CEFR-J Wordlist
- その開発手法 -

Yukio Tono (TUFS)

CEFR-J Wordlist ができるまで

PART 1

CEFR-J Wordlist の開発経緯

対象
- 中国・韓国・台湾の英語教科書分析

内容分析&CEFR紐付け
- 各国・地域の学習指導要領と教科書内容分析による CEFR レベル分けを暫定的に行う

コーパス化
- 教科書のテキストをすべて電子化する

語彙表作成
- それぞれの語彙表を作成し、共通語を抽出する

対象とした教科書の学年とCEFRレベル

CEFR レベル	中国	韓国	台湾
PreA1	小３－６	小３－６	小３－６
A1	中１	中１	中１
A2	中２－３	中２－３	中２－３
B1	高１－２	高１－２	高１－２
B2	高３	高３	高３
C1	n/a	n/a	n/a
C2	n/a	n/a	n/a

表３：CEFR レベル準拠コーパスの作成

日本の次期学習指導要領のレベルよりやや高いが、それがアジア圏の教科書内容分析の一致する見解だった

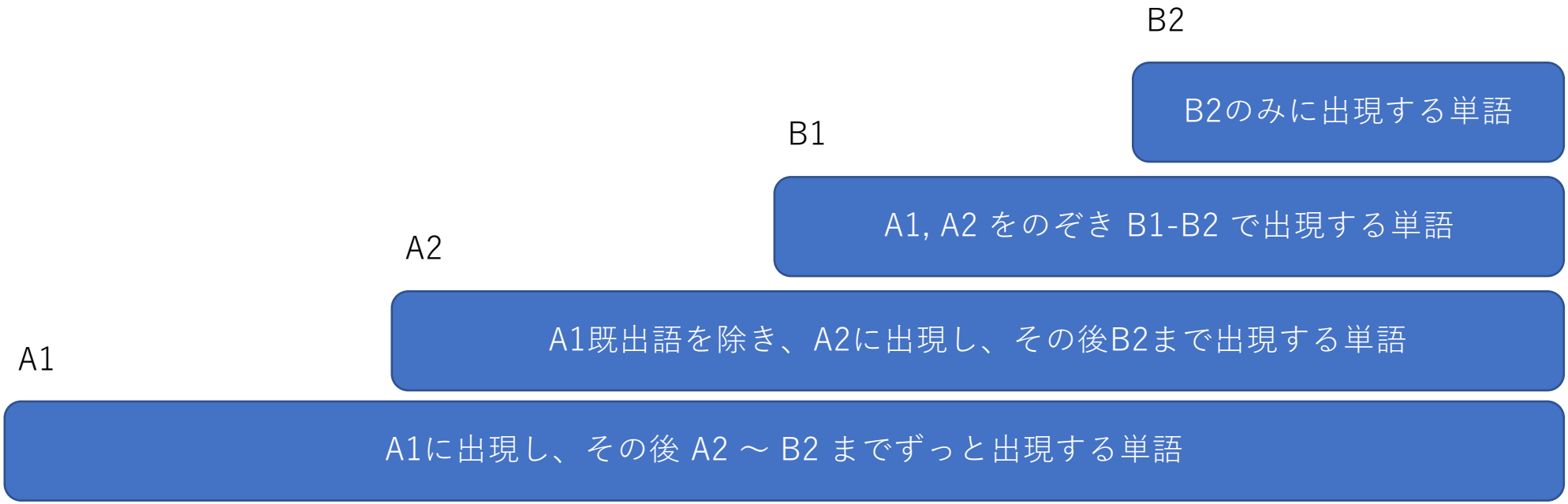
英語教科書コーパス

それぞれの国・地域で最も広範に使用されている教科書を小学校～高等学校まで１セット選定

	小学校用	中学用	高等学校用
韓国		*Middle School English* J.E.Feldt 他（斗山）	*High School English* P. A. O'Neill 他
中国	New Primary English for China（人民教育出版社）	*Go for It!*（人民教育出版社）	Senior English for China Student's Book（人民教育出版社）
台湾	『國小英語 Hello Drabie』（康軒文教事業）	『國中英語』（康軒文教事業）	高級中学『英文』（三民書局）

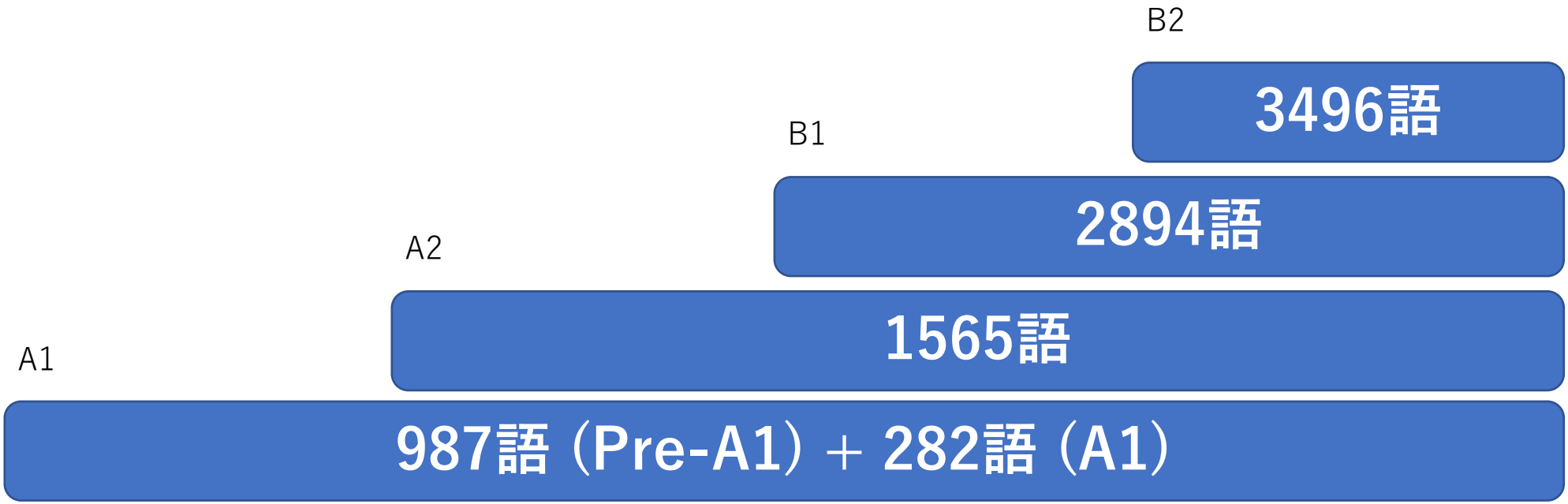
表１：研究調査対象の教科書

語彙選定の方法



最初、Pre-A1, A1 と区切りを入れたが、実質は A1（中１）では ほぼPre-A1 の繰り返しであった

語彙選定の方法（wordform 換算）



CEFR推定語彙サイズ

CEFR レベル	推定語彙サイズ	日本	中国・韓国・台湾
C1/2	8000 語～	社会人？	大学
B2	5500～8000 語	大学	高等学校
B1	3000～5500 語	高等学校/大学	高等学校
A2	1000～3000 語	高等学校	中学校
A1		高等学校	中学校
Pre A1	約 1000 語	中学校	小学校

図２：CEFR 推定語彙サイズと日本およびアジア各国の教科書レベルの比較

投野（2008）

再調整した語彙表サイズ（見出し語換算）

CEFR -Level	Pre-A1	A1	A2	B1	B2	Total
Text analysis	976		1057	1884	1722	5639
Our Target	1000		1000	2000	2000	6000

- 分析結果に準じて、使いやすいように各レベルを1000語単位に調整
- 各レベルで最低限身につけておきたい必修語彙を位置づけ
- B2レベルでA2までの2000語がproductive vocabulary、残りがreceptive vocabularyになるように指導する、というイメージ

Cambridge EVP との比較

CEFR -Level	Pre-A1	A1	A2	B1	B2	Total
Text analysis	976		1057	1884	1722	5639
Target vocabulary size	1000		1000	2000	2000	6000
+ EVP Integrated →Final Version	**1068**		**1358**	**2359**	**2785**	**7570**

A1レベルではほとんど差がなかったが、B2では約1000語ずれていた

見出し語	CEFRレベル	品詞	名詞の分野カテゴリー（Threshold Level）	カテゴリー２（Core Inventory）
activity	A0	n	Leisure activities	
actor	A0	n	Work and Jobs	Film
afternoon	A0	n		
age	A0	n	Personal information	
airplane	A0	n	Ways of travelling	
airport	A0	n	Travel and services vocab	Things in the town, shops and shopping
animal	A0	n		
answer	A0	n		
apple	A0	n	Food and drink	
apron	A0	n	Objects and rooms	
arm	A0	n	Personal information	
art	A0	n	Hobbies and pastimes	
aunt	A0	n	Family life	
baby	A0	n	Family life	
back	A0	n		
bag	A0	n	Shopping	Clo
ball	A0	n	Hobbies and pastimes	
banana	A0	n	Food and drink	
bank	A0	n	Things in the town, shops and shopping	

品詞・CEFRレベルで
フィルタできるだけでなく、
内容語の意味カテゴリーで
抽出が可能

語彙表作成の基礎

WORKSHOP 1

今日は基礎的な語彙表の作成方法を学びます

- サンプル・テキスト：絶版の中学・高校教科書のテキスト1セット
- ツール：MLTP (Multilingual Text Processor)
  - 同志社大学、金明哲先生の作成したソフト
  - 日本語、中国語、韓国語、英語に対応
  - 基礎的な語彙統計を求めるのに便利
  - Java ベースなので、Win/Mac/Linux どれでも動く（はず）
- 今日は品詞タグをつけて英語の語彙表を試作してみます

語彙表の多言語への転換

PART 2

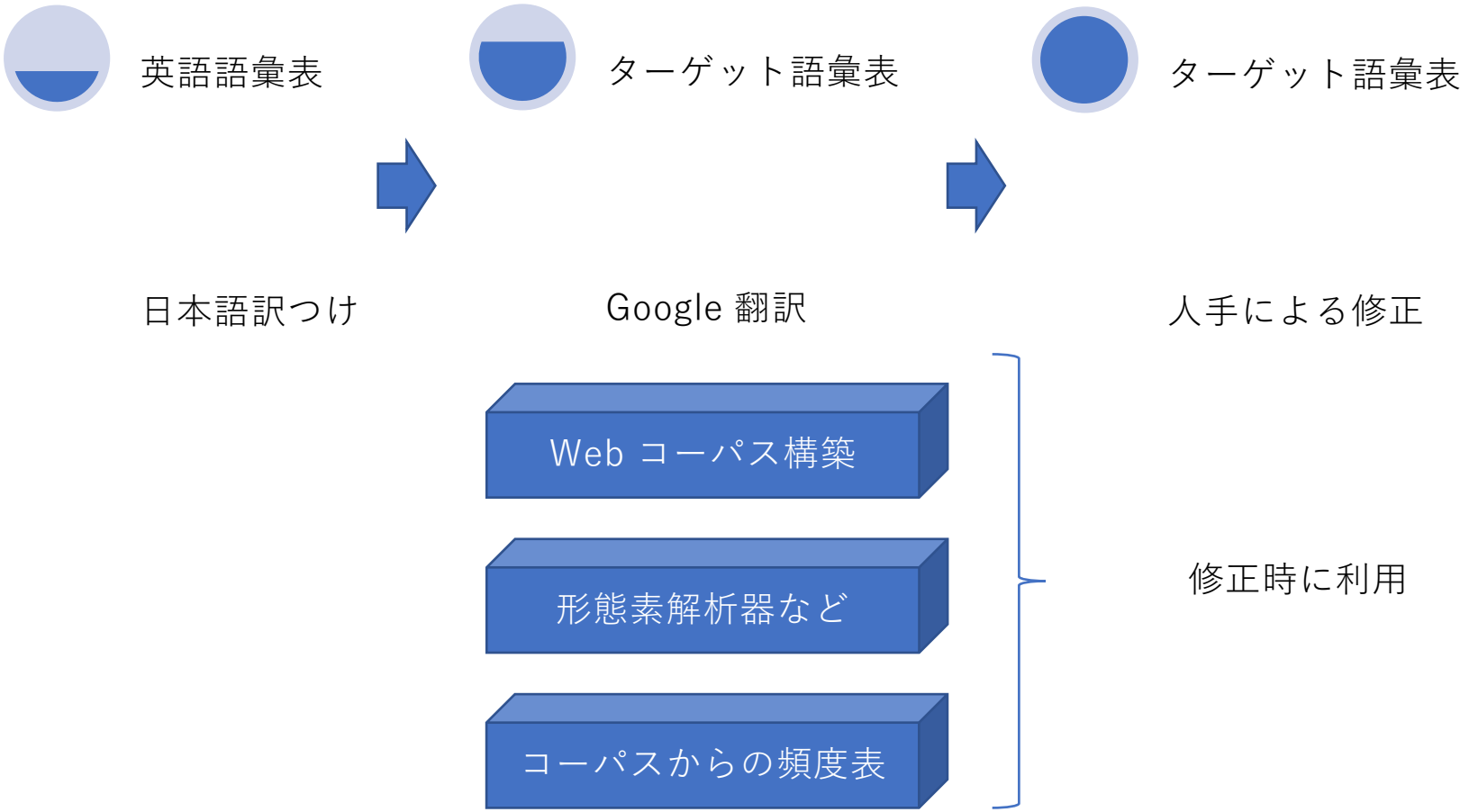
CEFR-J x 27プロジェクト

- Super Global 大学創成支援の補助金を受けて行われているプロジェクト
- 目的：CEFRを用いた多（複）言語教育の評価と可視化
- ワールドランゲージセンターの設立（2017年度）
- 科学研究費（基盤Ａ）の採択（2018年度）

CEFR-J x 27 言語リソースの構築

- CEFR-J x 27 Wordlist：CEFRベースの学習語彙表の整備
- CEFR-J x 27 Phrase List: CEFR(CAN-DO)ベースのフレーズ表整備
- CEFR-J x 27 CAN-DO Task List: CEFRベースのCAN-DOを実現する教室内（教材）タスクの整備
- CEFR-J x 27 CAN-DO Test: CEFRベースのCAN-DOタスクをどのくらいできるかを見る performance test の整備

CEFR-J x 27 語彙表整備プロセス



Google 翻訳


Google
翻訳
100以上の言語に対応
リアルタイ
英語
日本語
韓国語
言語を検出する
日本語
英語
韓国語
翻訳
テキストまたはウェブサイトのアドレスを
GOOGLE 翻訳コミュニティにぜ
言語を検出する
アイスランド語
アイルランド語
アゼルバイジャン語
アフリカーンス語
アムハラ語
アラビア語
アルバニア語
アルメニア語
イタリア語
イディッシュ語
イボ語
インドネシア語
ウェールズ語
ウクライナ語
ウズベク語
ウルドゥ語
エストニア語
エスペラント語
オランダ語
カザフ語
カタルーニャ語
ガリシア語
カンナダ語
ギリシャ語
キルギス語
グジャラト語
クメール語
クルド語
クロアチア語
コーサ語
コルシカ語
サモア語
ジャワ語
ジョージア(グルジア)語
ショナ語
シンド語
シンハラ語
スウェーデン語
ズールー語
スコットランド ゲール語
スペイン語
スロバキア語
スロベニア語
スワヒリ語
スンダ語
セブアノ語
セルビア語
ソト語
ソマリ語
タイ語
タガログ語
タジク語
タミル語
チェコ語
チェワ語
テルグ語
デンマーク語
ドイツ語
トルコ語
ネパール語
ノルウェー語
ハイチ語
ハウサ語
パシュト語
バスク語
ハワイ語
ハンガリー語
パンジャブ語
ヒンディー語
フィンランド語
フランス語
フリジア語
ブルガリア語
ベトナム語
ヘブライ語
ベラルーシ語
ペルシャ語
ベンガル語
ポーランド語
ボスニア語
ポルトガル語
マオリ語
マケドニア語
マラーティー語
マラガシ語
マラヤーラム語
マルタ語
マレー語
ミャンマー語
モンゴル語
モン語
ヨルバ語
ラオ語
ラテン語
ラトビア語
リトアニア語
ルーマニア語
ルクセンブルク語
ロシア語
英語
韓国語
中国語
日本語

Web コーパスの作成

- リソース不足の言語に関しては、web 上にあるテキストを自動取得するツールをもちいて、コーパスを作成。
- Sketch Engine (http://www.sketchengine.co.uk)
- WebBootCat の機能を使用
- ただし、アジア言語の一部は解析リソースが不足しているため、テキストは収集できても、形態素解析技術が利用できないものがある。

形態素解析

- 単語を分かち書きし、活用形・屈折等を辞書形に戻す、等の一連の形態素解析を行うツール

- https://langrid.org/playground/morphological-analyzer.html

- 多言語の形態素解析に関しても世界中で研究されている
- しかし、低資源言語はツールの精度なども低い

語彙表翻訳の基礎

Google 翻訳を使用した例

Google 翻訳を試してみる

- Google ドキュメントを開く（使用したことがなければ登録する）
- スプレッドシートを開く
- セルに好きな日本語の単語を１０個くらい打ってみる
- 隣のセルに以下の関数を書く：

=googletranslate(訳したいセル,"ja","en")

Google 翻訳：言語名一覧（1）

コード	言語名	コード	言語名	コード	言語名	コード	言語名
aa	アファル語（Afar）	bg	ブルガリア語（Bulgarian）	da	デンマーク語（Danish）	fi	フィンランド語（Finnish）
ab	アブハジア語（Abkhazian）	bh	ビハール語（Bihari）	de	ドイツ語（German）	fj	フィジー語（Fiji）
af	アフリカーンス語（Afrikaans）	bi	ビスラマ語（Bislama）	dz	ブータン語（Bhutani）	fo	フェロー語（Faeroese）
am	アムハラ語（Amharic）	bn	ベンガル語（Bengali）	el	ギリシャ語（Greek）	fr	フランス語（French）
ar	アラビア語（Arabic）	bo	チベット語（Tibetan）	en	英語（English）	fy	フリジア語（Frisian）
as	アッサム語（Assamese）	br	ブルターニュ語（Breton）	eo	エスペラント語（Esperanto）	ga	アイルランド語（Irish）
ay	アイマラ語（Aymara）	ca	カタラン語（Catalan）	es	スペイン語（Spanish）	gd	スコットランド・ゲール語（Gaelic [Scottish]）
az	アゼルバイジェン語（Azerbaijani）	co	コルシカ語（Corsican）	et	エストニア語（Estonian）	gl	ガリシア語（Galician）
ba	バシキール語（Bashkir）	cs	チェック語（Czech）	eu	バスク語（Basque）	gn	グワラニ語（Guarani）
be	白ロシア語（Byelorussian）	cy	ウェールズ語（Welsh）	fa	ペルシャ語（Farsi）	gu	グジャラート語（Gujarati）

Google 翻訳：言語名一覧（2）

コード	言語名	コード	言語名	コード	言語名	コード	言語名
gv	マン島ゲール語（Gaelic [Manx]）	ik	イヌピア語（Inupiak）	kn	カンナダ語（Kannada）	lv	ラトビア語（Latvian）
ha	ハウサ語（Hausa）	is	アイスランド語（Icelandic）	ko	韓国語（Korean）	mg	マダガスカル語（Malagasy）
he (iw)	ヘブライ語（Hebrew）	it	イタリア語（Italian）	ks	カシミール語（Kashmiri）	mi	マオリ語（Maori）
hi	ヒンディー語（Hindi）	iu	イヌクティトット語（Inuktitut）	ku	クルド語（Kurdish）	mk	マケドニア語（Macedonian）
hr	クロアチア語（Croatian）	ja	日本語（Japanese）	ky	キルギス語（Kirghiz）	ml	マラヤーラム語（Malayalam）
hu	ハンガリー語（Hungarian）	jv	ジャワ語（Javanese）	la	ラテン語（Latin）	mn	モンゴル語（Mongolian）
hy	アルメニア語（Armenian）	ka	グルジア語（Georgian）	li	リンブルガー語（Limburgish）	mo	モルダビア語（Moldavian）
ia	インターリンガ（Interlingua）	kk	カザフ語（Kazakh）	ln	リンガラ語（Lingala）	mr	マラーティー語（Marathi）
id (in)	インドネシア語（Indonesian）	kl	グリーンランド語（Greenlandic）	lo	ラオス語（Laotian）	ms	マレー語（Malay）
ie	インターリング（Interlingue）	km	カンボジア語（Cambodian）	lt	リトアニア語（Lithuanian）	mt	マルタ語（Maltese）

Google 翻訳：言語名一覧（3）

コード	言語名	コード	言語名	コード	言語名	コード	言語名
my	ビルマ語（Burmese）	ps	パシュト語（Pashto）	sg	サングロ語（Sangro）	ss	シスワティ語（Siswati）
na	ナウル語（Nauru）	pt	ポルトガル語（Portuguese）	sh	セルボ=クロアチア語（Serbo-Croatian）	st	セソト語（Sesotho）
ne	ネパール語（Nepali）	qu	ケチュア語（Quechua）	si	シンハラ語（Sinhalese）	su	スンダン語（Sundanese）
nl	オランダ語（Dutch）	rm	レト=ロマン語（Rhaeto-Romance）	sk	スロバキア語（Slovak）	sv	スウェーデン語（Swedish）
no	ノルウェー語（Norwegian）	rn	キルンディ語（Kirundi）	sl	スロベニア語（Slovenian）	sw	スワヒリ語（Swahili）
oc	オキタン語（Occitan）	ro	ルーマニア語（Romanian）	sm	サモア語（Samoan）	ta	タミール語（Tamil）
om	オロモ語（Oromo）	ru	ロシア語（Russian）	sn	ショナ語（Shona）	te	テルグ語（Telugu）
or	オーリア語（Oriya）	rw	キニャーワンダ語（Kinyarwanda）	so	ソマリ語（Somali）	tg	タジク語（Tajik）
pa	パンジャブ語（Punjabi）	sa	サンスクリット語（Sanskrit）	sq	アルバニア語（Albanian）	th	タイ語（Thai）
pl	ポーランド語（Polish）	sd	シンディー語（Sindhi）	sr	セルビア語（Serbia）	ti	チグリニャ語（Tigrinya）

Google 翻訳：言語名一覧（4）

コード	言語名	コード	言語名	コード	言語名	コード	言語名
tk	トルクメン語（Turkmen）	ur	ウルドゥー語（Urdu）				
tl	タガログ語（Tagalog）	uz	ウズベク語（Uzbek）				
tn	セツワナ語（Setswana）	vi	ベトナム語（Vietnamese）				
to	トンガ語（Tonga）	vo	ボラピュク語（Volapuk）				
tr	トルコ語（Turkish）	wo	ウォロフ語（Wolof）				
ts	ツォンガ語（Tsonga）	xh	コサ語（Xhosa）				
tt	タタール語（Tatar）	yi (ji)	イディッシュ語（Yiddish）				
tw	トウィ語（Twi）	yo	ヨルバ語（Yoruba）				
ug	ウイグル語（Uighur）	zh	中国語（Chinese）				
uk	ウクライナ語（Ukrainian）	zu	ズールー語（Zulu）				

Google 翻訳

- Google スプレッドシートに語彙リストをアップロードする
- 翻訳を記入したいセルに以下の関数を用いる：

  =googletranslate(訳したいセル,"ja","en")

- あとはこれをいろいろな言語に変えるだけ

Active Learning 課題（提出：１１月２７日）

- CEFR-J Wordlist A1レベルの語彙表をもとに：
  - 自分の専攻する言語に Google 翻訳を利用して翻訳してみる
  - 語彙表全体の翻訳の精度を自分なりに分析してみる
    - どういう単語が翻訳が正確に行われているか？
    - どういう単語が不正確か？
    - 不正確な翻訳の原因は何だと思うか？
    - 語彙表の変換を機械翻訳で行う際の利点・欠点は何か？
  - 全員に11/27 に発表、レポート（A4 2-3枚）を提出