

# 言語処理技術と語学教育の接点： ウェブ・コーパスと自律的学習の未来に向けて

文法コーパス SCoRE on Talkies

<http://www.mintap.com/talkies/?score>

TED コーパス selected360 on Talkies

<http://www.mintap.com/talkies/?ted360>

対訳コーパス Corpora on Talkies

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

ミント音声教育研究所 田淵 龍二  
[tabuchiryuji@nifty.ne.jp](mailto:tabuchiryuji@nifty.ne.jp)



# 英語表現の運用を仮想現実学ぶ

## 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

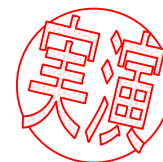
+ 近未来 4. 日常会話なら自動翻訳で事足りる

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 do you do? の定説と現実



# 研究の動機と課題

自然素材\*を語学教材にしたい  
利点

\*自然素材：あらかじめ語学教材として編集されたわけではない資源を自然素材と呼ぶ。

- 種類が豊富で、多様な表現を体験できる
- 生の言語表現に触れることができる

難点

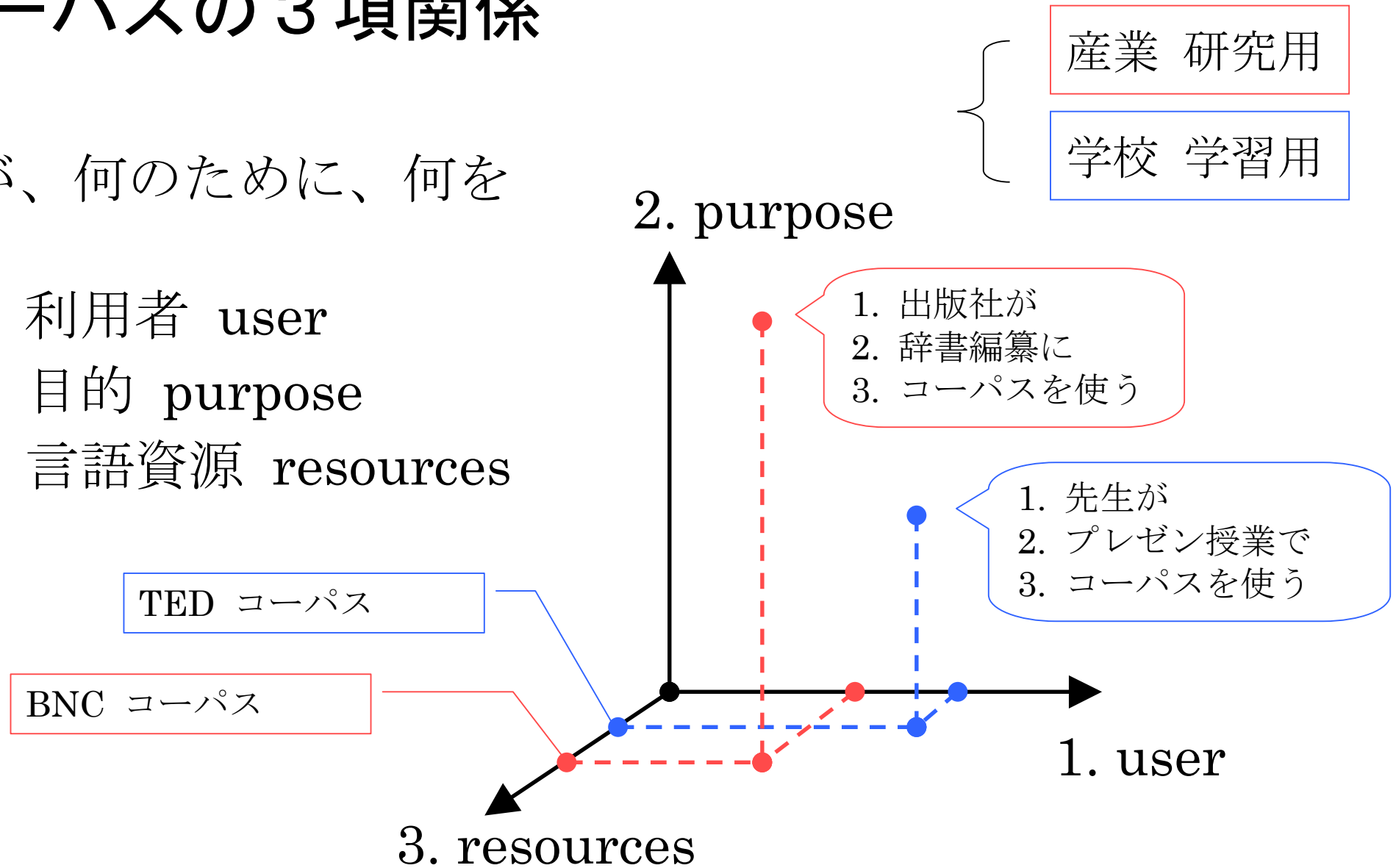
- 学力や目的に見合った素材かどうか不明
- 素材探しに時間がかかる
- 他言語だけだと意味理解に難があり敷居が高い
- 定着に欠かせないドリルがない

課題： 難点を克服したシステムの構築

# コーパスの3項関係

誰が、何のために、何を

1. 利用者 user
2. 目的 purpose
3. 言語資源 resources



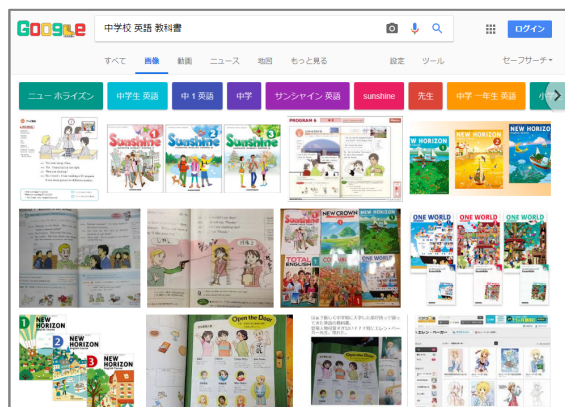
# 語学学習用コーパスの 4 大要件 (1)

おもちゃには対象年齢



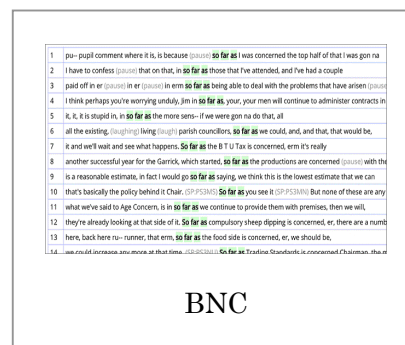
<https://www.kosodate-ryouhin.co.jp/SHOP/134684/list.html>

教科書には適応学年がある



<https://www.google.co.jp/>

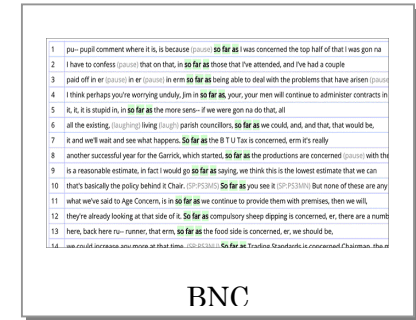
学習用コーパスにも  
適応学年が欲しい！



絵も音も訳もない  
コーパスなんて！！



# 語学学習用コーパスの 4 大要件 (2)



日本の英語学習者が コーパスを 自律的に  
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＞ | 1 字幕   | 学習の敷居を下げる意味理解     |
|     | 2 音映像  | 字幕（文字化）による欠落情報の復活 |
|     | 3 適応学年 | 学力ミスマッチでやる気をそがない  |
|     | 4 ドリル  | 運用と反復練習による定着促進    |

# コーパス事例の紹介 (オープンサイト)

- 1 BNC                      The British National Corpus (1994)  
<http://corpus.byu.edu/bnc/>
- 2 Seleaf                    Audio-Visual Word Corpus of Spoken English  
Through Movies (2011)  
<http://www.mintap.com/>
- 3 TCSE                    TED Corpus Search Engine (2014)  
<https://yohasebe.com/tcse/>
- 4 SCoRE                   文法用例コーパス (2014)  
<http://www.score-corpus.org/concordance/jp/>  
<http://www.mintap.com/talkies/?score>
- 5 selected360            TED Selections Most Viewed 360 (2017)  
<http://www.mintap.com/talkies/?ted360>
- 6 Corpora                Palarell Corpus Search Engine (2018)  
<http://www.mintap.com/talkies/?corpora>

# 会話と講演の表現比べ (1)

お題： 話し言葉でも 会話と講演で表現が違うか？

方法： "so ~ as" を検索

手順

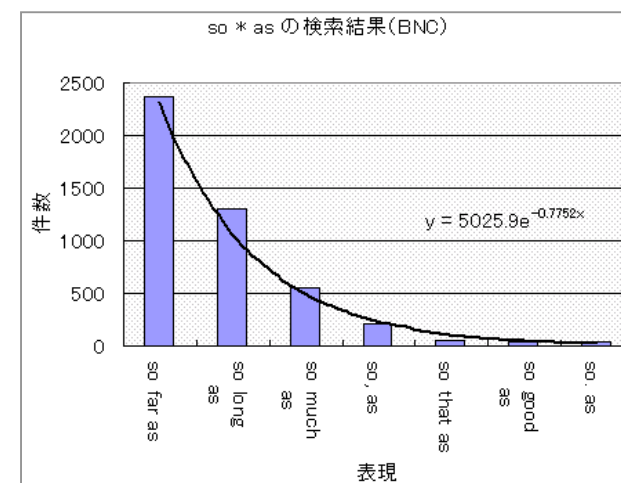
1. BNC で全般的な傾向を観察
2. 上位頻度表現を Corpora で再検索して比較

## 会話と講演の表現比べ (2)

### 手順 1 : BNC で "so ~ as" を検索

結果 : so far as が最大で、以下 指数的に減少

		CONTEXT	FREQ
1	<input type="checkbox"/>	SO FAR AS	2369
2	<input type="checkbox"/>	SO LONG AS	1297
3	<input type="checkbox"/>	SO MUCH AS	549
4	<input type="checkbox"/>	SO , AS	206
5	<input type="checkbox"/>	SO THAT AS	58
6	<input type="checkbox"/>	SO GOOD AS	43
7	<input type="checkbox"/>	SO . AS	35
8	<input type="checkbox"/>	SO KIND AS	25



The British National Corpus (BNC)  
<http://corpus.byu.edu/bnc/>

## 会話と講演の表現比べ (3)

### 手順 2：上位 2 つを Corpora で再検索

結果： 会話と講演では、文書と比べてそれぞれ個性的な振る舞いが見られた。”so far as”は会話では堅いのか？

	BNC	Seleaf	期待値	TED Talks	期待値
収録語数	100M	0.3M		5M	
so far as	2,369	0	7 (0%)	16	118 (14%)
so long as	1,297	4	4 (103%)	17	65 (26%)
so much as	549	1	2 (61%)	15	27 (55%)

note: \* 期待値は BNC での出現率を他のコーパス収録語彙数に適用したもの。丸括弧内は対期待比率。

## 会話と講演の表現比べ (4)

# so long as 検索結果のアプリによる提示

### 1. BNC

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

### 2. Seleaf



### 3. TCSE

#	ID	Line	Time	Text	Translation
1	4755	330	16:16	So long as these two wings are not equivalent in strength.	
2	2811	238	12:14	so long as we know	
3	2598	120	06:47	could never die, so long as they were eating from the tree of life.	
4	1960	75	03:17	so long as there's this fundamental corruption	政府の仕組みが
5	1934	172	07:03	So long as the questions are standardized,	質問が統一されていさえすれば
6	1935	165	07:57	but there will be famines so long as there are bad governments.	政府が改善しなければ 飢饉は続きます
7	1819	276	13:39	so long as you believe that you can handle it, in some sense?	
8	1702	155	07:49	so long as the funders are not the people.	一般市民への依存とは異質で 相対する 60です
9	1649	219	10:32	so long as you believe in yourself.	自分をずっと信じ続けて
10	1638	333	17:10	so long as we're on the gold standard, it should be okay.	「金本位制がある限りは大丈夫。ダメにはならない」と
11	1531	257	11:44	so long as you incentivize them properly	自分に満足を付けたり しないよう裏切りに

### 4. SCoRE

### 5. selected360

### 6. Corpora (Seleaf)

このコーパスで  
何が検索できるのかは  
あとで触れる



# 学年指標のあるその他のコーパス



文法コーパス

SCoRE on Talkies

<http://www.mintap.com/talkies/?score>



TED コーパス

selected360 on Talkies

<http://www.mintap.com/talkies/?ted360>



リーダビリティ公式： $MGJP = 0.07496 \cdot (3 \cdot S + 2 \cdot C) + 7.926 \cdot \text{LOG}(P) + 4.618$

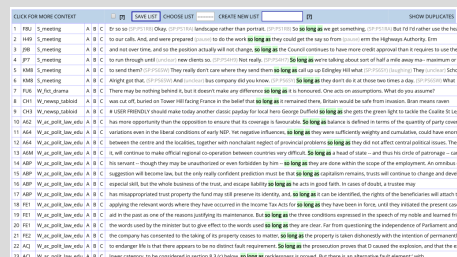
田淵龍二・湯舟英一 (2016). 「音韻符号化の予測時間に基づく日本人英語学習者向けリーダビリティ公式の開発」. Language Education & Technology, 52, 359-388.

語彙レベル公式： $VGL = 30.371 \cdot H + 8.7914$

田淵龍二 (2017). 「日本人英語学習者向け語彙レベル適応学年算出公式の試験的開発」. LET Kanto Journal, 1, 25-35.

# アプリによる検索結果提示法比較

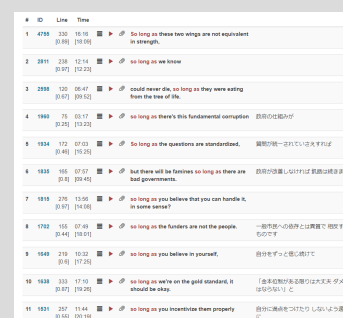
## 1. BNC



## 2. Seleaf



## 3. TCSE



## 6. Corpora



	対訳	音映像	学年	ドリル	文脈	その他
1 BNC	—	—	—	—	単語列	
2 Seleaf	○	○	—	○	シーン	
3 TCSE	○	△ YouTube	△	—	段落	形態素
4 SCoRE	○	○	○	○	用例文	文法項目
5 selected360	○	○	○	○	—	動画検索
6 Corpora	○	○	○	○	シーン	

# 学術研究団体における著作権処理の実態

## 言語処理学会年次大会予稿集掲載論文 23 本のまとめ

type 1: 著作権処理を実施

type 2: 著作権の切れた著作物

type 3: コーパス未公開

type 4: n-gram データを頒布

type 5: 元著作物へのリンクを公開

type 6: (疑似) テキストを生成して公開

type 7: 公開可能な著作物



文献検索サイト： NaCSE で「著作権;コーパス」を検索すると閲覧できる。

<http://www.mintap.com/nacse/nacse.html>

# お知らせ

## 日英対訳コーパスなどの 3 つの研究発表

言語処理学会第 24 回年次大会

日時： 3 月 12～16 日

プログラム： <http://www.anlp.jp/nlp2018/program.html>

会場： 岡山コンベンションセンター

## 言語教育と言語処理の接点

1. 3/13 自律学習を促す日英対訳コーパス構築
2. 3/13 文法コーパスと **TED** コーパス構築
3. 3/14 文脈検索可能な論文コーパス構築

# 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](mailto:tabuchiryuji@nifty.ne.jp)