NAIST同時通訳コーパスの構築：翻訳字幕との比較と通訳経験年数に基づく分析

土肥 康輔，須藤 克仁，中村 哲
奈良先端科学技術大学院大学

NAIST

本研究の概要

- NAIST同時通訳コーパス：大規模な英↔日同時通訳コーパス
  - 300時間以上を収録
  - 経験年数が異なる3名の同時通訳者が同一の原文を訳出したデータを含む [Shimizu+ 2014]
- 訳出遅延，品質，語順，の観点から分析
  - 経験を積んだ通訳者は，遅延時間と品質のバランスをよりよく保てている
  - 遅延時間があまりに長くなると，品質に悪影響が生じている
- コーパスの一部はウェブ上で公開しています

  https://dsc-nlp.naist.jp/data/NAIST-SIC/

本発表の内容

- 背景
- コーパス構築
- コーパス分析
- 結果
  - 訳出遅延
  - 品質
  - 人手評価
  - 語順
- まとめ

背景

- 自動音声翻訳システムの研究が加速
  - システムの構築には，音声翻訳コーパスを使用

- 同時通訳コーパス
  - 実際の同時通訳に基づく
  - 自動音声翻訳システムの開発だけでなく，通訳研究にも有益

【本発表】

NAIST同時通訳コーパスの分析結果について報告

本発表の内容

- 背景
- コーパス構築
- コーパス分析
- 結果
  - 訳出遅延
  - 品質
  - 人手評価
  - 語順
- まとめ

NAIST同時通訳コーパスの概要

- 300時間以上を収録
- * = 通訳比較用データ（ランクが異なる3名の同時通訳者が同一の原文を訳出，4時間×3人）
- その他 = SまたはAランクの通訳者
- データのおよそ半分は人手で書き起こし

	原文	2018	2019	2020
英 → 日	TED	67+12*	50	50
日 → 英	TEDx	12*	40	0
	CSJ	33	0	0
	JNPC	4	36.5	0
総時間		128	126.5	50
累積時間		128	254.5	304.5

経験年数	ランク
15年	S
4年	A
1年	B

原文

- 英 →日
  - TED: 科学や文化などの様々なトピック

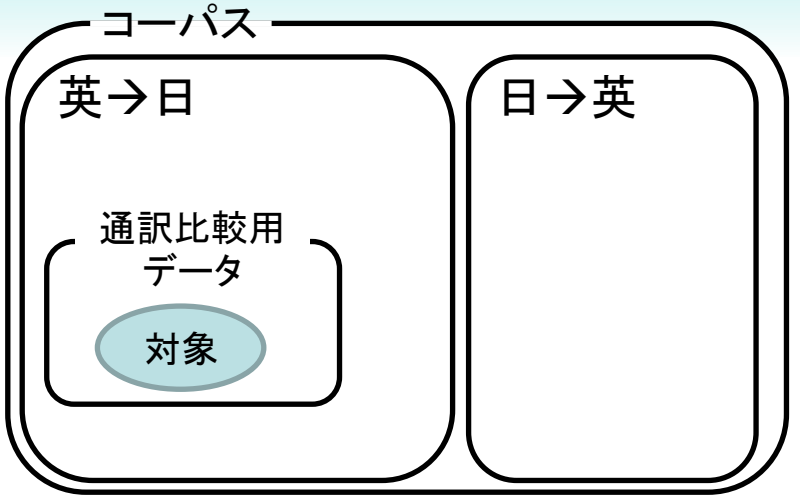
- 日 → 英
  - TEDx: TEDと同じフォーマット
  - CSJ: 講演や日常会話
  - JNPC: 記者会見

本発表の内容

- 背景
- コーパス構築
- コーパス分析
- 結果
  - 訳出遅延
  - 品質
  - 人手評価
  - 語順
- まとめ

データ

- 分析対象：TEDの14本の講演
  - 通訳比較用データの一部




```
EN_0001 13363 17427 Oliver was an extremely dashing,
EN_0002 17427 22248 handsome, charming and largely unstable male
EN_0003 22248 25433 that I completely lost my heart to.
JA_0001 14860 16416 (F えー)オリバーは<H>
JA_0002 17500 21555 (F えー)(F このー)凄くハンサムで魅力的な
JA_0003 22125 24347 (F えー)そして私が
JA_0004 24945 28556 (F えー)(?)大好きな<H>(F えー)男性です。
```

TEDにより配布されている英語字幕

同時通訳

英語字幕 ＋ 同時通訳x3

```
1089 5153  オリバーはとても威勢が良く、
5153 9974  ハンサムで魅力的で、じっとしていることがない。
9974 13159 私が完全に心奪われた男性でした。
```

TEDにより配布されている翻訳字幕

データ

- 同時通訳と翻訳字幕
  - 日本語形態素解析システム Juman++ [Morita+ 2015] と，構文・格・照応解析システム KNP [Kawahara+ 2006] を用いて文節に分割

- 現時点で公開中のコーパス
  - セグメントごとに原文と同時通訳を対応付け
  - 通訳者間で必ずしも一致していない

→ 本研究のために，文ごとの対応付けを作成（公開データには含まれていない）

文ごとの対応付けルール (1)

- 英語の文を基準に人手で対応付けを作成
  - ピリオド (.), またはピリオド＋二重引用符で終わる (.”)
  - 疑問符 (?), または疑問符＋二重引用符で終わる (?”)
  - 閉じ括弧で終わる

```
EN_0177 469789 471829 I've got two questions for you.
JA_0116 XXXXXX 473315 二つの質問がありますよ。

EN_0178 471829 473469 (Laughter)
JA_0000 XXXXXX XXXXXX __null__

EN_0179 473469 476069 You know what's coming now, right?
JA_0117 474778 476197 質問分かってるんですね。
```

文ごとの対応付けルール (2)

- 複数の文に対応するセグメント (赤字部分)
  - 文の境界で分割
  - 終了・開始時刻は XXXXXX に変更
- 訳出されていない語句 (青字部分)
  - タグ等の付与なし

JA_0116 466888 473315
(F え)こちらの方が匿名でやってるので誰も見られていない、そしてお金が入ってくる訳{です||で}。二つの質問がありますよ。

```
EN_0174 464229 465509 It's anonymous,
EN_0175 465509 467429 nobody's watching what you're doing,
EN_0176 467429 469789 and you get paid that money anonymously.
JA_0116 466888 XXXXXX (F え)こちらの方が匿名でやってるので誰も見られて
                       いない、そしてお金が入ってくる訳{です||で}。

EN_0177 469789 471829 I've got two questions for you.
JA_0116 XXXXXX 473315 二つの質問がありますよ。
```

文ごとの対応付けルール (3)

- 訳出されていない文: __drop__
- 意図的に訳出されていない(と思われる)文: __skip__

  （例）Thank you.

```
EN_0135 354509 357749 It's just wrong to lie, for example.
JA_0079 359795 362350 例えば、嘘をつくのは悪いと言う事です。

EN_0136 357749 360909 So, meet my friend Immanuel here.
JA_0000 XXXXXX XXXXXX __drop__

EN_0137 360909 363749 He knows that the sausage is very tasty,
EN_0138 363749 366229 but he's going to turn away because he's a good dog.
JA_0080 362690 367930 こちらの犬ですが、ソーセージがとてもおいしいと分かっているけれども、いい犬なので、
JA_0081 368775 370830 あそこに飛び上がろうとはしません。
```

文ごとの対応付けルール (4)

- 訳出の必要がない文: __null__
- 対応する英語がないセグメント: __null__

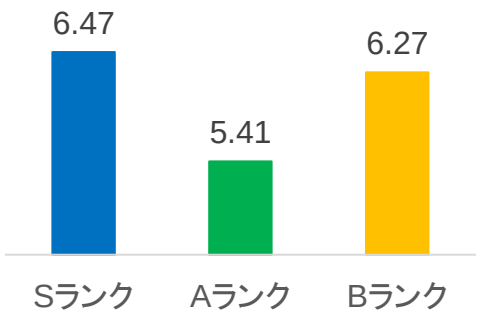
```
EN_0019 71325 72688 (Laughter)
JA_0000 XXXXX XXXXX __null__

EN_0000 XXXXX XXXXX __null__
JA_0024 72267 72882 突然

EN_0020 72688 74839 It's a curious shape for a normal condition.
JA_0025 73687 74417 これは面白い
JA_0026 74820 77850 形ですね、{||こどげ}普通の状況と考えるには。
```

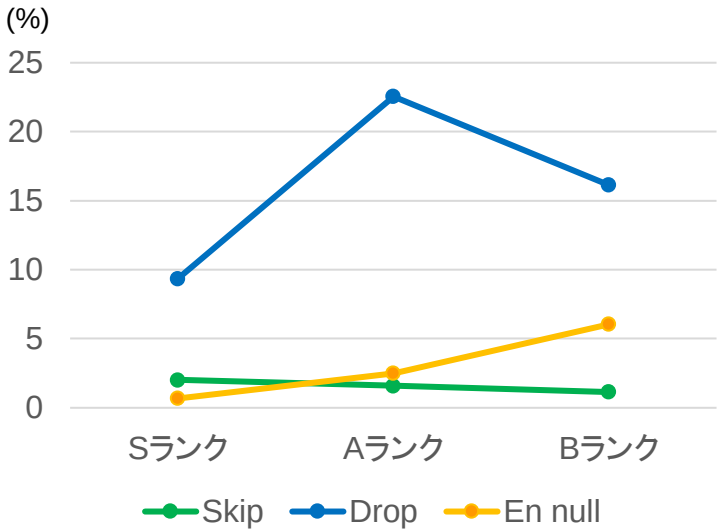
本発表の内容

- 背景
- コーパス構築
- コーパス分析
- 結果
  - 訳出遅延
  - 品質
  - 人手評価
  - 語順
- まとめ

全体傾向

- 高ランクの通訳者のほうが品質が高い
  - 1文あたりの通訳の長さ: S = B > A ($p < 0.001$)
  - 英語の「null」: B > A > S
    → Bランク通訳者は，原文にない内容を付け加えている可能性
  - Drop: A > B > S
    → Sランク通訳者は訳抜けが少ない
  - Skip: 有意差なし





訳出遅延：指標

- Ear-Voice Span (EVS)
  - 原発話と対応する同時通訳のラグ

```
EN_0006 31877 33347 But I like to say,
EN_0007 33347 36951 okay, let's look at the modern human condition.
JA_0005 32643 33450 でも私は、
JA_0006 35761 36901 (F え)(F このー)
JA_0007 37951 39960 現代の人間の状態と言うのを見てみましょう。
```

EVS_{start}: 開始時刻の差

$32643 - 31877 = 766$

EVS_{end}: 終了時刻の差

$39960 - 36951 = 3009$

  - 以下のような場合は分析から除外

```
EN_0009 38643 41045 This is the normal way for things to be.
JA_0012 XXXXX 42236 これが
JA_0013 42465 44085 極普通の状況です。
```

開始・終了時刻のデータがない

```
EN_0033 115144 118838 And Ed Witten unleashed the second
                      superstring revolution.
JA_0044 117516 118421 エドウイットの
```

EVS_{end} が負の値
(= 途中で訳出を断念した場合)

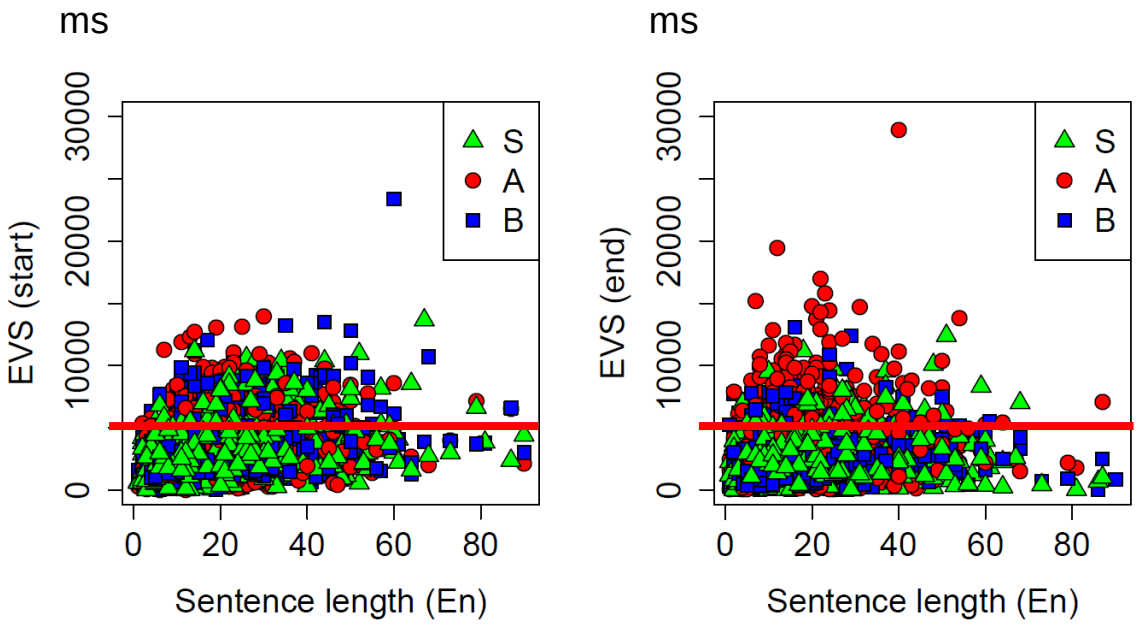
$118421 - 118838 = -417$

訳出遅延：結果

- EVS_{start} ， EVS_{end} の傾向
  - A > B > S
  - 遅延は2～4秒程度
    - 先行研究の結果と一致

通訳者ランク	EVS_{start}	EVS_{end}
S	2.95	2.48
A	**3.57**	**3.89**
B	3.46	2.79

- EVS_{start} ， EVS_{end}が大きい場合がある（5秒以上）
  - 原発話の文長は EVSに影響しない
  - 相関：$r = 0.2584, 0.1206$
    - [Lee 2002] の結果と一致しない



EVSが大きくなる原因の考察

- EVS_{start}
  - 文の前半が訳出されていない

    (En) A week later, Ping was discovered in the apartment alongside the body of her owner, and the vacuum had been running the entire time.

    (A-rank) そしてずっと掃除機がオンになったまま残されていたんですけれども

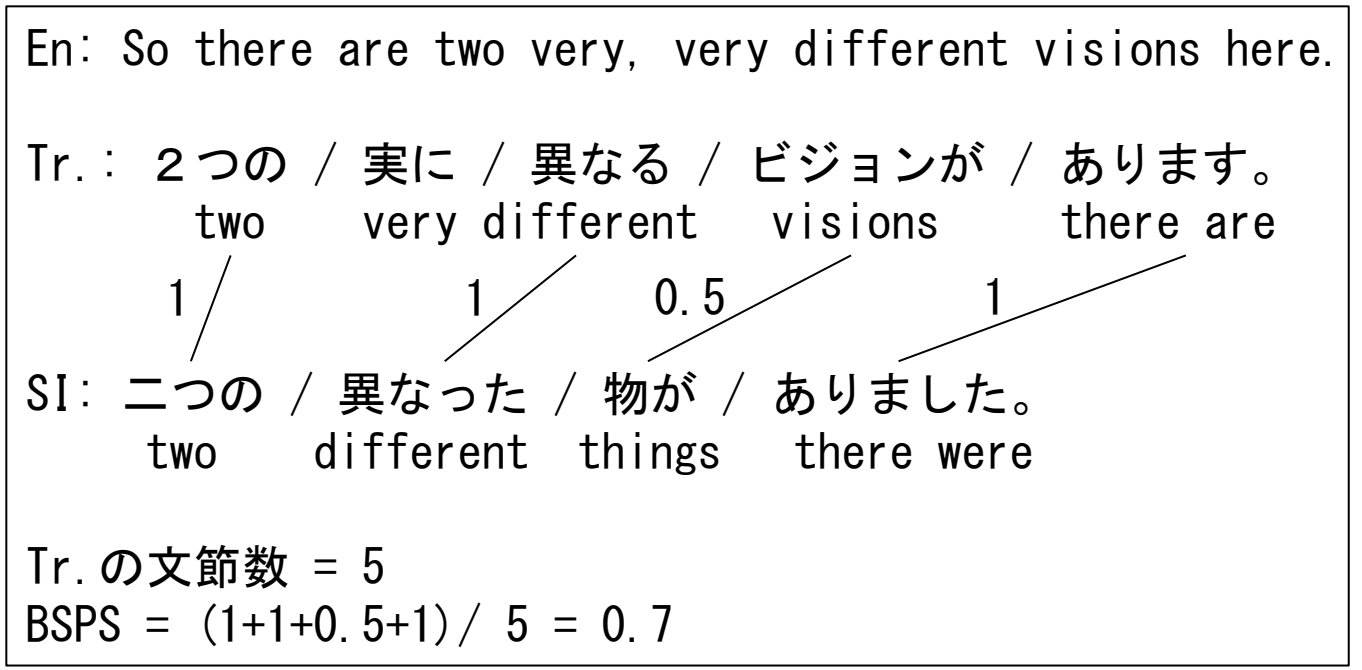
    [And the vacuum had been running the entire time.]

- EVS_{end}
  - 次の文が始まっているにも関わらず，現在の文の訳出を続ける
  - （Aランク）EVS_{end}が大きくなっている上位10%
    - 56.68% で次の文が __drop__

  → EVS_{end} 大きくなると，品質に悪影響が生じる

品質：指標

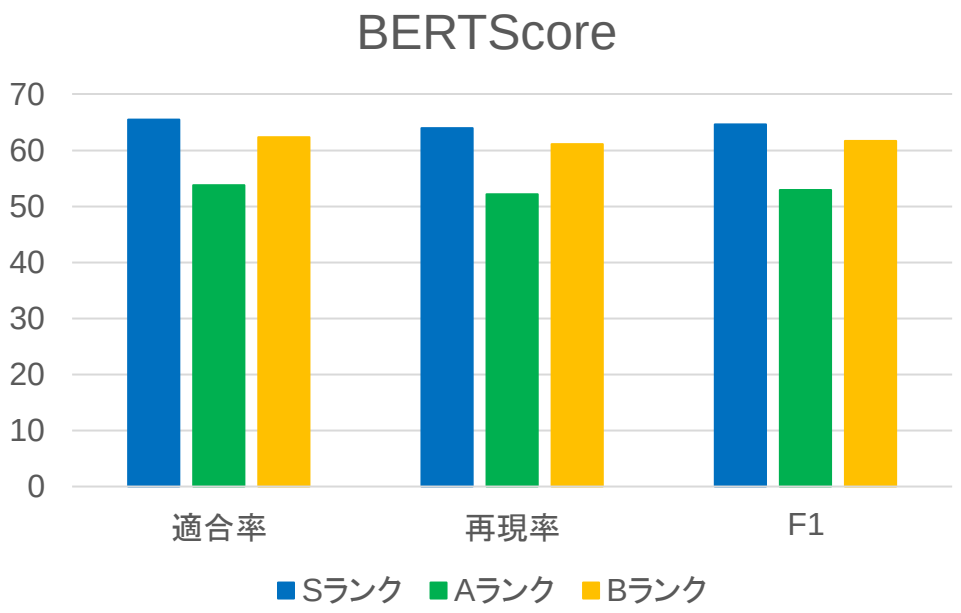
- BERTScore [Zhang+ 2019]
  - 文脈を考慮して，文をベクトル化
  - 意味の類似度を数値化できると言われている
- 文節レベルでの意味保持スコア (BSPS) [稲生ら 2008]
  - 同時通訳の「忠実さ」を翻訳字幕と比較
    - 人手で文節ごとの対応付けを作成
    - 3つの講演で算出



品質：BERTScore

- 適合率：同時通訳の語が，翻訳字幕の語とどの程度一致するか
- 再現率：翻訳字幕の語が，同時通訳にどの程度出現するか
- F1スコア：適合率と再現率の調和平均

- 適合率 > 再現率
  - 「要約」や「一般化」などのストラテジーの使用
- S > B > A (p<0.05)
  - Aランクでは「drop」の割合が高い
- 人手評価との相関は0.3程度
  - p.27 で詳細を提示



品質：BERTScore

- BERTScoreで適切に評価できている例: F1 = 83.25

  (En) We did this experiment for real.
  (Ref) 実際にこの実験を行ってみました。
  (A-rank) これを実際にしました。 [Did this for real.]

- BERTScoreでの評価に問題がある例: F1 = 55.19

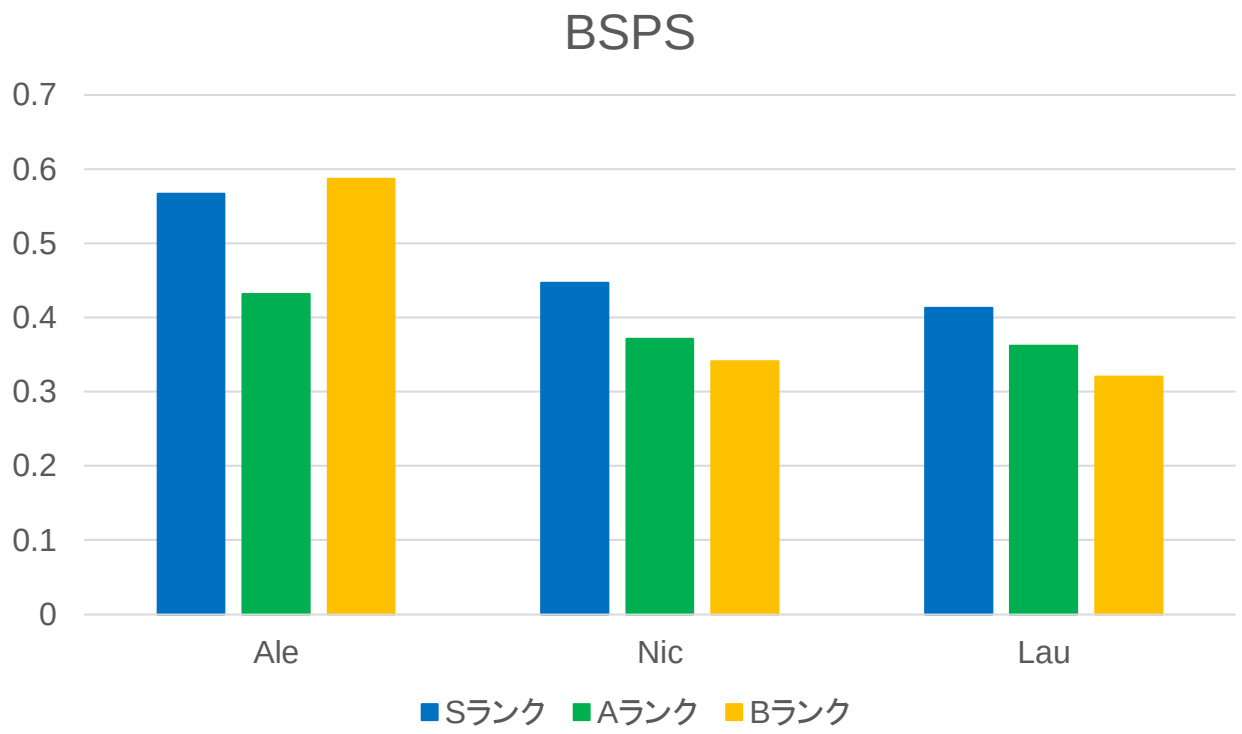
  (En) We can all think of some examples, right?
  (Ref) 例を挙げる事ができると思います。
  (S-rank) 例えば、 [For example,]

品質：BSPS

- 訳出難易度の異なる，以下の3つの講演で算出
  - Ale (易), Nic (中), Lau (難)

- S > A > B (Aleを除く)
  - Ale
    - 「drop」，英語の「null」が
      Bランク通訳者で少ない
    - 人手評価結果と一致

- 人手評価との相関は0.45程度
  - P. 27で詳細を提示



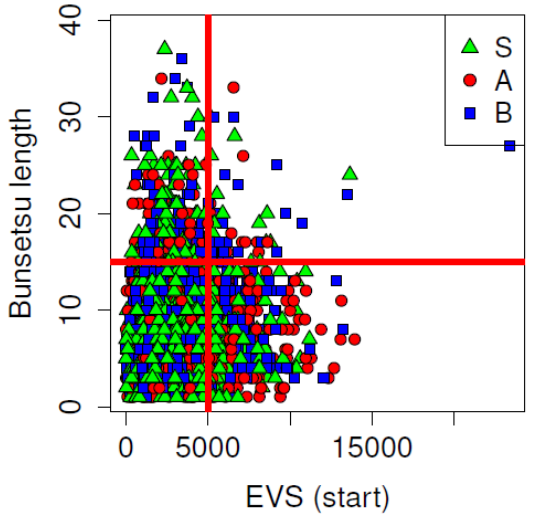
訳出遅延と品質の関係

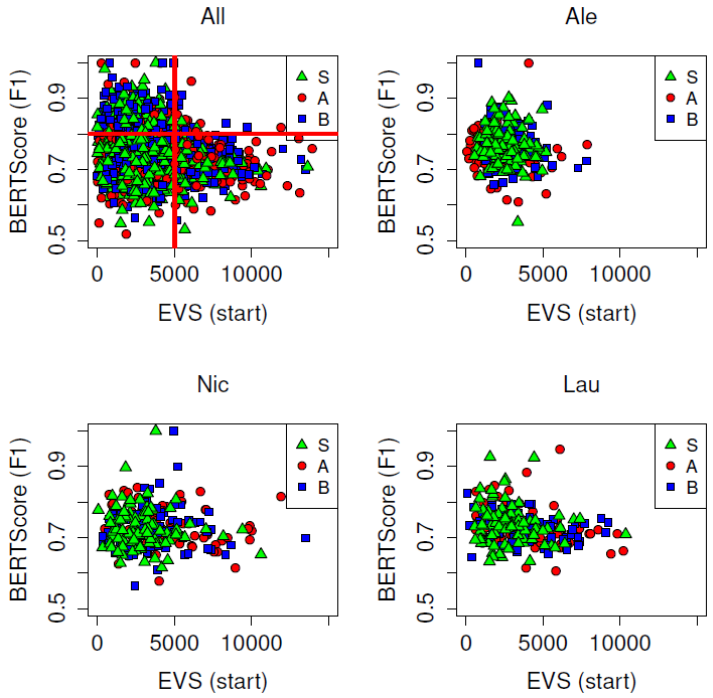
- EVS_{start}が大きくなると(5秒以上)，訳出中の文の品質に悪影響
  - 15文節以上の同時通訳の数は少ない
  - BERTScore，およびBSPSが低い
    - EVS_{start}とBERTScore(F1)，およびBSPSとの相関は，それぞれ-0.15，-0.30程度

- EVS_{start}の標準偏差

Ale	Nic	Lau
1.33	2.25	2.16

Sランク	Aランク	Bランク
1.06	1.68	2.16

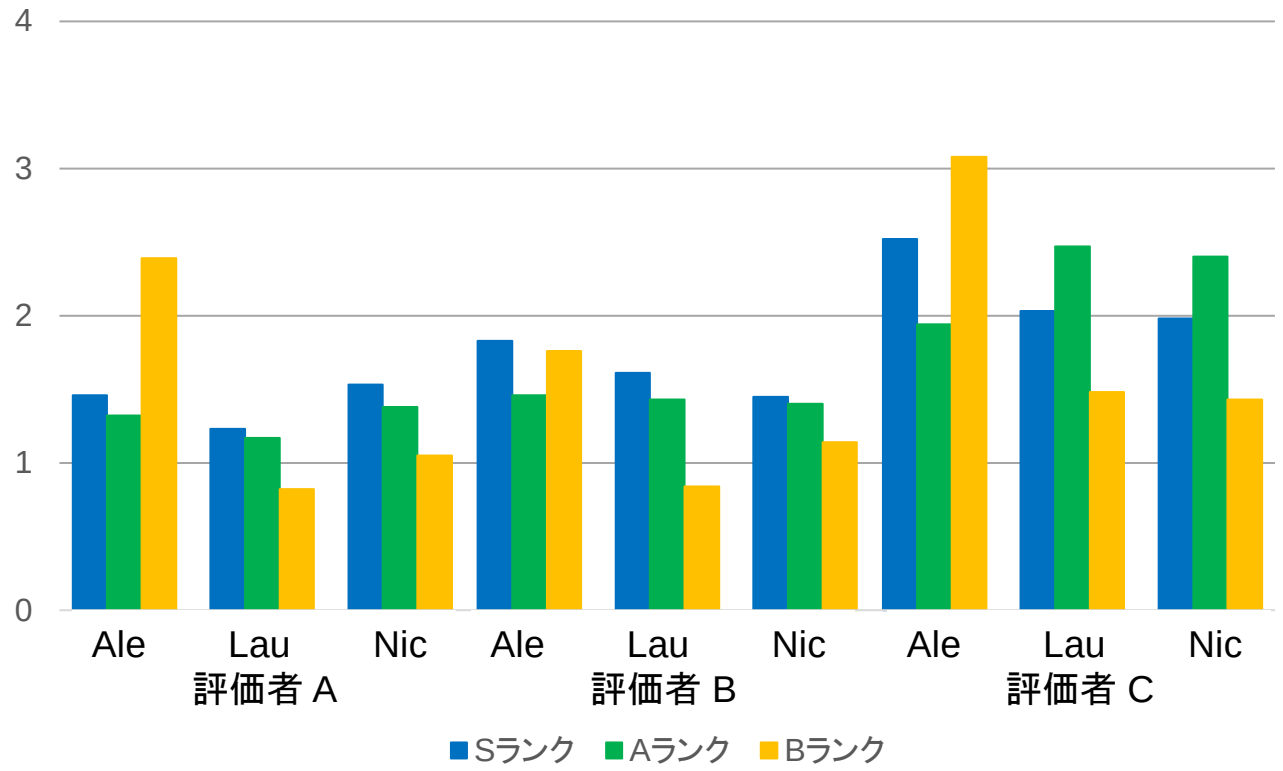




人手評価

- 3名のプロの翻訳者(≠通訳者)による主観評価
  - 「忠実さ」のみ評価 (= デリバリや文法性などは評価しない)
  - 1 (incomprehensible), 2 (poor), 3 (minor errors), 4 (acceptable)
- 高ランクの通訳者が高い評価を受ける傾向
  - 例外: AleのBランク
  - 個人差の影響

    （例）背景知識
- スコアが全体的に低い



人手評価のスコアが低かった原因の考察

- 翻訳者は原文での文構造に厳格

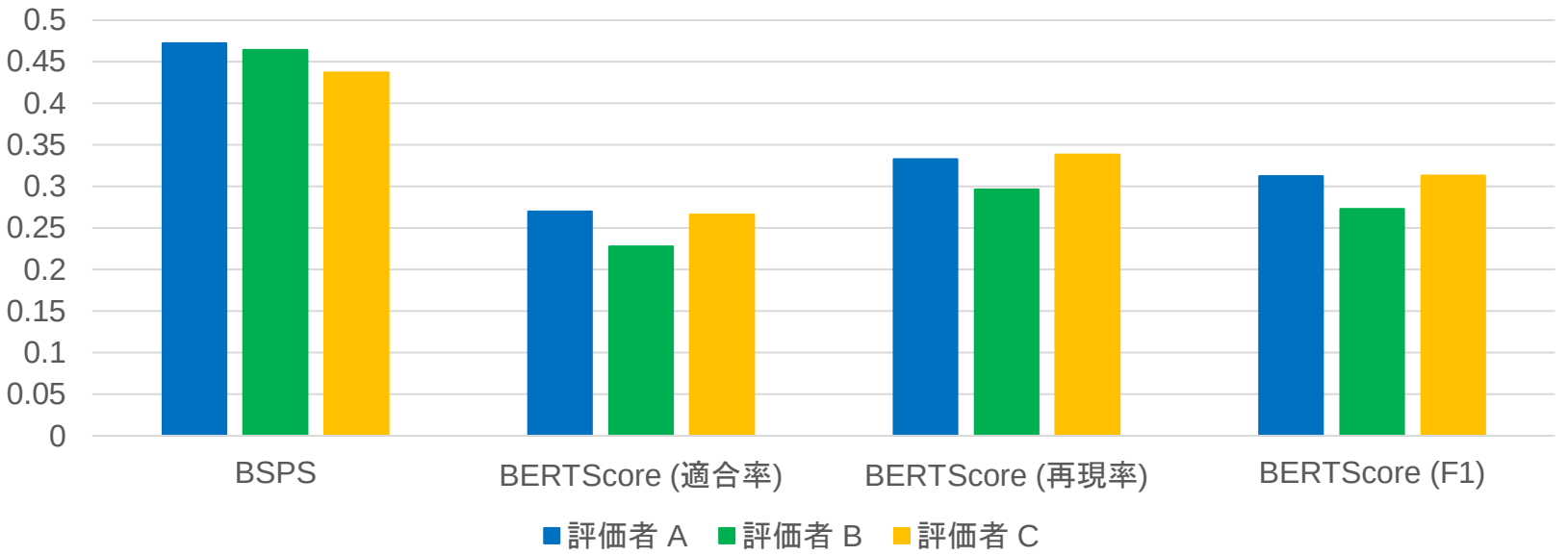
(En) People are motivated by the different values perhaps.
(A-rank) 人のモチベーションは／違う物によって／起こってきます
(Scores) 1, 3, 2

→ 今後の課題: 同時通訳者による人手評価

人手評価

- 人手評価と品質指標との相関
  - (全体傾向) BSPS > BERTScore

  ⇔ 発話ごと
    - Nic_S: BSPS (≈ 0.3) < BERTScore (≈ 0.45)



語順 : 指標

- ケンドール距離（K）[Kendall 1938]
  - 0~1の値
  - 0: 2つのリストの順序が完全一致 （「a b c d」「a b c d」）
  - 1: 一方のリストがもう一方のリストの逆順 （「a b c d」「d c b a」）

$$d_K(\pi, \sigma) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij}}{n(n - 1/2)}$$

$$\text{where } z_{ij} = \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ and } \sigma(i) > \sigma(j) \\ 0 & \text{otherwise} \end{cases}$$

語順

- 通訳者のランクによる明確な違いは見つからず
- 例: K=0.75

En	That’s a huge problem if you think about, especially, an economy like Switzerland, which relies so much on the trust put into its financial industry.
Ref	金融業界の／信用に／大きく依存する／スイスのような／経済を／考えると／これは巨大な問題です。 [put into financial industry / the trust / which relies so much on / like Switzerland / an economy / if you think about / that’s a huge problem]
B-rank	これは、大きな問題です。／特に、／スイスの様な／経済を／考えてみると／そうでしょう。／金融業界に対する／信頼／によって成り立っている／国だからです。 [that’s a huge problem / especially / like Switzerland / an economy / if you think about / it’s true / on its financial industry / the trust / based on / it’s a country]

結論

- NAIST同時通訳コーパス：大規模な英↔日同時通訳コーパス
  - 通訳比較用データを含む

- 同時通訳と翻訳字幕を，通訳者のランクごとに比較
  - 経験を積んだ通訳者は，遅延時間と品質のバランスをよりよく保てている
  - 遅延時間があまりに長くなると，品質に悪影響が生じている

- 公開データ

  https://dsc-nlp.naist.jp/data/NAIST-SIC/