

ConLoan: A Contrastive Multilingual Dataset for Evaluating Loanwords

Sina Ahmadi Micha David Hess Elena Álvarez Mellado Alessia Battisti Cui Ding
Anne Göhring Yingqiang Gao Zifan Jiang Andrianos Michail Peshmerge Morad
Joel Niklaus Maria Christina Panagiotopoulou Stefano Perrella Juri Opitz
Anastassia Shaitarova Rico Sennrich

[ACL 2025](#)

Presented by TANAKA Kunitomo
(Sasano Lab., Nagoya University)
2025-09-26

概要

- 借用語を本来語へ書き換えた多言語データセット **ConLoan** を構築
- ConLoan を用いて言語モデルの選好や機械翻訳への影響を調査

Portuguese for “Indeed, this is one advantage of the present system.”

Source

É, na verdade,
uma **vantagem** do sistema actual.

Target

É, na verdade,
uma **benefício** do sistema actual.

目次

- 背景 4
 - 借用 4
 - NLP からみる借用語 5
- データセット構築 7
- モデル評価 14

背景 借用

借用 : 他言語から表現が採り入れられる言語現象
borrowing

- **語彙借用** : 借用のうち語彙レベルで起こるもの [補遺]
lexical borrowing

▶ **借用語** : 他言語の語彙を音ごと借用 ↔ **本来語** (native word)
loanword

例: トークナイゼーション < tokenization ↔ 分かち書き

社会的な動機

他の言語の影響をなるべく省こうとする動きは 言語純粹主義と呼ばれる
language purism

- 優勢な言語から危機言語を保護
- 新しい概念・技術の導入

背景

NLP からみる借用語

タスクとしては、借用語の特定 + 本来語への置換

課題

- (特に少数言語の) 注釈ありデータの不足
- 借用語の置換にまで目が向けられてこなかった
 - せいぜい辞書ベースの逐語訳レベル

→ 文脈ありの多言語データセットを構築して NLP の視点から分析

目次

- 背景 4
- データセット構築 7
 - 概要 7
 - 事例収集 9
 - アノテーション準備 10
 - アノテーション 11
 - 統計情報 12
- モデル評価 14

データセット構築 概要

- 原語文：借用語と本来語を対照する対 + 英訳文
 - 原語文中の借用語/本来語はマークアップ

対象言語: 10 言語

- 中国語・フランス語・ドイツ語・ギリシャ語・アイスランド語・イタリア語・北部クルド語・ポルトガル語・ロシア語・スペイン語

Portuguese for “Indeed, this is one advantage of the present system.”

Source

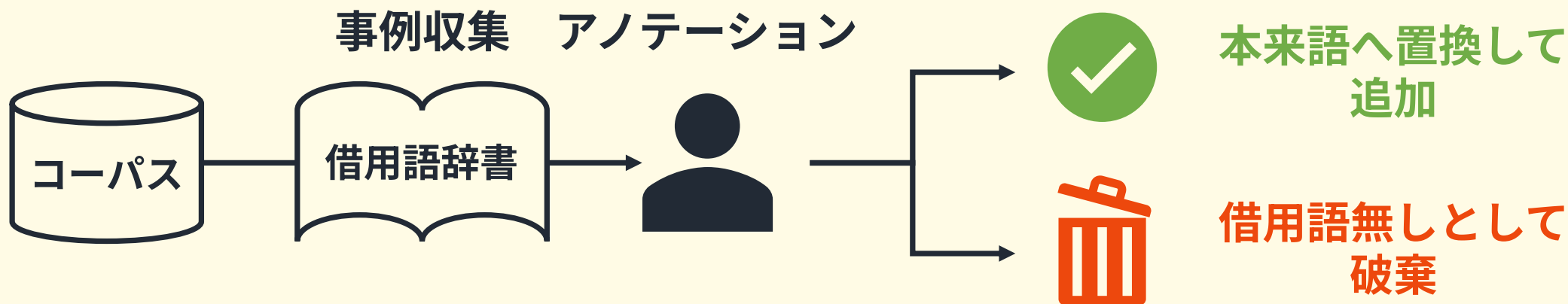
É, na verdade,
uma **vantagem** do sistema actual.

Target

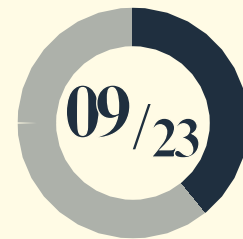
É, na verdade,
uma **benefício** do sistema actual.

データセット構築 概要

1. コーパスから借用語が入っていそうな事例を選定
2. 実際に借用語が含まれているかを判定し、できるだけ本来語へ置換



データセット構築 事例収集



コーパスから借用語が入っていそうな事例を選定

各言語 — 英語の対訳コーパスから借用語を含む事例を抽出

- コーパス: OPUS-MT ([Tiedemann & Thottingal EAMT'20](#))
- 辞書: 主に Wiktionary

検出された借用語はマークアップされる

- が、事例中の全ての借用語がマークアップされるわけではない
(事例が複数の借用語を含み、かつ一部が立項されていない場合)

アノテーション準備

事例に含まれる借用語を特定し、できるだけ本来語へ置換

- その補助として特定された借用語の類義語をアノータに提供
 - ▶ WordNet から取得
- その他借用語特定のためのガイドラインが与えられる [補遺]

Portuguese for “Indeed, this is one advantage of the present system.”

Source

É, na verdade,
uma **vantagem** do sistema actual.

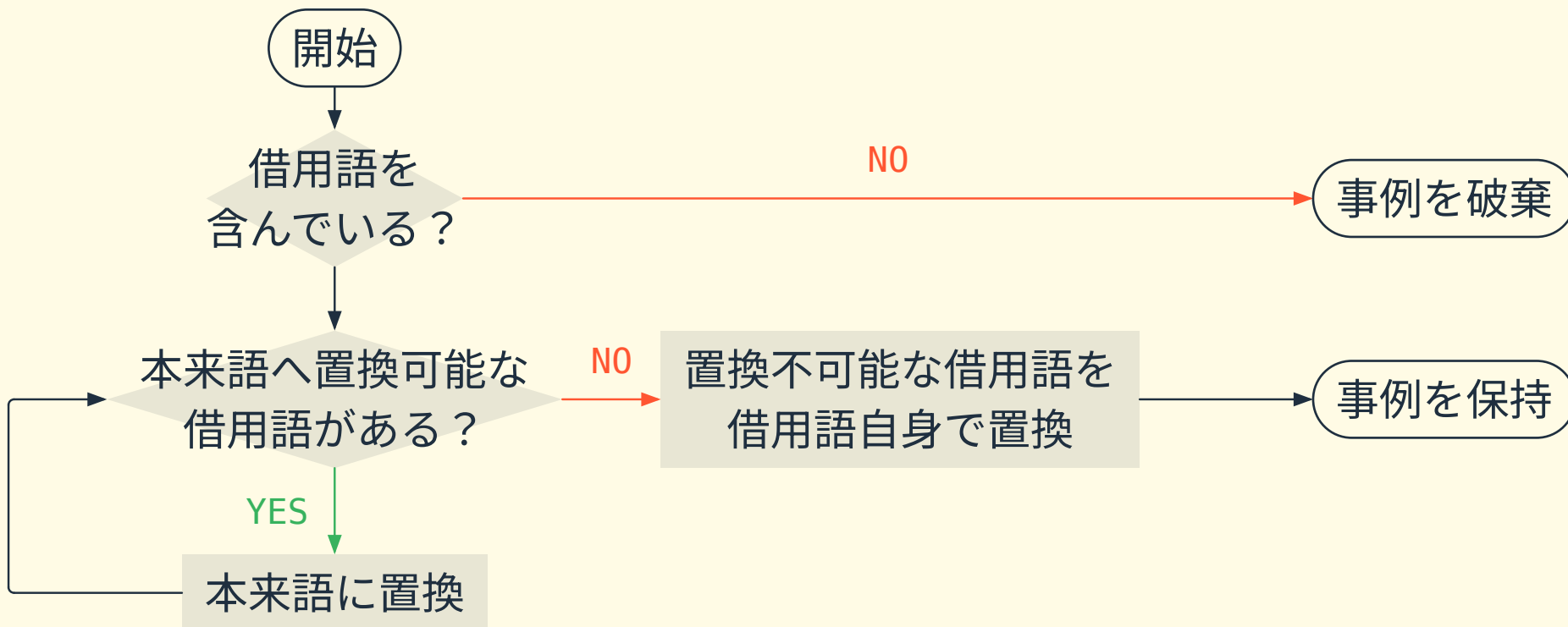
Target

É, na verdade,
uma **?** do sistema actual.

Suggestions: ‘propriedade’, ‘recurso’, ‘benefício’, ‘bem’, ‘handicap’, ‘utilidade’, ‘lucro’

アノテーション

事例に含まれる借用語を特定し、できるだけ本来語へ置換
(辞書で検出されなかった借用語も含む)



データセット構築 統計情報

- 9599 / 16870 例で借用語を確認
 - ▶ 辞書ベースの検出精度は約 57%
- 英語から借用している言語が多い
 - ▶ 元データの偏りでありうることに注意

言語	人手検出事例	辞書検出事例	検出された借用語数	置換した借用語数	借用元 1 位
ZHO	639	1328	779	186	English
FRA	413	1315	442	237	English
DEU	1162	1861	1282	1095	Latin
ELL	503	1813	515	395	French
ISL	2902	4023	3076	1130	Latin
ITA	595	1780	652	589	English
KMR	401	630	628	528	English
POR	2336	2509	2568	862	French
RUS	298	300	1051	1039	Latin
SPA	350	1311	387	287	English
Total	9599	16870	11380	6348	English

目次

- 背景 4
- データセット構築 7
- モデル評価 14
 - 置換前後の surprisal 14
 - 翻訳モデルの評価 16

置換前後の surprisal

言語モデルは借用語を含む文と本来語のものどちらを予測しづらいか？

Surprisal (sentence-level)

「文 x 中の各語 x_t が文脈 $x_{<t}$ からどれだけ予測されにくいかを合計」

$$\text{Surprisal}(x) \stackrel{\text{def}}{=} - \sum_t \log p(x_t \mid x_{<t})$$

使用モデル

Llama-3.1-8B (8-bit quantized), Llama-2.7-7B (8-bit quantized), EuroLLM

- どのモデルも同様の傾向を示す

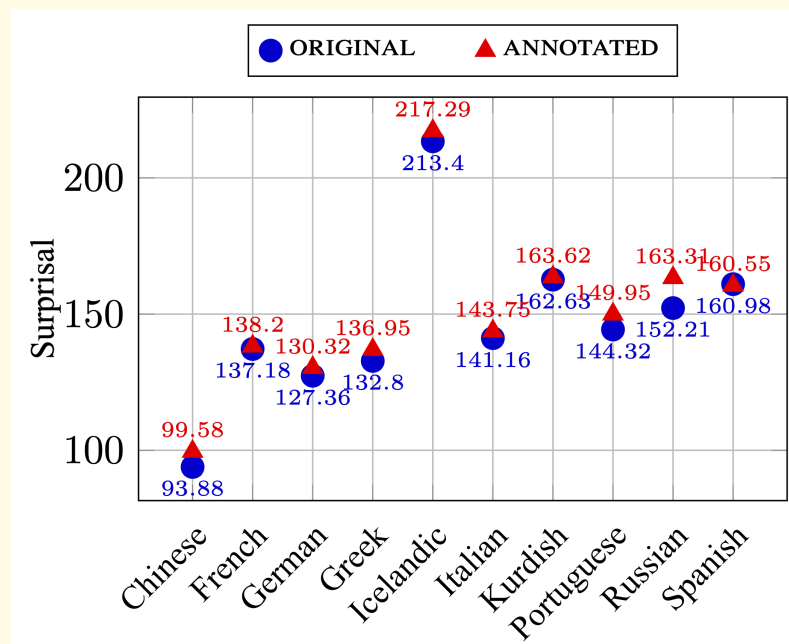
置換前後の surprisal

言語モデルは借用語を含む文と本来語の文のどちらを“好む”か？

結果・考察

置換前（借用語）の方が値が小さい ($p < 0.01$)

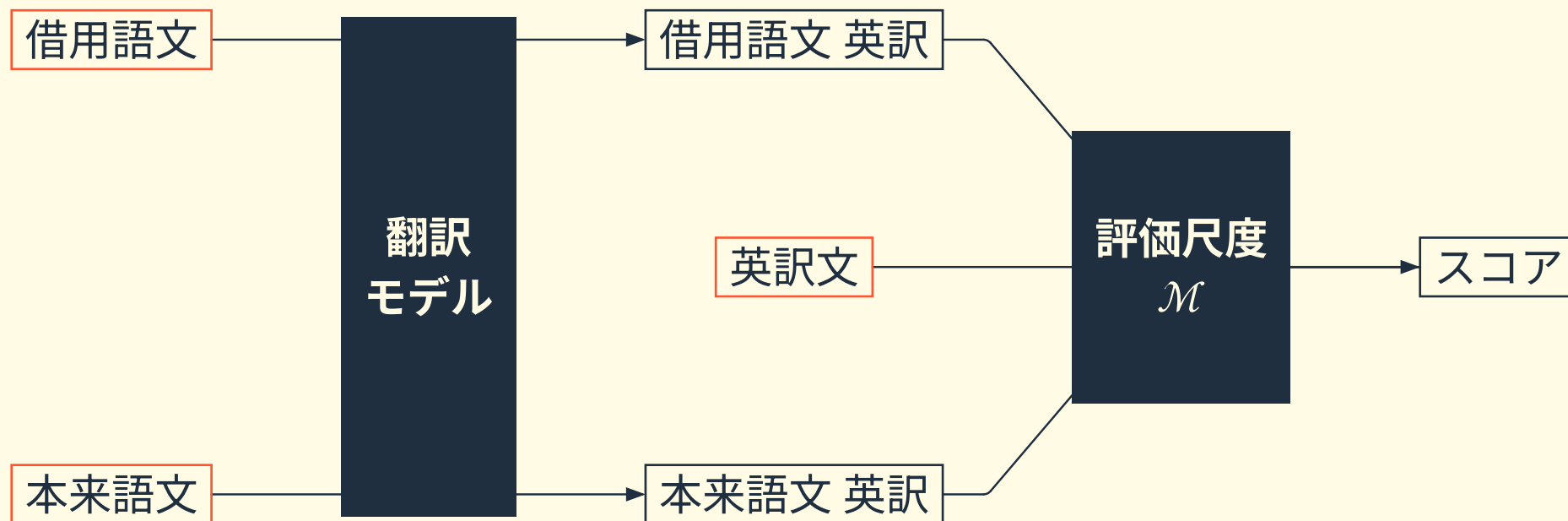
- モデルの中心言語が英語であるのに対し、英語の借用語が多いため
- 借用語が慣例的な文脈に登場しやすいため
- 借用語には学術用語が多く、訓練データに多く含まれているため



各言語の置換前後の Surprisal 平均
(Llama-3.1-8B 8bit-quantized)

翻訳モデルの評価

借用語の文と本来語の文が同じ英訳に翻訳されるか？



翻訳モデルの評価

借用語の文と本来語の文が同じ英訳に翻訳されるか？

モデル

nllb-200-distilled-600M ([NLLB Nature'24](#))

評価指標

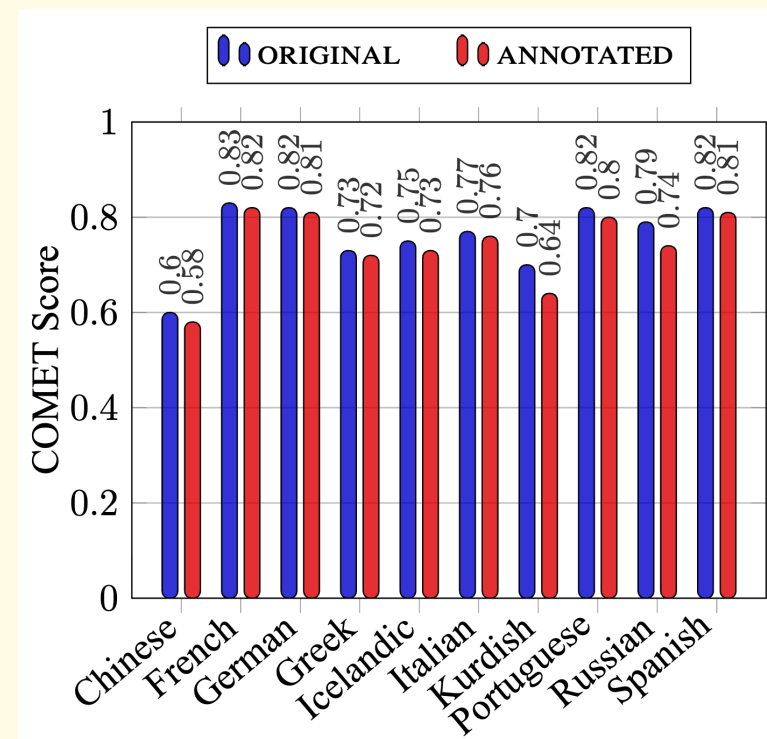
- COMET ([Rei+ EMNLP'20](#)): 翻訳文の意味的な類似度を評価
- 参照訳無しの評価も実施したが、結論は同様

翻訳モデルの評価

借用語の文と本来語の文が同じ英訳に翻訳されるか？

結果

- 一貫して借用語の文のほうがより忠実に翻訳される
 - サプライザルの結果と同様の傾向
 - 多言語事例の学習が借用語を選好を促す可能性
- 中国語が低いのは、
北京語と広東語が混じっているから



置換前後の COMET 平均

まとめと疑問

- 多言語の借用語データセット ConLoan を構築
- 言語モデル・翻訳モデルの選好を分析
 - ▶ 借用語を含む文を好む傾向
 - ▶ 多言語学習の影響か？
- 自然に選択された語を置き換えているので、性能は下がるの自然では？

Portuguese for “Indeed, this is one advantage of the present system.”

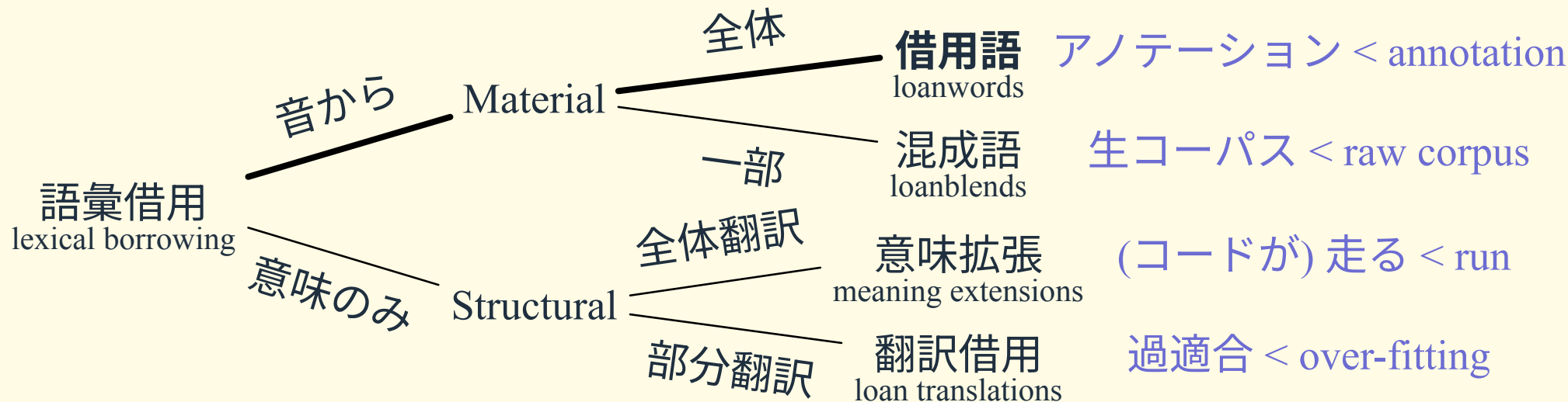
Source

É, na verdade,
uma **vantagem** do sistema actual.

Target

É, na verdade,
uma **benefício** do sistema actual.

補遺



Haspelmath 2009 による語彙借用の分類

アノテーションガイドライン

特に借用語特定に際して、アノテータには借用語の特徴が提供される

- **外来語**

あきらかに対象言語に馴染んでいない (特に英語からの借用語)

Anglicism

- **形態的特徴**

接辞などが借用語に特有のものである

- **音韻的特徴**

その言語に特有の音韻規則に従わない

借用語と本来語の頻度

どの言語でも置換後の本来語のほうが“よく見る”語である傾向

