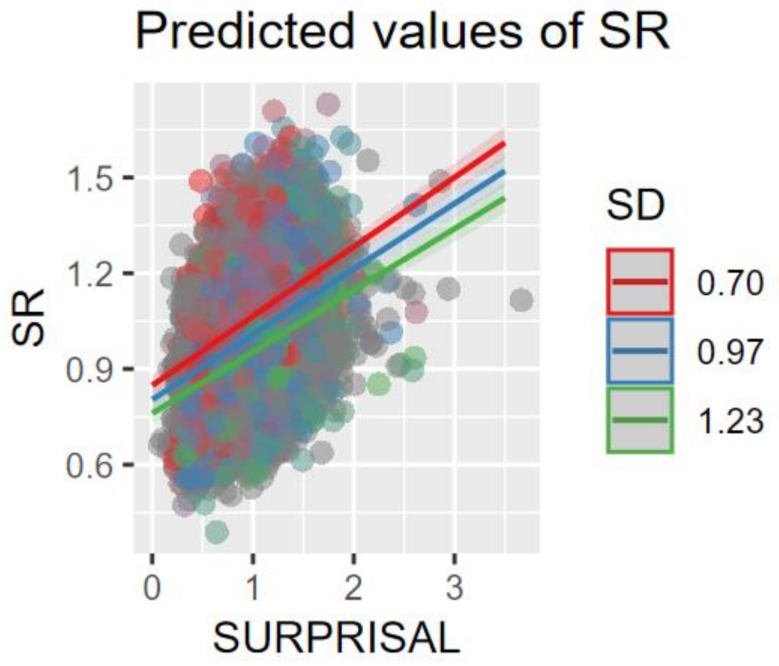
Accounting for interactions: DURATION

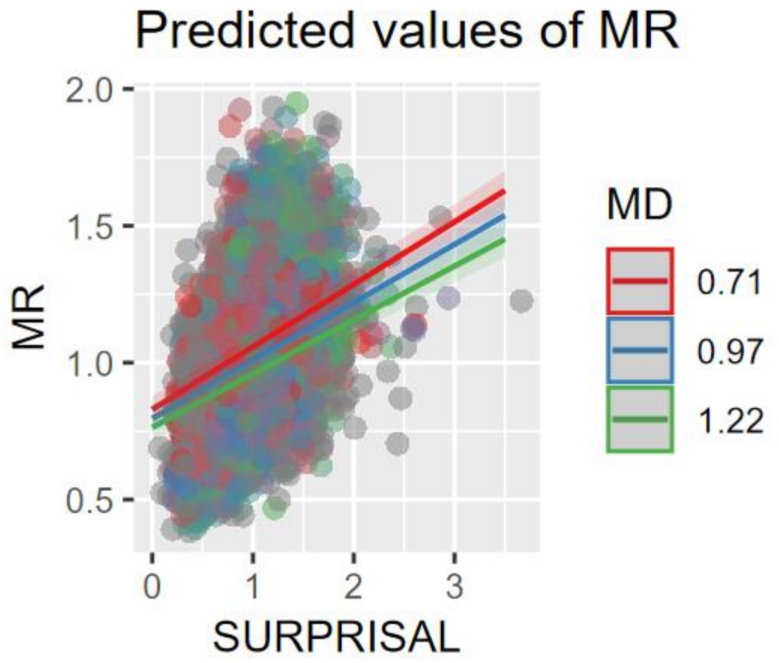


Duration negatively impacts speed regardless of the sex and age of the speaker, and regardless of the value of SURPRISAL

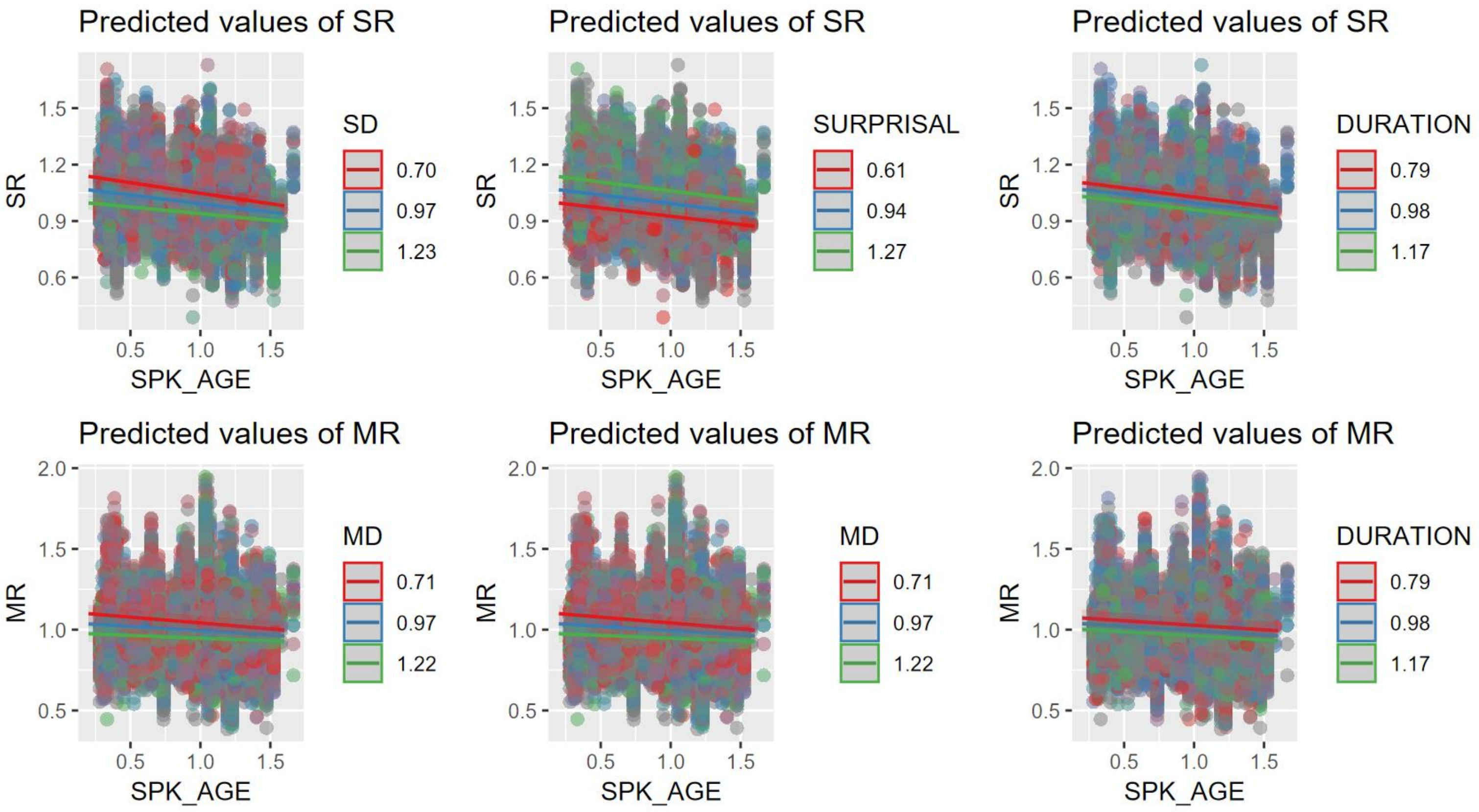
Accounting for interactions: density and surprisal

The positive effect of SURPRISAL on speed weakens as density increases, both at syllabic and morphological level.





Interactions with age



The positive effect of CUMUL_DURATION on speed weakens as age increases, but age doesn’t module the effects of SURPRISAL or DURATION