Compound Probabilistic Context-Free Grammars for Grammar Induction

Yoon Kim
Harvard University
Cambridge, MA, USA
yoonkim@seas.harvard.edu

Chris Dyer
DeepMind
London, UK
cdyer@google.com

Alexander M. Rush
Harvard University
Cambridge, MA, USA
srush@seas.harvard.edu

最先端 NLP 勉強会 2019/9/28

能地宏
産総研 AI センター

良い解説

A Review of “Compound Probabilistic Context-Free Grammars for Grammar Induction”

Ryan Cotterell Follow
Jun 30 · 7 min read

Kim et al. (2019) is one of the most exciting papers I’ve had the pleasure to read in a while. Why? It challenges long-established wisdom in the field. I’ll explain in what follows.

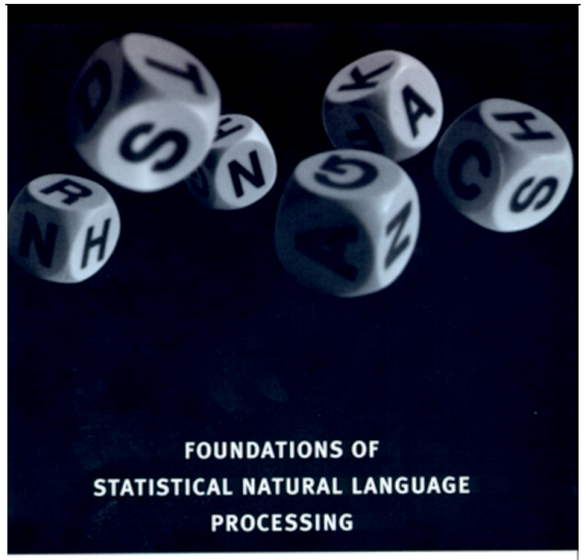
https://medium.com/@ryancotterell/a-review-of-compound-probabilistic-context-free-grammars-for-grammar-induction-2cd24ad060cf

Exciting? Why?

- **単語列から PCFG の教師なし学習が EM である程度可能**なことを初めて示した

- **EM で PCFG が学習できない**ことは常識だった

- PCFG が悪いのでは？文法に対する制約が弱いのでは (Manning & Schütze, 1999)



- 他の文法/良い文法の制約を模索するように
  - Dependency (Klein & Manning, 2004)
  - CCG (Bisk & Hockenmaier, 2013)
  - Shorter dependency (K&M), Center-embedding (Noji et al., 2016)

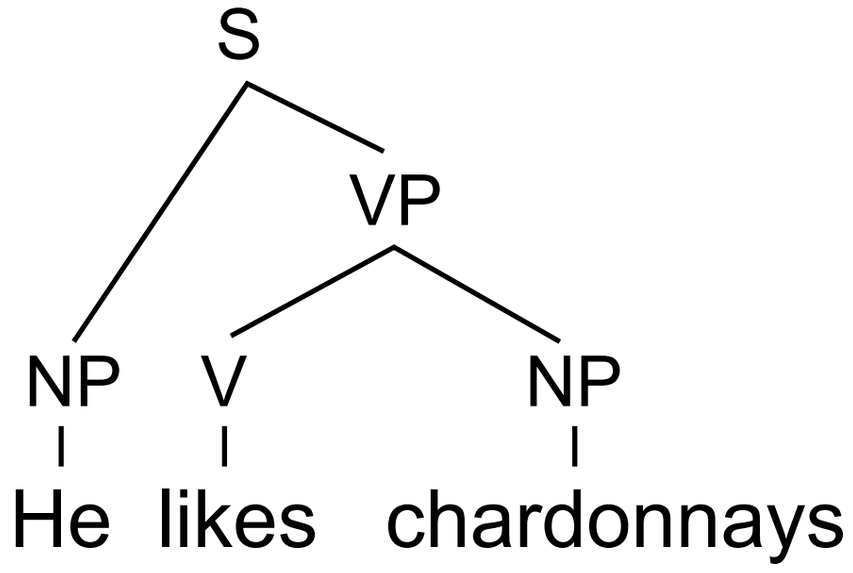
やったこと (モデル)

通常の PCFG:

$$p(\mathrm{VP} \to \mathrm{V\ NP}) = \pi_{\mathrm{VP} \to \mathrm{V\ NP}} = 0.2$$

$$p(\mathrm{VP} \to \mathrm{VP\ PP}) = \pi_{\mathrm{VP} \to \mathrm{VP\ PP}} = 0.15$$

$$p(\mathrm{V} \to \mathrm{likes}) = \pi_{\mathrm{V} \to \mathrm{likes}} = 0.002$$



提案モデル (各ルール確率をニューラルでパラメタ化)

$$p(\mathrm{VP} \to \mathrm{V\ NP}) = \frac{\exp(\mathbf{u}_{\mathrm{V,NP}}^{\top} \mathbf{w}_{\mathrm{VP}})}{\sum_{BC} \exp(\mathbf{u}_{BC}^{\top} \mathbf{w}_{\mathrm{VP}})}$$

- Painless unsupervised learning with features
  (Berg-kirkpatrick et al., 2010) のニューラル化とみなせる

- もう少し工夫したモデル (compound PCFG) も提案

学習について

- 入力: 生文 (not 品詞列) の集合
- 目的関数: 文の対数尤度の最大化
  - 従来の PCFG: inside-outside でルールの期待値を計算して EM
  - 本モデル: 内側確率だけ求めて、 SGD でそれを最大化
    - SGD だが EM とやっていることはほぼ同じ

      (Inside-Outside and Forward-Backward Algorithms Are Just Backprop)
- なぜうまくいくか？
  - 巨大な (over-parameterize) ニューラルモデルは SGD で最適解が探しやすいため、と著者らは主張

Gold tree と比較 (シンボルは無視)

Model	PTB Mean	PTB Max	CTB Mean	CTB Max
Left Branching	8.7		9.7	
Right Branching	39.5		20.0	
Random Trees	19.2	19.5	15.7	16.0
PRPN (tuned)	47.3	47.9	30.4	31.5
ON (tuned)	48.1	50.0	25.4	25.7
Neural PCFG	50.8	52.6	25.7	29.5
Compound PCFG	55.2	60.1	36.0	39.8

Ordered neurons に使われる木構造復元アルゴリズムは right-branching に過剰なバイアスをかけるという批判

https://arxiv.org/pdf/1909.09428.pdf (Dyer, et al., 2019/9/23)

教師なし構文解析の闇

ほとんどの手法はハイパーパラメータの変化に敏感
しかしこれを調整すると教師なしでなくなってしまう

> Relatedly, the models were quite sensitive to parameterization (e.g. it was important to use residual layers for f_1, f_2), grammar size, and optimization method. Finally, despite vectorized GPU im-

本モデルがどれぐらいパラメータに敏感かは回してみないと分からない

実装は公開済み: https://github.com/harvardnlp/compound-pcfg

What Kind of Language Is Hard to Language-Model?

Sebastian J. Mielke[1] **Ryan Cotterell**[1] **Kyle Gorman**[2,3] **Brian Roark**[3] **Jason Eisner**[1]
[1] Department of Computer Science, Johns Hopkins University
[2] Program in Linguistics, Graduate Center, City University of New York [3] Google
{sjmielke@,ryan.cotterell@}jhu.edu kgorman@gc.cuny.edu
roark@google.com jason@cs.jhu.edu

著者スライド

https://sjmielke.com/docs/MieCotGor19What_slides.pdf

興味の中心:

現在のNLP技術はどれぐらい言語の特性に依存するか？

図表は論文と著者スライドから抜粋

去年の研究との関係

NAACL 2018

Are All Languages Equally Hard to Language-Model?

Ryan Cotterell[1] and **Sebastian J. Mielke**[1] and **Jason Eisner**[1] and **Brian Roark**[2]
[1] Department of Computer Science, Johns Hopkins University [2] Google

⇒ 形態素の活用が複雑な言語ほど難しい

What Kind of Language Is Hard to Language-Model? ACL 2019

Sebastian J. Mielke[1] **Ryan Cotterell**[1] **Kyle Gorman**[2,3] **Brian Roark**[3] **Jason Eisner**[1]
[1] Department of Computer Science, Johns Hopkins University
[2] Program in Linguistics, Graduate Center, City University of New York [3] Google

去年の発見は間違っていました

多言語を通じた分析は丁寧に行いましょう

Research questions

What Kind of Language Is Hard to Language-Model?

- Q1: (そもそも) 言語によって言語モデルの難しさは変わる？

  YES. 現在の言語モデルにとって英語よりドイツ語の方が難しい

- Q2: どのような言語が難しいか？言語学的な特徴はある？

  Unclear. 活用の複雑さなど分かりやすい指標では説明できない

  テクニカルな要素でしか説明できない（語彙数など）

- Q3: Translationese は native の言語より簡単か？

  NO. 言語モデルにとって、両者に違いはあるが、簡単ではない

準備: 異なる言語をどう比較するか

		$p(\cdot)$	$\Rightarrow$ NLL
en	I love Florence!	0.03	$\Rightarrow$ 5 bits
de	Ich grüße meine Oma und die Familie dahein.	0.008	$\Rightarrow$ 7 bits
nl	Alle mensen worden vrij en gelijk in waardigheid en rechten geboren.	0.0004	$\Rightarrow$ 11 bits

- 問題点:
  - 言語モデルの性能 (パープレキシティ) はコーパス依存
  - 言語が異なるとコーパスが異なる ⇒ 言語間の比較？
- 提案する方法論 (Cotterell et al., 2018)
  - 内容が同じコーパス = 翻訳データを使う (Europerl, Bible)
  - UNK なし言語モデルを使う (BPE-LM, char-LM)

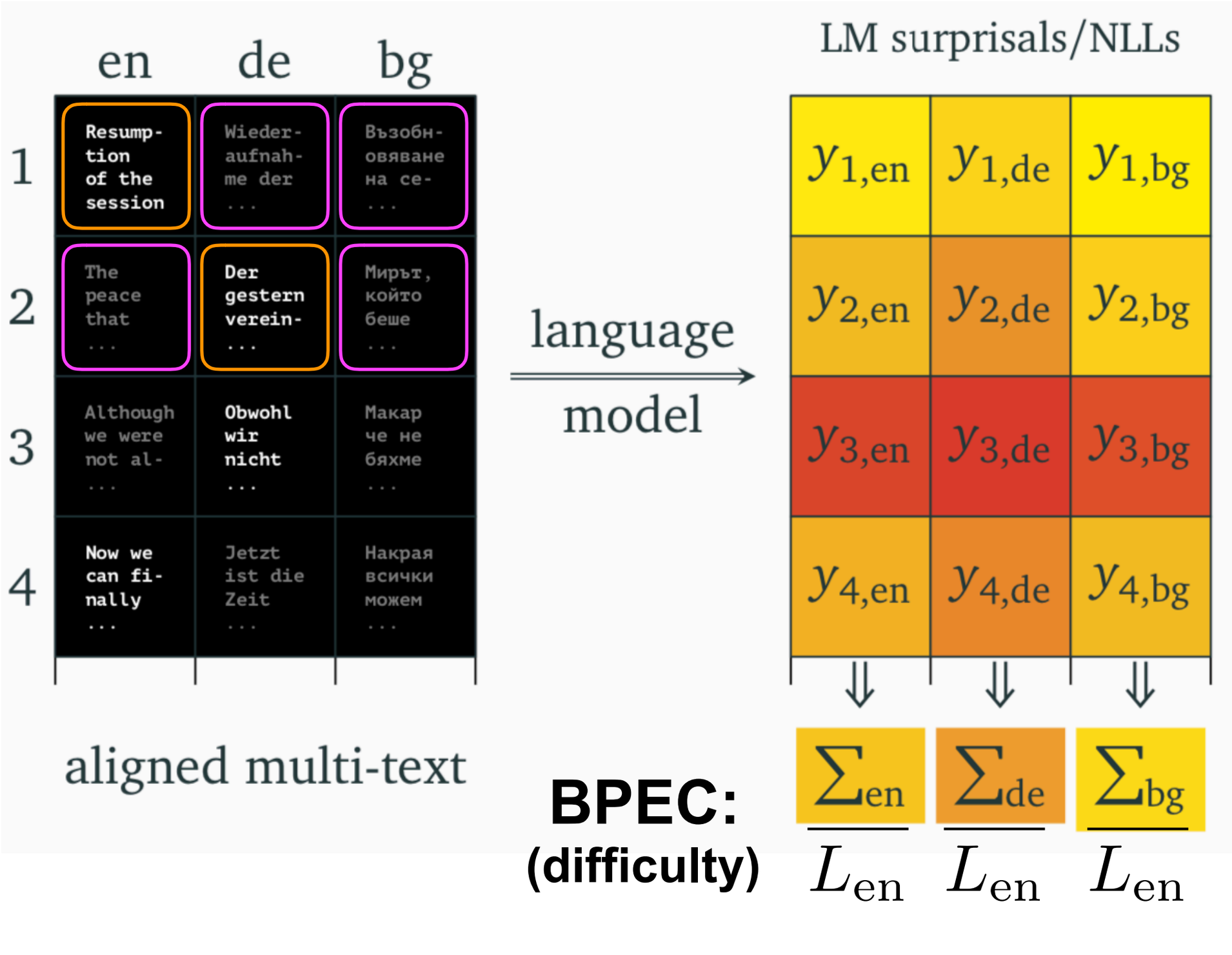
“難しさ” の計算方法

		$p(\cdot)$	$\Rightarrow$ NLL
en	Resumption of the session.	0.013	$\Rightarrow$ 6.5 bits
de	Wiederaufnahme der Sitzung.	0.011	$\Rightarrow$ 6.3 bits
nl	Hervatting van de sessie.	0.012	$\Rightarrow$ 6.4 bits

- 翻訳文だとしても … 語彙が異なる; 単語長/文長が異なる
  - 言語間の差を吸収した尺度 (言語間で比較できる尺度) が欲しい
- 過去の方法: Bits per English character (BPEC) (Cotterell 18)
  - 対数尤度を “英語コーパスの文字数” で割って補正
  - 全ての言語に渡って文の alignment が取れている必要がある
  - Europerl ではある程度可能; Cotterell 18 は Europerl のみ使用
  - 今年は言語を拡張したい ⇒ Bible ⇒ alignment が不完全

Fully parallel corpora

Original Translated

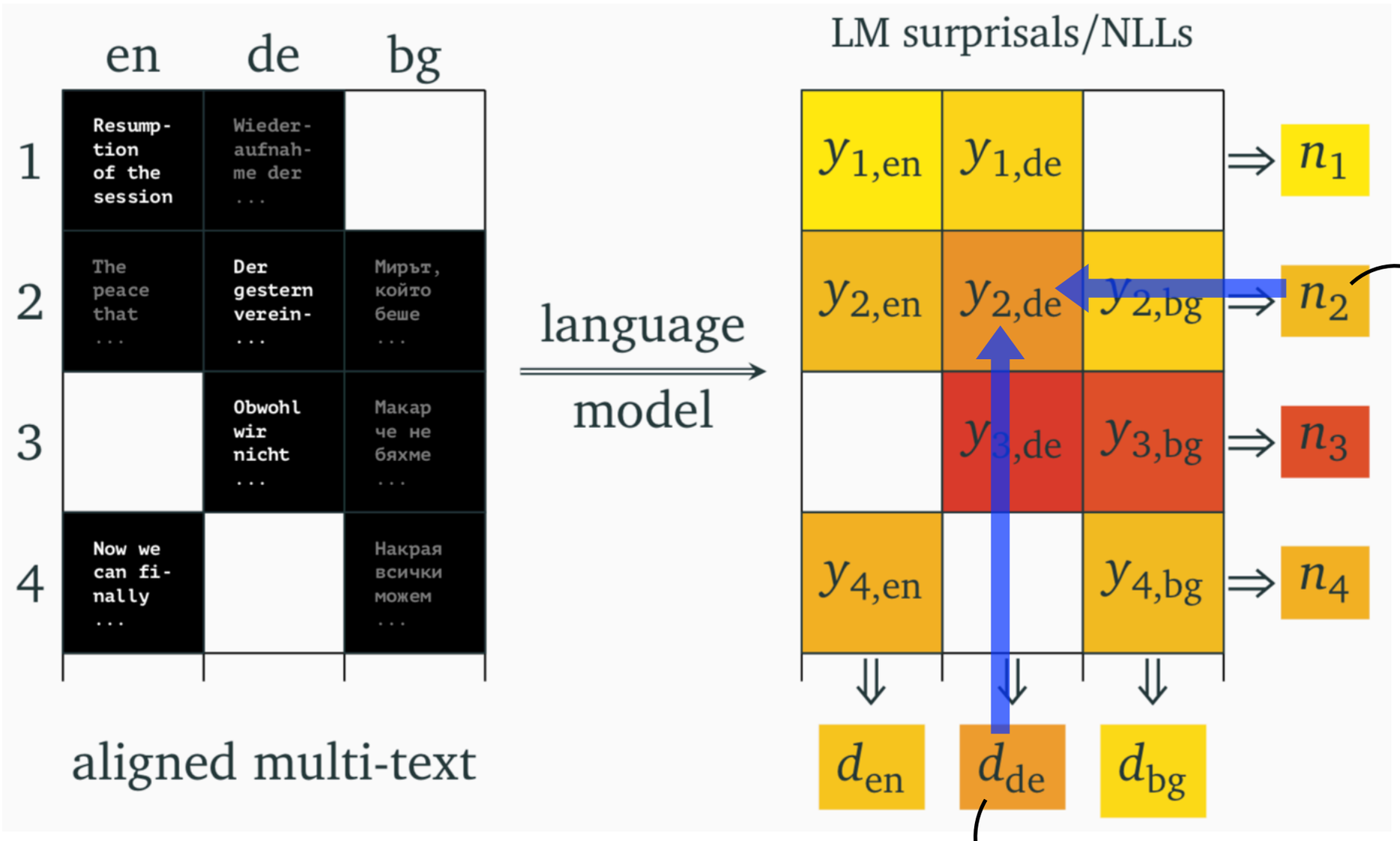


BPEC:
(difficulty)

欠損値 (文) が存在 ⇒ 統計モデルで回帰

混合効果積モデル (multiplicative mixed-effects)

$$y_{2,\mathrm{de}} \sim n_2 \cdot \exp d_{\mathrm{de}}$$

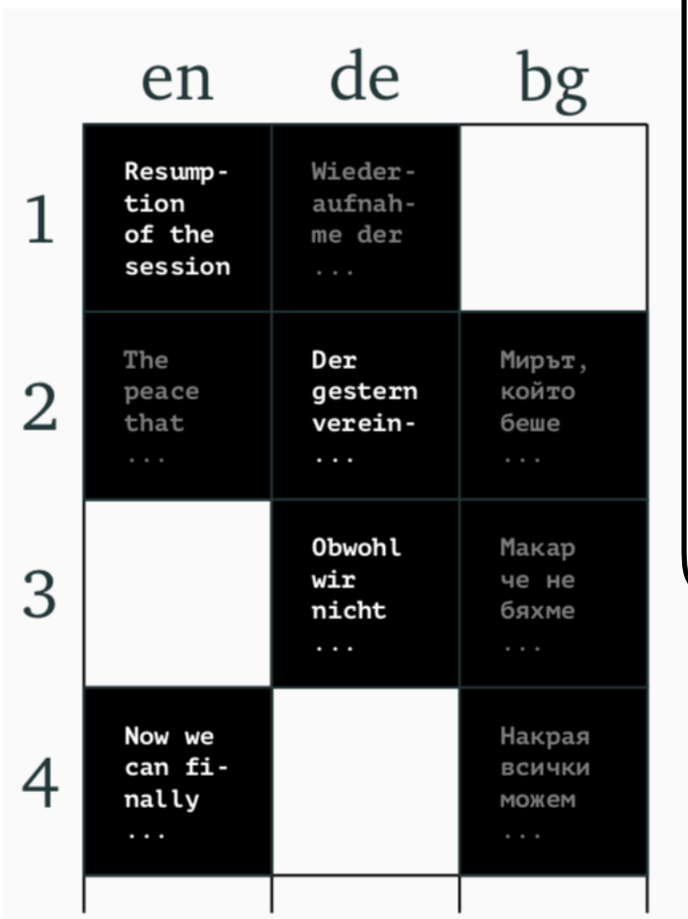


2文目の抽象的な”情報量”

ドイツ語の”難しさ”を表す量

欠損値 (文) が存在 ⇒ 統計モデルで回帰

混合効果積モデル (multiplicative mixed-effects)



aligned multi-text

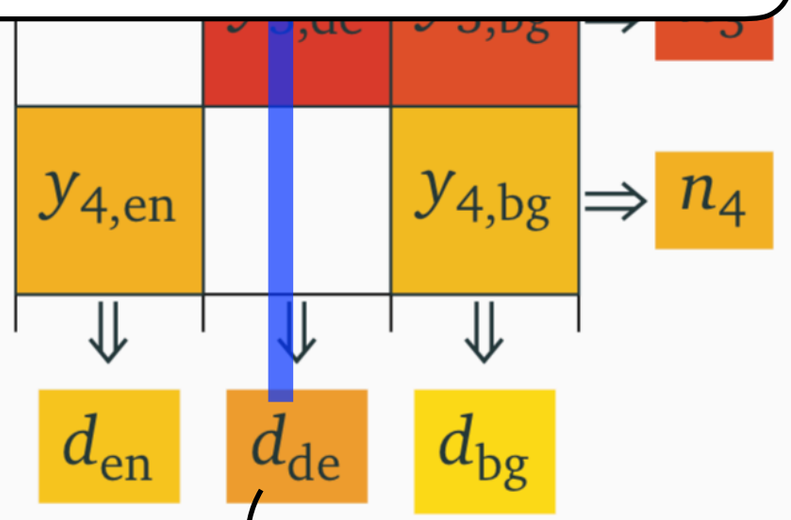
実際のモデル (モデル2)

$$y_{ij} = n_i \cdot \exp(d_j) \cdot \exp(\epsilon_{ij})$$

$$\sigma_i^2 = \ln\left(1 + \frac{\exp(\sigma^2)-1}{n_i}\right)$$

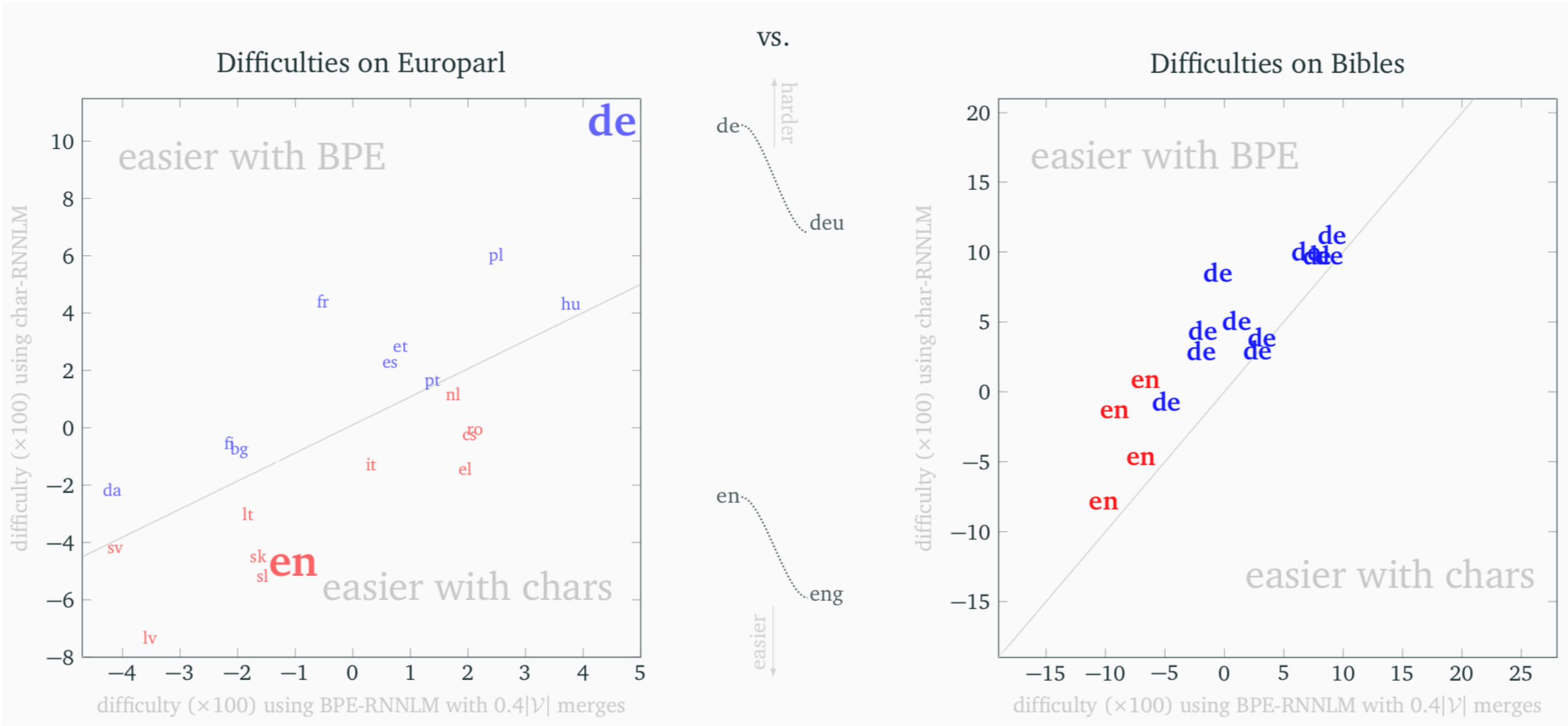
$$\epsilon_{ij} \sim \mathcal{N}\left(\frac{\sigma^2-\sigma_i^2}{2},\ \sigma_i^2\right),$$

2文目の抽象的な”情報量”



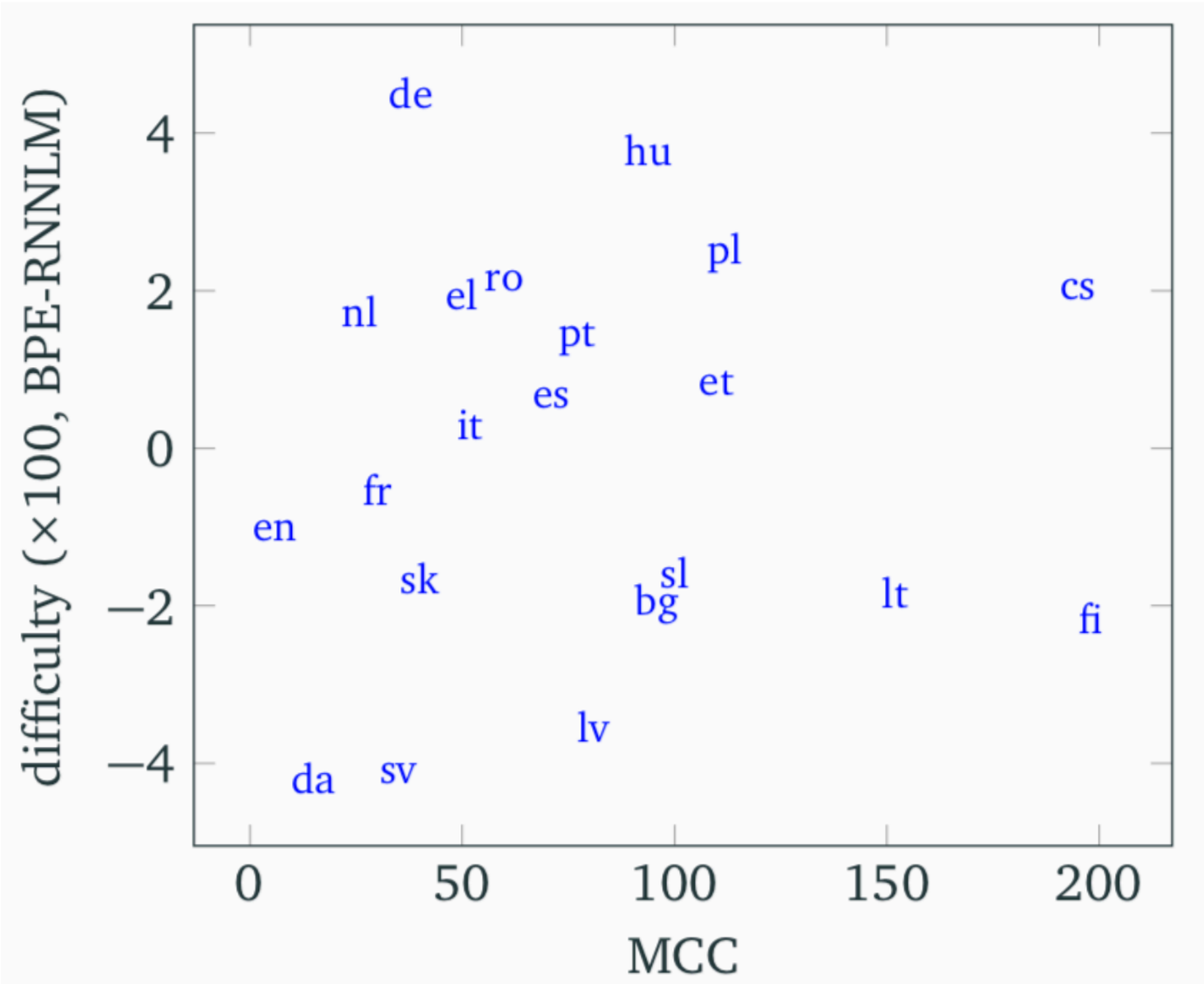
ドイツ語の”難しさ”を表す量

どの言語が難しいか (with BPE/char-RNNLM)



- 複数の言語モデル、複数のコーパスで一貫した結果
  ⇒ より信頼性のある観察と言える (皆さんもしましょう)

難しさは言語の特性によるか？



MCC = morphological counting complexity

(Cotterell 18 で説明能力のあった量)

⇒ 相関関係は認められない (良い統計モデルにより結果が変わった)

他の複雑性の指標ではどうか？

- WALS features / dependency length / Head-POS entropy
  - … どれも優位な説明能力なし

- より単純な指標はどうか？
  - Difficulty for char-RNNLM: raw-character length と相関
    - Bibles: p < .001; Europerl: p<.01
    - RNN の記憶能力のため（長い履歴がより生じるため）
  - Difficulty for BPE-RNNLM: 語彙サイズと相関
    - Bibles: p < .00000000001; Europerl: 相関なし
    - 語彙が多いと BPE が切っても低頻度語が多くなるから

どちらも RNN の特性により生じた影響といえる (テクニカル)
言語学的に面白い結果は得られなかった (Cotterell 18 の否定)

Takeaway

- Cross-linguistic study は丁寧に行いましょう
- 少量の一部の言語のみをみて主張を一般化してはいけない
- でも ACL community は英語にしか興味ないし、
  cross-lingual なリソースが全然増えなくて悲しいね