CEFR-J準拠教育用webコーパスの開発

投野由紀夫†・川原田将之††・渡辺亮嗣†††・星野守†††・奥村学††

†東京外国語大学　　††東京工業大学　　†††（株）ネットアドバンス

英語コーパス学会第46回大会（JAECS2020）, 2020年10月3-4日（オンライン開催）

小学館コーパスネットワーク

https://scnweb.japanknowledge.com/

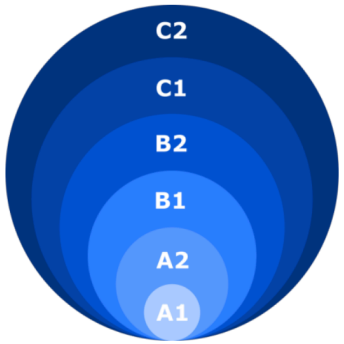
- 2003年、世界初の BNC と WordbanksOnline の共通IFでの検索サービス提供
- PERC：科学技術英語
- JEFLL：日本人学習者作文
- Webコーパス LECS

CEFRレベル付き英語コーパス構築プロジェクト

2001年発表

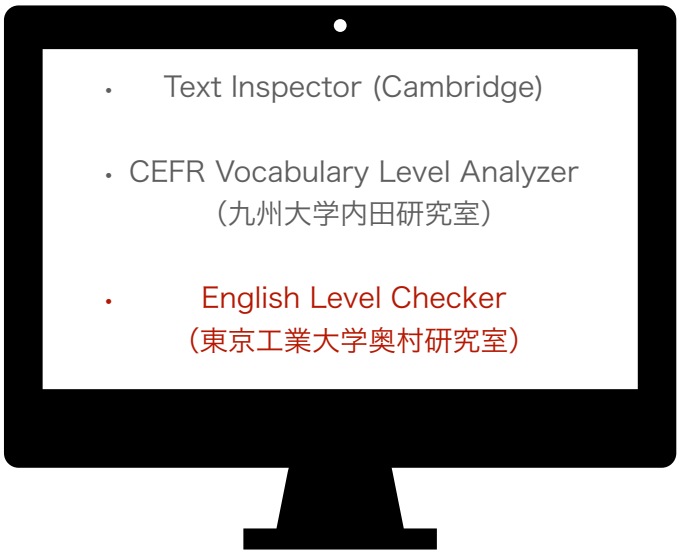
COUNCIL OF EUROPE
CONSEIL DE L'EUROPE
COUNCIL OF EUROPE

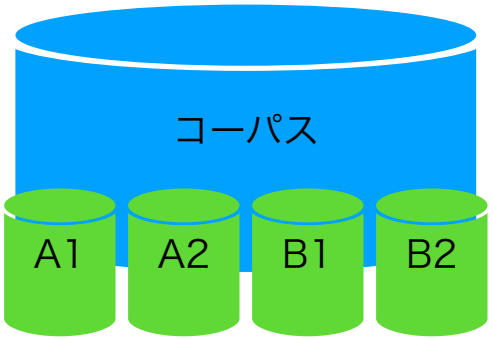
Common European Framework of Reference for Languages (CEFR)



日本の英語教育に特化した枠組

CEFRレベル判定ツール





学習用レベル調整済み英語コーパス

- CEFRレベルでテキスト難易度を限定可能
- 多様なジャンル（13分野；下位項目39分野）
- 文法事項を選択して例文検索が可能

コーパス作成の流れ

- **収集対象URLの決定**
  - 多様なジャンルを含むようにURLを決定
- **ファイル収集**
  - スタートページからクローリングを行いURLリストを作成
  - HTMLファイルをダウンロード
- **本文テキストの抽出**
  - タグ単位、ブロック単位で本文テキストの抽出方法を検証
  - BootCatをベースとして性能評価
- **CEFR-J レベル判定**
  - ６段階のCEFR-Jレベルの自動判定
- **フィルタリング**
  - テキストの言語判定、POSタグなどの情報を追加
  - それを元に用いるファイルを決定

収集対象URLの決定

決定方法

1. 取得分野の仮ジャンルを設定（全68 ジャンル）
2. webページを直接確認して、**人手で取得するURL**とクローリングを行う**スタートページURL**を決定

URLの種類

A) 人手で取得したURL
- 人手で直接確認しているので**取得したいテキストが必ず存在**
- ページ内のノイズ（本文以外のテキストや記号）は比較的少ない

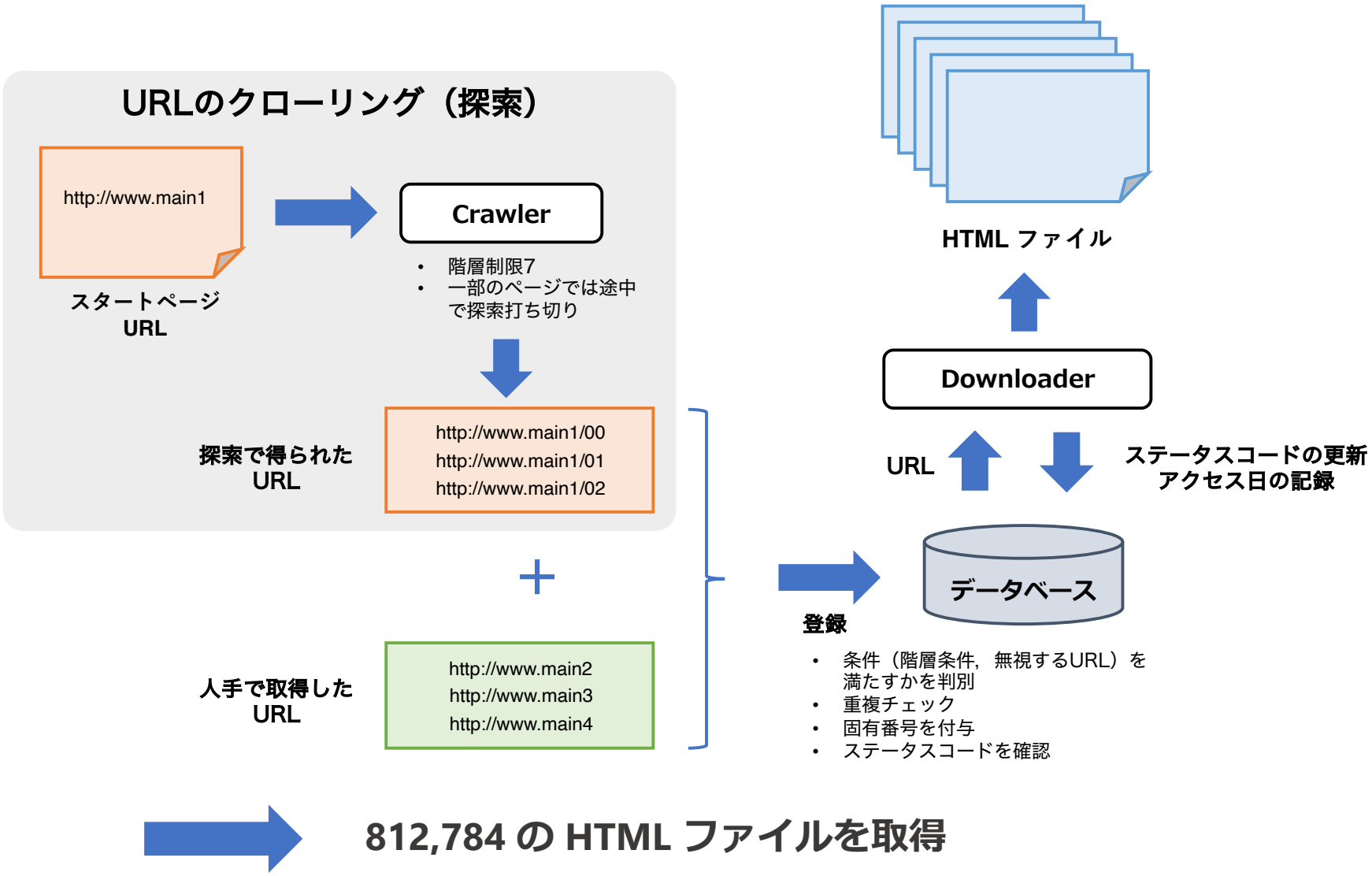
B) スタートページURL
- クローリングする際に探索開始位置となるURL
- どのようなジャンルのURLが得られそうかを予め人手で確認

C) 探索で得られたURL
- スタートページからのクローリングによって自動取得されたURL
- テキストが存在しない**不必要なページも含まれる**

  ➡ **取得した後にフィルタリングが必要**

	人手で取得したURL	スタートページURL	探索で得られたURL
URLの数	551	243	988,599

ファイル収集

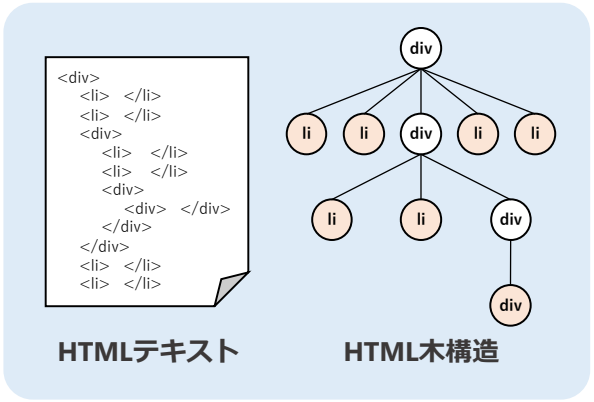

URLのクローリング（探索）
http://www.main1
スタートページ URL
Crawler
• 階層制限7
• 一部のページでは途中で探索打ち切り
探索で得られた URL
http://www.main1/00
http://www.main1/01
http://www.main1/02
人手で取得した URL
http://www.main2
http://www.main3
http://www.main4
HTML ファイル
Downloader
URL
ステータスコードの更新
アクセス日の記録
データベース
登録
• 条件（階層条件，無視するURL）を満たすかを判別
• 重複チェック
• 固有番号を付与
• ステータスコードを確認
812,784 の HTML ファイルを取得

本文テキストの抽出

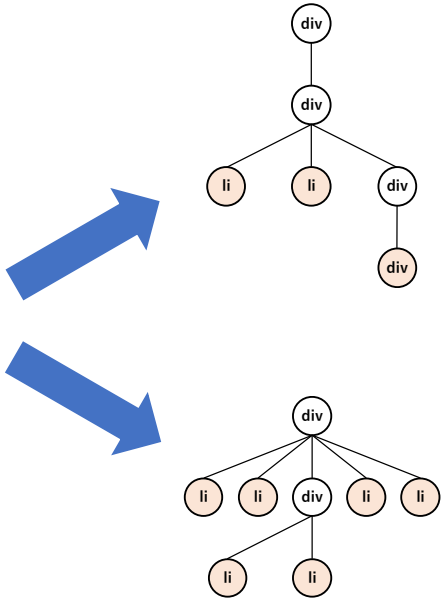
複数の本文抽出方法で抽出精度を検証

1. BootCat
2. HTMLの構造を利用した抽出法（タグ単位）
3. HTMLの構造を利用した抽出法（ブロック単位）

HTMLの構造を利用した抽出方法



取得したHTMLファイル



タグ単位

- htmlのliタグ、divタグなどテキストが存在する可能性が高いタグ構造のルールを作成
- ルールに基づいて不要タグの刈り込みを行う
- **テキストを刈り込みすぎる傾向**

ブロック単位

- タグ単体ではなくタグの集合からテキスト位置の推定
- 複数のタグに含まれているテキストをまとめて抽出
- **テキストがまとまっていない場合は抽出できない**

本文テキストの抽出

抽出方法の評価

抽出方法の選定を行うために以下の方法で評価を行った

1. 抽出できたテキストの量
   - 文数と単語数による定量評価
   - 多くのテキストを抽出した方が評価が高い

2. テキスト品質の人手評価
   - 複数のジャンルが含まれるように**サンプリング**
   - 『webページのフッダ/ヘッダなどの不要な部分が含まれていないか』,
     『抽出対象のテキストに取りこぼしがないか』の２点で評価

BootCatとブロック単位のテキスト抽出方法を採用
（それぞれのファイルに対して2つの方法で抽出を行い、得られた単語数の多い方を採用）

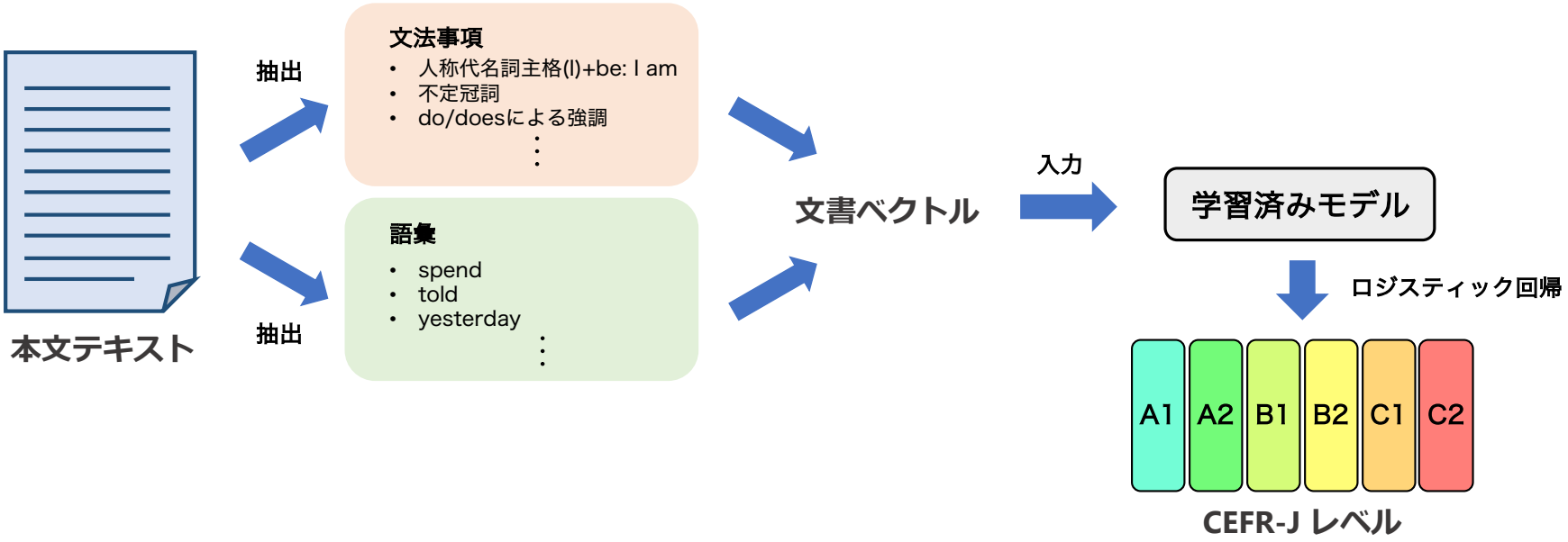
抽出方法を採用したファイル数

抽出方法	採用ファイル数
BootCat	328,464
ブロック単位	253,217
合計	**581,681**

CEFR-J レベル判定

CEFR-J レベル判定の流れ

1. 本文テキストから文法事項・語彙を抽出しベクトル化
2. 文書ベクトルを学習済みモデルに入力
3. ロジスティック回帰を用いてレベル判定



※レベル判定器の学習に用いたデータの特性上、C2レベルの信頼性は低い

フィルタリング

本文テキストに対して、実際にコーパスに含めるテキストを決めるフィルタリングを行う

フィルタリングの為の追加情報を付与

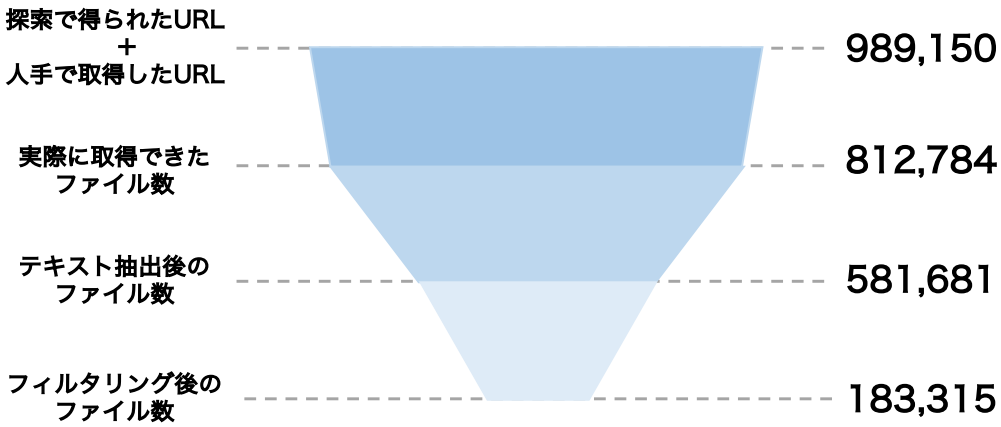
- POSタグ付（SpaCy を利用）
- 言語判定（SpaCy-Langdetectを利用）

フィルタリングの基準

フィルタリング項目	コーパスに含める条件
CEFRクラス	**A1, A2, B1, B2**
有効文数（Punctuationで終わる文）	15 以上
動詞の数	10 以上
最大単語数/文	300 以下
最大単語長	120 以下
代名詞の数	6 以上
言語判定結果	英語

コーパスの統計情報

URL数・ファイル数の推移

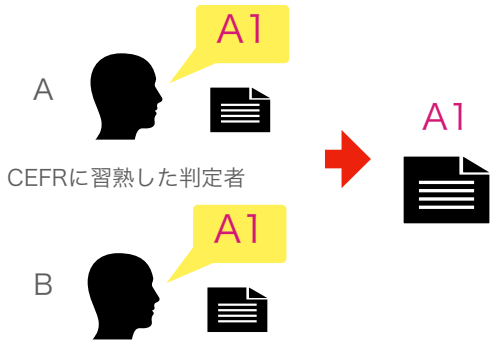


- フィルタリング後のファイル数は **183,315ファイル**
- 今後、コーパスに含めるファイルを更に選定するので、最終的なファイル数はもう少し減少する見込み

CEFR-J クラス別のファイル数・単語数

CEFR-J クラス	ファイル数	単語数	有効文数	単語数の割合 (%)
A1	394	422,194	29,513	**0.22**
A2	3,861	3,451,683	210,715	1.59
B1	37,553	52,161,365	2,728,760	20.61
B2	141,507	204,339,112	10,272,449	77.58
合計	**183,315**	**260,374,354**	**13,241,437**	100

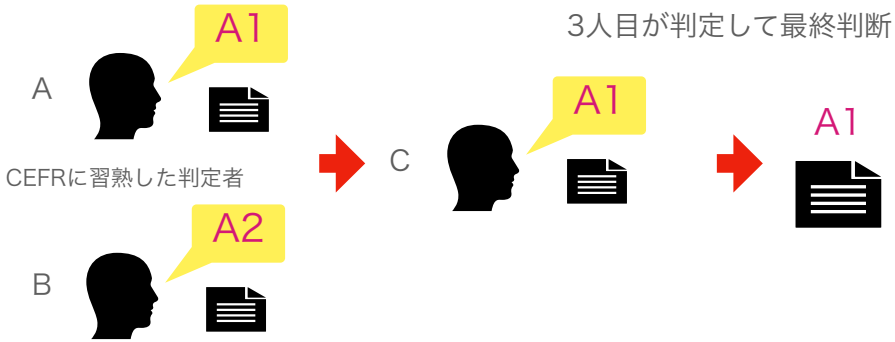
CEFRレベル判定精度の評価



基本は2名のCEFR判定

Machine ratings

Human Ratings	A1	A2	B1	B2	Total
A1	5	6	2		13
A2	13	8	8	4	33
B1	2	5	9	10	26
B2		1	1	6	8
Total	20	20	20	20	80



2名が異なる判定の場合,
3人目が判定して最終判断

最終判定と各判定者の Cohen's Kappa = 0.7

この人手の判定ファイル80件をベースに
ツールの判定結果を評価した

CEFR	人手の判定	ツール判定	一致率	コメント
A1	13	5	38.46%	機械では A2と判定したものが6件あった
A2	33	8	24.24%	機械では A1と判定したものが13件あった 機械では Bレベルに判定したものが12件あった
A-level	46	33	71.74%	Aレベル全体では７割の正答率
B1	26	9	34.62%	機械ではB2と判定したものが10件あった
B2	8	6	75.00%	機械ではA2と判定したものが1件あった
B-level	34	26	76.47%	Bレベル全体の推定は7割の正答率

判定ツールの精度はCEFRの6段階ではまだそれほど高くはないが，Aレベルとレベルを分ける部分に関しては7割強の精度を出せているので，webコーパスのテキスト難易度分類を自動で行った後のレベル調整方法を今後考案して，人手の分類に近い分類に近づけたい.

Hit	KWIC			File
1	ules/5/categories/5000/articles/5102 My home is	in	a large town. My town has many buildings	b1_1_10400.I
2	lots of streets and homes. My town is	in	a state in the United States. Many different	b1_1_10400.I
3	and homes. My town is in a state	in	the United States. Many different towns can be	b1_1_10400.I
4	United States. Many different towns can be found	in	my state. Some towns are small. Only a	b1_1_10400.I
5	are small. Only a few hundred people live	in	them. Big towns are called cities. Large cities	b1_1_10400.I
6	Large cities have millions of people. Towns are	in	every state and country in the world. My	b1_1_10400.I
7	people. Towns are in every state and country	in	the world. My town has grocery stores. We	b1_1_10400.I
8	buy bread and milk. We also have restaurants	in	my town. My favorite restaurant serves pancakes	b1_1_10400.I
9	te restaurant serves pancakes and French toast.	In	my town, there is a post office. I	b1_1_10400.I
10	get healthy. I know how to be safe	in	my town. I learn how to read a	b1_1_10400.I
11	town. I learn how to read a map	in	case I get lost. My parents tell me	b1_1_10400.I
12	parents tell me which adults I can trust.	In	an emergency, I call 911.	b1_1_10400.I
13	for lunch? B: Sure! Where did you have	in	mind? A: I was thinking of Joe's	b4_1_0017.me
14	mind? A: I was thinking of Joe's	in	the village. B: I love that place. Sure,	b4_1_0017.me
15	: Good morning, this is Ray speaking. Is Lee	in?	B: Hi, Ray. This is Le	
16	would be great. What restaurant did you have	in	mind? A: We could go to	
17	This is a pilot. Have you ever flown	in	an airplane? If so, you	
18	s pictures for a living. Some photographers work	in	a studio. They use diffe	
19	wonderful student. A: She isn't messing up	in	class? B: Of course not	
20	. B: It is my pleasure to have her	in	my class. 2. A: We f	
21	omething wrong? B: I enjoy having your daughter	in	my class. A: I'm glad t	
22	. B: I'm more than happy having her	in	my class. 3. A: How	
23	discuss your daughter. A: Is she acting up	in	class? B: Not at all. S	
24	at all. She's a joy to have	in	my class. A: Is she rea	
25	education. Her first class was math. She went	in	. She was nervous. The te	
26	to study a lot. She tried to study	in	her room. Her baby broth	
27	brought her book to the bathroom. She studied	in	there.	
28	asks to see their garden. They let her	in	. Their garden is beautif	
29	ooms usually had magazines. There were magazines	in	Spanish. Eileen can't re	

A1-level テキストのコンコーダンス

B2-level テキストのコンコーダンス

Hit	KWIC			File
1	main streets of days gone by and hiking winding trails	in	secluded forests make you feel like you're in anothe	a1_1_10254.M
2	g trails in secluded forests make you feel like you're	in	another time. With more than 150 miles of shoreline	a1_1_10254.M
3	an overnight experience, pitch a tent under the stars	in	Parvin State Park in Pittsgrove. More into glamping?	a1_1_10254.M
4	nce, pitch a tent under the stars in Parvin State Park	in	Pittsgrove. More into glamping? The park has 18 furn	a1_1_10254.M
5	Camden's USS New Jersey, the most decorated battleship	in	U.S. Navy history and Princeton's War Memorial. Othe	a1_1_10254.M
6	istory and Princeton's War Memorial. Other attractions	in	the Delaware River Region include Camden Children's	a1_1_10254.M
7	ousel or climb the treehouse and Grounds for Sculpture	in	Hamilton, home to 270 outdoor sculptures. Award-win	a1_1_10254.M
8	und throughout the region, including Coda Rossa Winery	in	Franklinville, which produces an Ameritage (it's an	a1_1_10254.M
9	s like Berlin Brewing Company offering microbrews made	in	-house. Major annual events in this area include the	a1_1_10254.M
10	ffering microbrews made in-house. Major annual events	in	this area include the Fire & Ice Festival, held in l	a1_1_10254.M
11	nts in this area include the Fire & Ice Festival, held	in	late January in Mount Holly and the Cranberry Festiv	a1_1_10254.M
12	include the Fire & Ice Festival, held in late January	in	Mount Holly and the Cranberry Festival, held in earl	a1_1_10254.M
13	anuary in Mount Holly and the Cranberry Festival, held	in	early October in Bordentown.	a1_1_10254.M
14	olly and the Cranberry Festival, held in early October	in	Bordentown.	a1_1_10254.M
15	i/laws03.htm Small Claims Court Occasionally, people	in	the United States get into conflict with other peopl	a1_3_0002.Sm
16	where disputes are resolved quickly and inexpensively.	In	this court, the rules are simplified and the hearing	a1_3_0002.Sm
17	not need to be a U.S. citizen to file a suit	in	this court. Here are some of the things individuals	a1_3_0002.Sm
18	are some of the things individuals can file a suit for	in	Small Claims Court. If your former landlord refused	a1_3_0002.Sm
19	and refuses to pay for the repair of the victim's car.	In	the U.S. the person who is at fault of the accident	a1_3_0002.Sm
20	are just a few examples of what a person can sue for	in	Small Claims Court. For more information, visit the	a1_3_0002.Sm
21	,7-247-49025-34612--,00.html Why do we tip the server	in	restaurants? There is an often told beginning for	a2_1_10322.M
22	now. But, alas, the story isn't true. There were boxes	in	English inns and pubs that held coins to "tip" the w	a2_1_10322.M
23	en, which means "to tap." The expression "hot tip," as	in	a sure winner in a horse race, also comes from the a	a2_1_10322.M
24	to tap." The expression "hot tip," as in a sure winner	in	a horse race, also comes from the act of tapping. In	a2_1_10322.M
25	r in a horse race, also comes from the act of tapping.	In	the old days, during card games, gamblers would have	a2_1_10322.M
26	ays, during card games, gamblers would have an partner	in	the room. This partner would signal the player of	a2_1_10322.M
27	ilton Village Council has bought a small piece of land	in	the hopes of making an intersection safer for driver	b1_1_10418.S
28	ed to purchase property at 219 Stillwater St. for $200	in	order to improve the intersection of Jay, Market and	b1_1_10418.S
29	ntil August, Miller said, for those who are interested	in	learning more about the local police department. Co	b1_1_10418.S
30	t there will be no vendor fees for children interested	in	selling items on those days.	b1_1_10418.S

まとめ

CEFR-J準拠webコーパス開発

- 教育用コーパスの自動構築の試み
- CEFR-Jリソースを活用することでCEFRレベル別のテキストをある程度自動収集できる可能性を示した
- 多様なジャンルでCEFRレベル表示があり、かつ文法事項単位の検索ができるコーパス
  → 外国語教育支援への強力な武器になる
- 課題：
  - Aレベル，Bレベル内の詳細レベルの判定精度
  - A1レベルのテキストがwebでは少ない