会話とスピーチの
映像による日英対訳コーパス構築
― 自律学習を促す適応学年レベル のあるコンコーダンス ―

Chrome
(recommended)

Corpora　http://www.mintap.com/talkies/?corpora

ミント音声教育研究所　田淵　龍二
tabuchiryuji@nifty.ne.jp

言語処理学会第 24 回年年次大会(NLP2018)
2018 年 3 月 13 日（火）15：10～

英語表現の運用を仮想現実に学ぶ

日本英語教育の歴史的課題

1. 運用する場がない・・・あるのは試験だけ
2. 実社会で使わない・・・5%もいる？
3. 朝令暮改な国家政策に振り回される

+近未来 4. 日常会話なら自動翻訳で事足りる

課題： しかし・・・今 この瞬間にどう対応する？
方策： 動画場面で仮想体験を積む
手段： 対訳コーパス Corpora
事例： How are you? と How do you do? の使い分け

発表の流れ

1. 研究の動機と目的
   問い＞ これまで主に研究者向けだったコーパスを
   授業や学習に使えるようにするには何が必要か？
   答え＞ 意味理解を補助する日英対訳字幕と音映像
   ＋適応学年とドリル
2. 対訳コーパス構築 Corpora http://www.mintap.com/talkies/?corpora
   Corpora＝Seleaf＋TED
   ＝会話（dialogue）＋講演（speech）
3. 自律学習を促す学習用コーパスの諸要件
   適応学年、その理論と実践

1. 研究の動機と目的（1）

対象： 日本における英語学習
自律学習： 各人の問題意識を自ら解決する意思と手段と
　　方法による学び方

これまでは、主に研究者向けコーパスだった その事例

BNC　The British National Corpus（1994）
http://corpus.byu.edu/bnc/
収録語 100M
英語文献を広範囲に集めた最初の大規模コーパス

1. 研究の動機と目的 (2)　so ~ as 検索

1. お題：　so ~ as について BNC で調べる

		CONTEXT	FREQ
1	☐	SO FAR AS	2369
2	☐	SO LONG AS	1297
3	☐	SO MUCH AS	549
4	☐	SO , AS	206
5	☐	SO THAT AS	58
6	☐	SO GOOD AS	43
7	☐	SO . AS	35

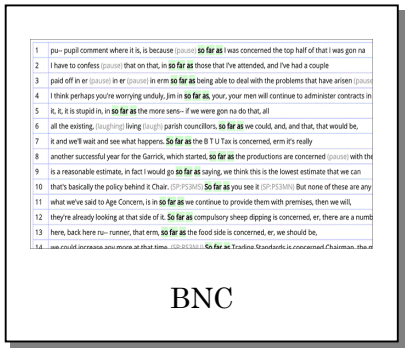
2. ヒット数最大の so far as をひらく

1	pu-- pupil comment where it is, is because (pause) **so far as** I was concerned the top half of that I was gon na
2	I have to confess (pause) that on that, in **so far as** those that I've attended, and I've had a couple
3	paid off in er (pause) in er (pause) in erm **so far as** being able to deal with the problems that have arisen (pause
4	I think perhaps you're worrying unduly, Jim in **so far as**, your, your men will continue to administer contracts
5	it, it, it is stupid in, in **so far as** the more sens-- if we were gon na do that, all
6	all the existing, (laughing) living (laugh) parish councillors, **so far as** we could, and, and that, that would be,
7	it and we'll wait and see what happens. **So far as** the B T U Tax is concerned, erm it's really
8	another successful year for the Garrick, which started, **so far as** the productions are concerned (pause) with the
9	is a reasonable estimate, in fact I would go **so far as** saying, we think this is the lowest estimate that we can
10	that's basically the policy behind it Chair. (SP:PS3MS) **So far as** you see it (SP:PS3MN) But none of these are any
11	what we've said to Age Concern, is in **so far as** we continue to provide them with premises, then we will,
12	they're already looking at that side of it. **So far as** compulsory sheep dipping is concerned, er, there are a numb
13	here, back here ru-- runner, that erm, **so far as** the food side is concerned, er, we should be,
14	we could increase any more at that time. (SP:PS3NU) **So far as** Trading Standards is concerned Chairman, the m

学習者には扱いにくい
　理由
　1. 意味理解が困難
　2. 読解した結果を保証できない

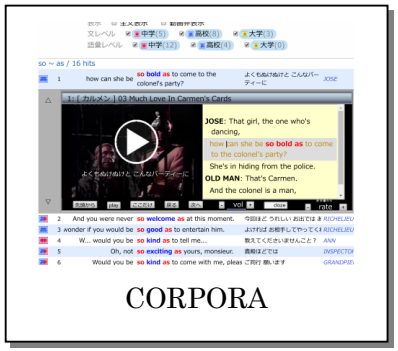
1. 研究の動機と目的 (3)

語学学習用コーパスの諸要件



日本の英語学習者が コーパスを
授業や学習で使えるようにするには 何が必要か？

答え＞ **意味理解**を補助する字幕と音映像
＋適応学年
＋ドリル



解題＞

字幕	学習の敷居を下げる意味理解
音映像	字幕（文字化）による欠落情報の復活
適応学年	学習者の学力に合った情報提供
ドリル	反復練習による定着促進

2. 対訳コーパス構築 Corpora (1)

http://www.mintap.com/talkies/?corpora

Corpora＝Seleaf（会話 dialogue）＋TED（講演 speech）

コーパス	Seleaf	TED
言語資源	名作映画	講演会
作品数	24 本	2,000 本
収録語数	0.3M	5M
検索単位	字幕	字幕
字幕特性	呼気段落	文字数
訳文体	意訳	直訳
	チャンク訳	文訳

検索単位： 対訳の単位であり、ヒットした結果を表示する時の単位となる。

呼気段落（breath group, BG）： 吸気などにより区切られた音声連続（平均 2 秒程度）で、無音区間で区切られることを特徴とする。

字幕の文字制限： テレビなどの字幕には時数制限があり、TED は 1 字幕 84 字 4 秒までとしている。

意訳と直訳： 原文の語句を忠実に翻訳する直訳に対し、意訳は翻訳語の文化と筋書きを重視する。

チャンク訳と文訳： 字幕（BG）ごとに訳すチャンク訳に対し、文訳は文ごとに訳して字幕に配分する。

2. 対訳コーパス構築 Corpora (2)

TED の文訳と、Seleaf のチャンク訳

TED 対訳コーパスにおける日英字幕のズレとフレーズ訳（右）

#	英語字幕	日本語字幕（文訳）	フレーズ訳
80	And having said to your dad, Nic,	君のお父さんのニックに	お父さんと約束したから
81	that I would try to teach you, I was then slightly confused	君にピアノの演奏を教えると約束しておきながらも	君に教えようと思ったけど困ったことに
82	as to how I might go about that	ピアノに近づくことができないのに	どうしたらいいかわかんないよ
83	if I wasn't allowed near the piano.	どうやって教えるんだと困ってしまいました	ピアノに近寄れないんだから

memo: from TED Talks: *In the key of genius* by Derek Paravicini and Adam Ockelford

2. 対訳コーパス構築 Corpora (3)　so ~ as 検索

1. お題： so ~ as について調べる
2. 結果

手順

1. コーパス選択
2. 検索語 so ~ as 入力
3. 結果一覧
4. 先頭項目のスキット

- 20～30 秒の文脈提示
- 映像音声視聴
- 検索語ハイライト
- 和文対訳並列

2. 対訳コーパス構築 Corpora (4)

対訳コーパス構築の 課題と実装結果

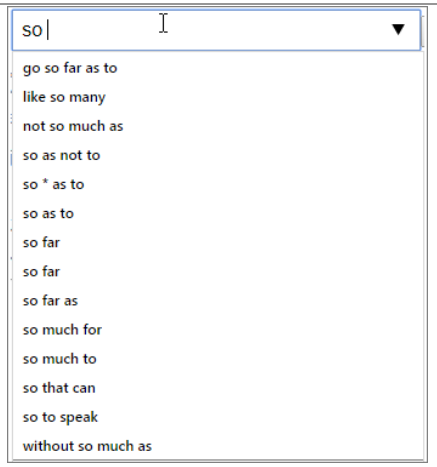
	課題	実装
1	実用性と品質	社会的評価の高い映画と講演
2	対訳による意味理解	英語字幕と日本着字幕の並列
3	会話とスピーチ	映画と講演
4	検索語を見つけやすい	数文字入力で候補を提示
5	文脈単位で視聴可能	日英字幕付きビデオ埋め込む
6	自動生成問題演習	Talkies との連携で各種ドリル
7	自律的学習を促す工夫	文と語彙の中高大レベル分け

2. 対訳コーパス構築 Corpora (5)

入力補助の効用

- うろ覚えの綴りや熟語でも検索可能にする
- イディオムを想起・発見させる

入力欄に "so "（エス・オー・空白）と入力した時に表示される候補の様子

中学のみの場合	高校のみの場合	大学受験のみの場合
		

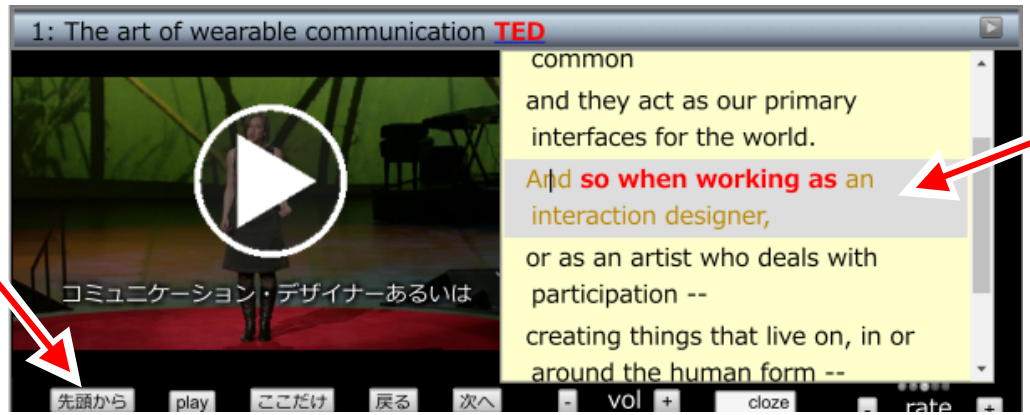
3. 自律学習を促す学習用コーパス（1）

自律学習を促す学習用コーパスの諸要件

1. 字幕＋映像音声で意味理解補助
2. 自動生成問題演習で定着促進
3. 中高大レベル分けで学力にあったテキスト提供

1 文脈単位で視聴可能　日英字幕付きビデオ埋め込む

文脈を通して視聴するには、**先頭から**をクリック（タップ）



ひとつの字幕だけを視聴するには、字幕をクリック（タップ）

3. 自律学習を促す学習用コーパス (2) 自動生成問題演習

2 自動生成問題演習　Talkies との連携で各種ドリル

リンクボタンをクリックしてひらく

聴き取り語順整序

あるいは直アドレスで Corpora on Talkies をひらく

http://www.mintap.com/talkies/talkies.html?corpora

3. 自律学習を促す学習用コーパス (3) 学年レベル分け

3 文と語彙の中高大レベル分け

高校レベルの用例だけ学習したい ＞大学をオフにする

大学レベルの用例
が除かれる

＞ 学年レベル分けの理論は

この１つあと 20 分後に この教室で 詳しく説明

4. 文化（教育研究）と著作権

映画： Seleaf は著作権の保護期間が終了した名作映画

講演： TED Talks は ideas worth spreading（価値ある考えを広めよう）の理念のもと、クリエイティブ・コモンズ・ライセンスを宣言

コーパス： Google などの検索エンジンと同様、公開文書へのアクセスリンク一覧を提供する仕組み

著作権法の目的：第一条 この法律は・・・文化的所産の公正な利用に留意しつつ、著作者等の権利の保護を図り、もつて文化の発展に寄与することを目的とする。

萎縮懸念： 経済利益と文化発展の均衡ある合意が必要

ありがとうございました

アンケートは回収箱へ

ハンドアウト（PDF 版）や TED コーパスなどの記事を配信するメルマガをご希望の方はメールでお知らせください

tabuchiryuji@nifty.ne.jp