国立大学法人
電気通信大学
Unique & Exciting Campus

言語資源の設計・再設計と言語資源を活用した実習授業の設計

電気通信大学大学院
情報理工学研究科 情報学専攻
助教　松吉俊

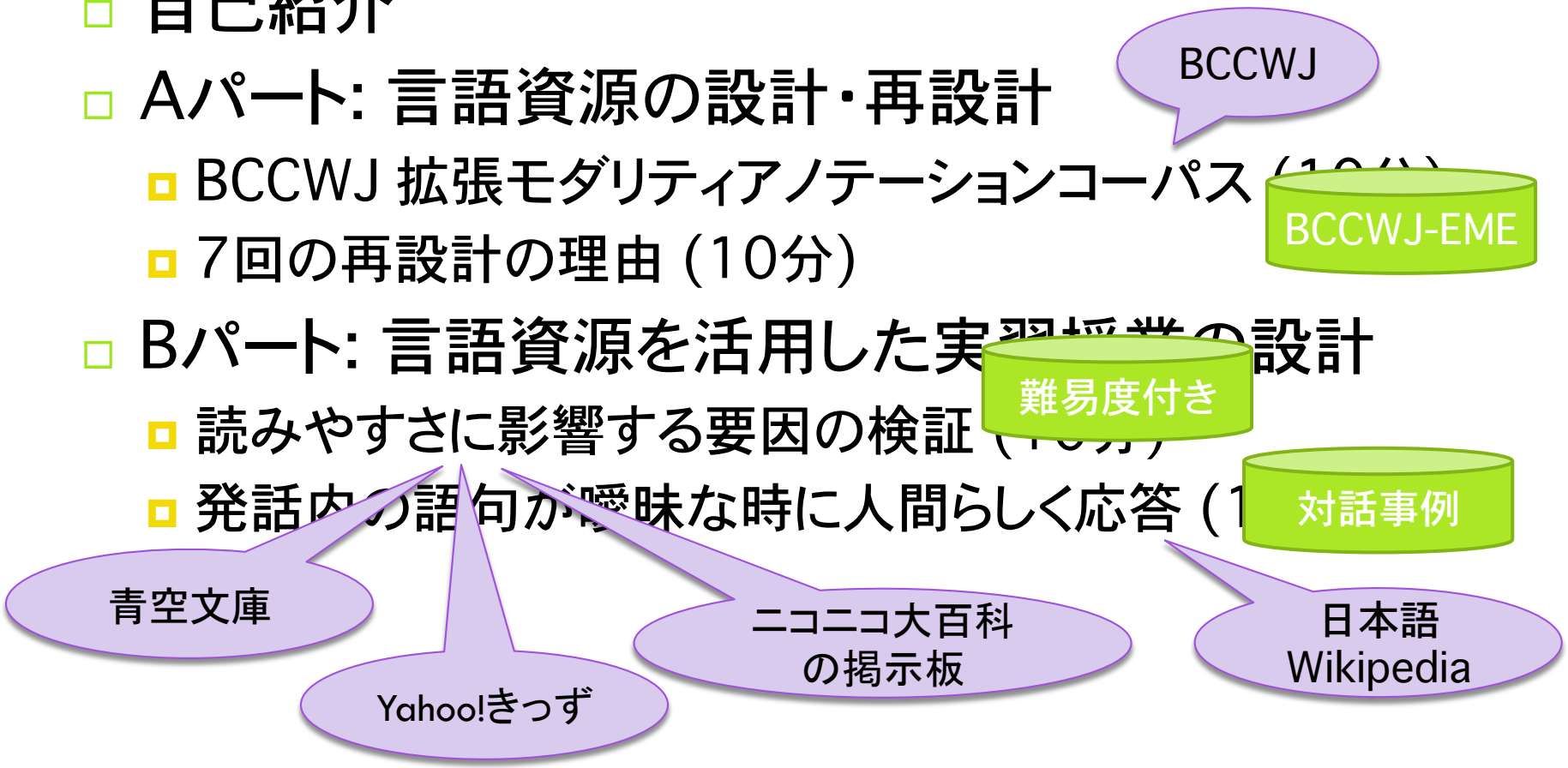
2017年3月8日(水)

発表の流れ

- 自己紹介
- Aパート: 言語資源の設計・再設計
  - BCCWJ 拡張モダリティアノテーションコーパス (10分)
  - 7回の再設計の理由 (10分)
- Bパート: 言語資源を活用した実習授業の設計
  - 読みやすさに影響する要因の検証 (10分)
  - 発話内の語句が曖昧な時に人間らしく応答 (10分)

発表の流れ

- 自己紹介
- Aパート: 言語資源の設計・再設計
  - BCCWJ 拡張モダリティアノテーションコーパス（10分）
  - 7回の再設計の理由（10分）
- Bパート: 言語資源を活用した実習授業の設計
  - 読みやすさに影響する要因の検証（10分）
  - 発話内の語句が曖昧な時に人間らしく応答（1



関わったアノテーション

名称	元データ	対象	作業者	グループ
日本語複合辞用例データベース	毎日新聞	複合辞	複数	佐藤・宇津呂
全教科日本語教科書コーパス	市販の教科書	文字	1人	佐藤
日本語言明間意味的関係コーパス	Webページ	文の対	1人	乾・村上
拡張モダリティアノテーションコーパス	BCCWJ	事象	1人	乾
否定の焦点情報アノテーションコーパス	BCCWJ 楽天データ	事象	複数	
評価視点別レビュー要約コーパス	楽天データ	文	複数	福本

発表の流れ

- 自己紹介
- Aパート: 言語資源の設計・再設計
  - BCCWJ 拡張モダリティアノテーションコーパス (10分)
  - 7回の再設計の理由 (10分)
- Bパート: 言語資源を活用した実習授業の設計
  - 読みやすさに影響する要因の検証 (10分)
  - 発話内の語句が曖昧な時に人間らしく応答 (10分)

言語資源の設計→再設計

拡張モダリティアノテーションコーパス

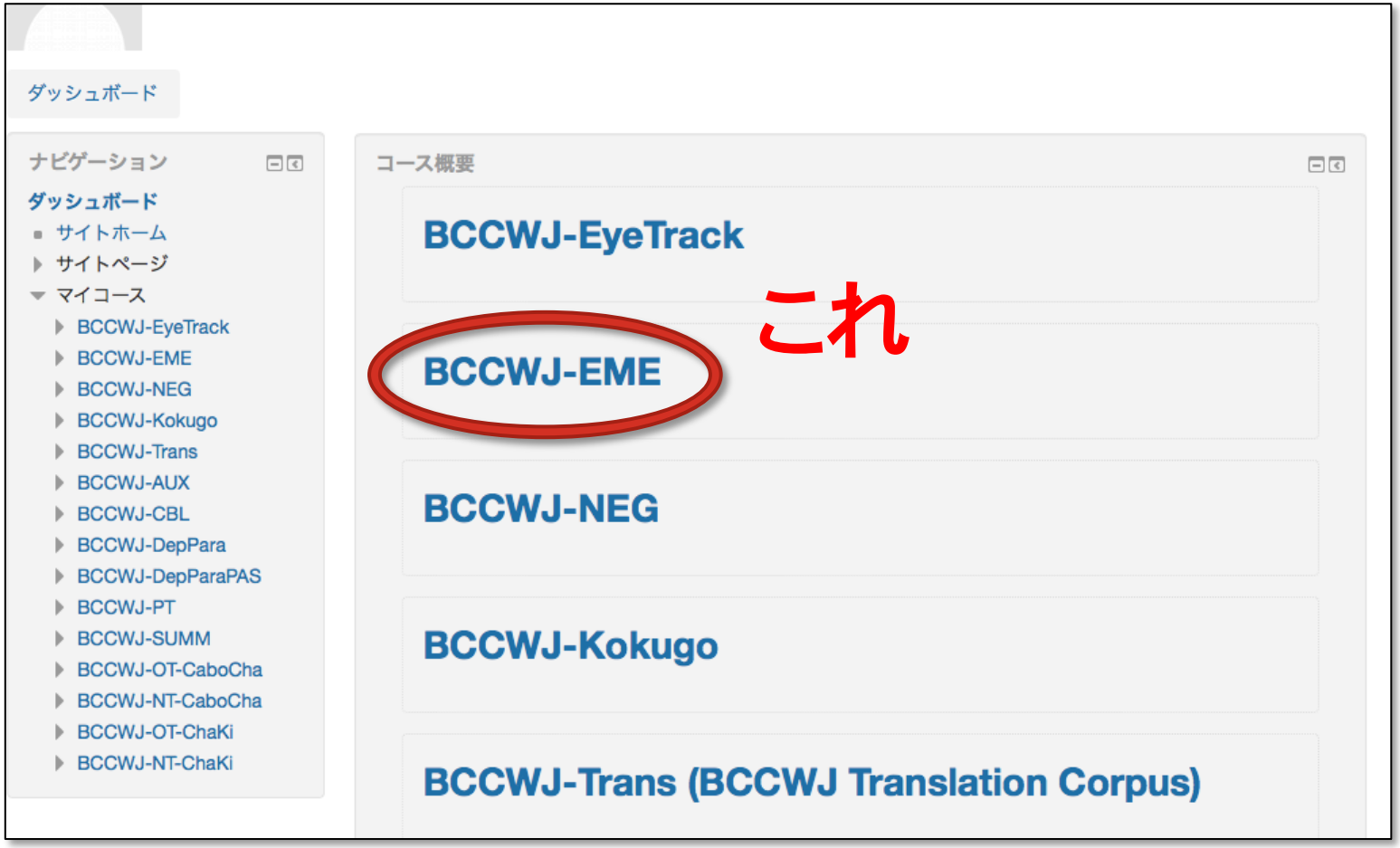
- テキストに対して、**事象のモダリティ**の情報を付与
  - 事象 = 述語項構造
  - 述語 = 動詞、形容詞、形状詞、名詞述語
- (現在は) 6項目のモダリティ関連ラベルを付与

- 2016年3月から、中納言にてダウンロード可能

中納言


中納言 ロク
コーパス検索アプリケ
コーパス検索アプリケーション
「中納言」とは？
国立国語研究所で開発されたコーパスを検索することができる Web アプリケーションです。短単位・長単位・文字列の 3 つの方法によってコーパスに付与された形態論情報を組み合わせた高度な検索を行うことができます。
ご利用には登録（無償）が必要です。詳しくはユーザ登録の申請をご覧ください。
■問い合わせ先
kotonoha@ninjal.ac.jp
Copyright © National Institute for Japanese Language a
中納言 コーパス選択
コーパス検索アプリケーション
ご利用になりたいコーパス名をクリックしてください。
ログアウト
コーパス名
略称
状態
関連サイト
備考
現代日本語書き言葉均衡コーパス 通常版
BCCWJ-NT
利用可能
関連データ配布
より利用いただいている BCCWJ のデータです。コーパスの紹介ページへのリンク
現代日本語書き言葉均衡コーパス 非 numTrans 版
BCCWJ-OT
利用可能
非 NumTrans 版（OT）とは
日本語歴史コーパス
CHJ
利用できません
利用申請
日本語話し言葉コーパス
CSJ
利用可能
多言語母語の日本語学習者横断コーパス
I-JAS
利用できません
利用申請
発話の音声ファイルや書き起こしたプレインテキストは、関連データ配布サイトからダウンロードできます。その他の資料については、I-JAS 関連資料を参照ください。
名大会話コーパス
名大会話コーパス
利用可能
Copyright © National Institute for Japanese Language and Linguistics.

BCCWJ関連データ配布サイト



BCCWJ-EME


BCCWJ-EME
ダッシュボード ▶ BCCWJ ▶ BCCWJ-EME
ナビゲーション
ダッシュボード
サイトホーム
サイトページ
現在のコース
BCCWJ-EME
参加者
一般
マイコース
『BCCWJ 拡張モダリティアノテーションコ
version 0.81_20160310 for BCCWJ 1.0
・概要
本コーパスは、『現代日本語書き言葉均衡コ
・謝辞
本研究は、以下に挙げ
です。
- 独立行政法人 情報通信研究機構 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」
- 科研費若手研究(B)「高精度モダリティ解析のための言語資源構築に関する研究」(課題番号: 23700176、研究代表者: 松吉俊)
packages
一番下
BCCWJ-EME
ダッシュボード ▶ BCCWJ ▶ BCCWJ-EME ▶ 一般 ▶ packages
ナビゲーション
ダッシュボード
サイトホーム
サイトページ
現在のコース
BCCWJ-EME
参加者
一般
packages
マイコース
packages
BCCWJ-EME_0.81.zip
ダウンロード
できます

データの形式

- XMLファイル
  - オリジナル
- 拡張CaboChaフォーマット簡易形式ファイル
- TSVファイル
  - 自動処理向け
- **Excelファイル**
  - **人間向け**
  - TSVフォルダーの中にあります

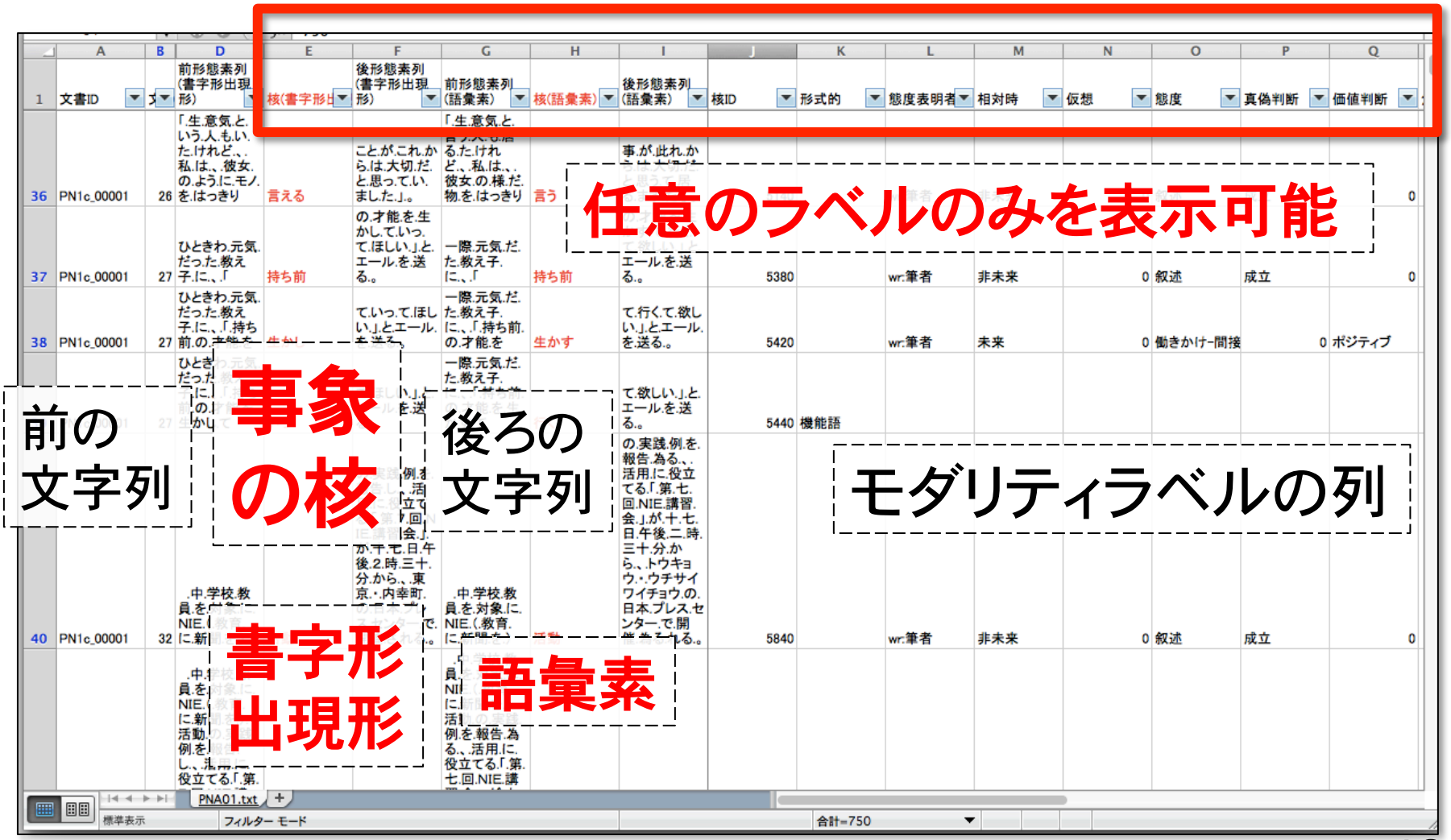
Excelデータの見方

	A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	文書ID	文	前形態素列(書字形出現形)	核(書字形出	後形態素列(書字形出現形)	前形態素列(語彙素)	核(語彙素)	後形態素列(語彙素)	核ID	形式的	態度表明者	相対時	仮想	態度	真偽判断	価値判断
36	PN1c_00001	26	「.生.意気.と.いう.人.も.い.た.けれど、.私.は、.彼女.の.よう.に.モノ.を.はっきり	言える	こと.が.これ.から.は.大切.だ.と.思っ.て.い.ました.」。	「.生.意気.と.言う.人.も.居る.た.けれど、.私.は、.彼女.の.様.だ.物.を.はっきり	言う	事.が.此れ.から.は.大切.だ.と.思う.て.居る.ます.た.」。	5140		wr:筆者	非未来	0	叙述	成立	0
37	PN1c_00001	27	ひときわ.元気.だっ.た.教え子.に、.「	持ち前	の.才能.を.生かし.て.いっ.て.ほしい.」.と.エール.を.送る.。	一際.元気.だ.た.教え子.に、.「	持ち前	の.才能.を.生かす.て.行く.て.欲しい.」.と.エール.を.送る.。	5380		wr:筆者	非未来	0	叙述	成立	0
38	PN1c_00001	27	ひときわ.元気.だっ.た.教え子.に、.「.持ち前.の.才能.を	生かし	て.いっ.て.ほしい.」.と.エール.を.送る.。	一際.元気.だ.た.教え子.に、.「.持ち前.の.才能.を	生かす	て.行く.て.欲しい.」.と.エール.を.送る.。	5420		wr:筆者	未来	0	働きかけ-間接	0	ポジティブ
39	PN1c_00001	27	ひときわ.元気.だっ.た.教え子.に、.「.持ち前.の.才能.を.生かし.て	いっ	て.ほしい.」.と.エール.を.送る.。	一際.元気.だ.た.教え子.に、.「.持ち前.の.才能.を.生かす.て	行く	て.欲しい.」.と.エール.を.送る.。	5440	機能語						
40	PN1c_00001	32	.中.学校.教員.を.対象.に.NIE.(.教育.に.新聞.を.)	活動	の.実践.例.を.報告.し、.活用.に.役立てる.「.第.7.回.NIE.講習.会.」.が.十.七.日.午後.2.時.三十.分.から、.東京.・.内幸町.の.日本.プレス.センター.で.開催.さ.れる.。	.中.学校.教員.を.対象.に.NIE.(.教育.に.新聞.を.)	活動	の.実践.例.を.報告.為る.、.活用.に.役立てる.「.第.七.回.NIE.講習.会.」.が.十.七.日.午後.二.時.三十.分.から、.トウキョウ.・.ウチサイワイチョウ.の.日本.プレス.センター.で.開催.為る.れる.。	5840		wr:筆者	非未来	0	叙述	成立	0
			.中.学校.教員.を.対象.に.NIE.(.教育.に.新聞.を.).活動.の.実践.例.を.報告.し、.活用.に.役立てる.「.第.			.中.学校.教員.を.対象.に.NIE.(.教育.に.新聞.を.).活動.の.実践.例.を.報告.為る.、.活用.に.役立てる.「.第.七.回.NIE.講										

PNA01.txt

標準表示　フィルター モード　合計=750

Excelデータの見方



ver0.8のラベルの例

あの時彼女に真実を伝えるべきだったなと太郎が言った。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者_13:太郎	非未来	0	叙述	不成立	ポジティブ
wr:筆者	非未来	0	叙述	0	0

それ以来、医師たちはその薬を使い始めました。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	不成立から成立	0

今夏に予定している販売開始のめどが立たない状況に陥っている。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	未来	0	叙述	低確率	0

急激にその使用を中止するとリバウンド現象が起こります。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	条件	叙述	成立から不成立	0

ver0.8のラベルの例

あの時彼女に真実を伝えるべきだったなと太郎が言った。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
r:筆者_13:太郎	非未来	0	叙述	不成立	ポジティブ
wr:筆者	非未来	0	[illegible]	0	0

情報源の入れ子

叙述, 意志, 欲求,
働きかけ-直接, 働きかけ-間接,
働きかけ-勧誘, 許可, 問いかけ

事象成立の望ましさ

それ以来、[illegible]。

真偽判断	価値判断
不成立から成立	0

今夏に予定している販売開始のめどが立たない状況に陥っている。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	未来	0	叙述	低確率	0

急激にその使用を中止するとリバウンド現象が起こります。

事実性とその変化

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	条件	叙述	成立から不成立	0

BCCWJコアデータ内の対象データ

	Yahoo! 知恵袋 (OC)	白書 (OW)	新聞 (PN)	書籍 (PB)	計
文数	6,404	5,835	16,433	9,869	38,541
形態素数	110,649	228,651	360,814	234,540	934,654
拡張モダリティの事象数	15,781	7,733	8,819	9,466	41,799

✓ 雑誌とYahoo!ブログには未着手

- ファイル名の規則:
  - OCA01.xlsx

レジスター　優先順位

コーパス作成目的

- 自然言語処理の情報抽出への応用
  - 特に、事実文や要望文の抽出
- 既存の技術:
  - 単に述語項構造を認識するのみ
- やりたいこと:
  - 事実かそうでないかの区別
  - 文に(陽に)含まれる事象の認識
- 機械学習の訓練データとしたい

ラベルの例(2/5)

県内で新型インフルエンザが発生した。 (事実)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	成立	0

県健康推進課が県内で新型インフルエンザが発生したと報告した。 (伝聞)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者_4:課	非未来	0	叙述	成立	0
wr:筆者	非未来	0	叙述	0	0

県内で新型インフルエンザが発生したとみられる。 (推量)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	高確率	0

県内で新型インフルエンザが発生した可能性がある。 (推量)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	高確率	0

ラベルの例(3/5)

県内で新型インフルエンザが発生した可能性は低い。(推量)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	低確率	0

まるで県内で新型インフルエンザが発生したようなパニックが起こっている。(比況)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
--	--	--	--	--	--

対象外
比況

県内で新型インフルエンザが発生したわけではない。(通常否定)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	不成立	0

県内で新型インフルエンザが発生したというのは正しくない。(メタ否定)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	不成立	0

ラベルの例(4/5)

県内で新型インフルエンザが発生したら、どう対応するべきか。(仮定)

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	未来	条件	叙述	成立	0

あの時県内で新型インフルエンザが発生していたら、パニックになっていただろう。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	不成立	0

(反実仮想)

あの時県内で新型インフルエンザが発生していたら、パニックになっていただろう。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	不成立	0

(反実仮想)

1時間後に駅に集合したら、その足でいつもの居酒屋へ直行しよう。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	未来	0	働きかけ-勧誘	高確率	ポジティブ

ラベルの例(5/5)

太郎と花子の両方が買ったわけではない。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	不成立	0

焦点
否定-5:両方

太郎とおそらく**花子が買った**。

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	高確率	0

焦点
推量-4:花子

太郎が買ったのは、週刊誌ではない。「太郎が週刊誌を買うコト」

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	不成立	0

焦点
否定-8:週刊誌

太郎はCDをほとんど買わない。「太郎がCDを買うコト」

態度表明者	相対時	仮想	態度	真偽判断	価値判断
wr:筆者	非未来	0	叙述	成立	0

程度
5:ほとんど,7:ない

発表の流れ

- 自己紹介
- Aパート: 言語資源の設計・再設計
  - BCCWJ 拡張モダリティアノテーションコーパス (10分)
  - 7回の再設計の理由 (10分)
- Bパート: 言語資源を活用した実習授業の設計
  - 読みやすさに影響する要因の検証 (10分)
  - 発話内の語句が曖昧な時に人間らしく応答 (10分)

7回の再設計

1. ver0.1 = 最初の設計
2. ver0.2
3. ver0.3
4. ver0.4
5. ver0.5
6. ver0.6
7. ver0.7
8. ver0.8 ← イマココ

再設計の理由

	言語学的観点	工学側からの要請
0.1→0.2	整理	
0.2→0.3	整理	単純化
0.3→0.4		情報追加
0.4→0.5	整理	
0.5→0.6		情報追加 体系整理
0.6→0.7	整理 用語整理	
0.7→0.8		単純化

- 更新履歴の詳細は、同梱のmanual.pdfをご参照ください
  - 付録C (p.30 ～ p.34)

再設計の理由

	言語学的観点	工学側からの要請
0.1→0.2	整理	
		単純化
		情報
		情報 体系
0.6→0.7	整理 用語整理	
0.7→0.8		単純化

以下を求める
- 体系的に記述すること
- 網羅的に記述すること
- 例外がないこと

以下を求める
- 体系的に記述すること
- 目的に合った最低限の粒度を持っていること
- 機械学習訓練データが早く完成すること

目的の相違であり、どちらが良い悪いというものではない

- 更新履歴の詳細は、同梱のmanual.pdfをご参照ください
  - 付録C (p.30 ～ p.34)

言語学的観点

- ラベルの**整理** <=「**気持ち悪いから**」
  - 「あれがあるのに、これがないのはおかしい」
  - 「ここは細分する必要がある」
  - 「今の体系ではカバーできない事例に遭遇した」

- **用語整理** <=「正式な公開が近い」
  - 先行研究における名称と統一する
  - 必要のない修飾語を落とす

工学側からの要請

- 特別な目的がある
  - **単純化**: 細分はほどほどでよい
  - **情報追加**: 関連研究などにより刺激を受けた
  - **体系整理**: 1つの項目に押し込みすぎである
- 自動ラベル付けツールを作りたい
  - **単純化**、**体系整理**: 機械学習手法の枠組みによる

再設計

	言語学的観点	工学側からの要請
0.1→0.2	整理	
0.2→0.3	整理	単純化
0.3→0.4		情報追加
0.4→0.5	整理	
0.5→0.6		情報追加 体系整理
0.6→0.7	整理 用語整理	
0.7→0.8		単純化

- 「アノテーション学」?

発表の流れ

- 自己紹介
- Aパート: 言語資源の設計・再設計
  - BCCWJ 拡張モダリティアノテーションコーパス (10分)
  - 7回の再設計の理由 (10分)
- Bパート: 言語資源を活用した実習授業の設計
  - 読みやすさに影響する要因の検証 (10分)
  - 発話内の語句が曖昧な時に人間らしく応答 (10分)

言語資源を活用した実習授業の設計

動機付け(1/3)

- 大学のイベントで、高校生向け講義を6回行った
- 高校生は、「言語学」や「計算言語学」を知らない
- 説明しても、
  「もう研究することはない」と思っているもよう

動機付け(2/3)

- 新4年生の研究室配属時、
  優秀な学生が他の(華やかな)研究室に行ってしまう
  - 画像処理
  - 音声処理
  - 大規模計算

動機付け(3/3)

- 若い先生も、その分野の面白さを伝えるべきでは?
  - cf. 高校数学教師の日常的な悩み

- 機会を活かす
  - 計算言語学を教える時に工夫する
- 機会を作る
  - 関連する授業で、計算言語学について話す

実践中

1. 読みやすさに影響する要因の検証
   - 山梨大学工学部コンピュータ理工学科 3年生後期
   - 2014, 2015
   - のべ履修者: 37人

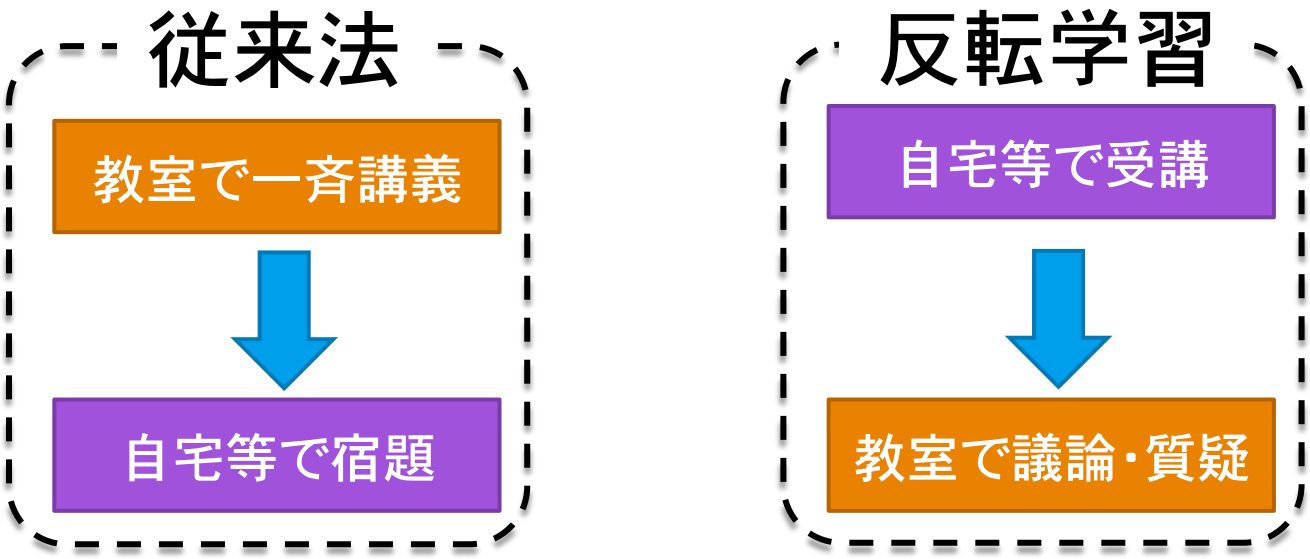
2. 発話内の語句が曖昧な時に人間らしく応答
   - 電気通信大学情報理工学部総合情報学科 3年生後期
   - 2016
   - のべ履修者: 49人

演習授業の設計指針

- 計算言語学や自然言語処理に興味を持ってもらえる
- 学生にテキストアノテーションを体験させる
- プログラミングが得意でない学生に配慮
- 留学生にも配慮
- レポートで評価
  - 学生自らが目標点を設定できる
    (60点を目指すも、100点を目指すも自由)
  - コピペ対策
  - きっちりしたレポートの書き方も指導する
- 反転学習を取り入れる

反転学習(flipped learning)

- 授業資料や説明動画を授業前に学生に公開
- 学生は事前にしっかり予習する
- 90分の授業は、議論や質疑に当てる



実習授業

1. 読みやすさに影響する要因の検証
   - 山梨大学工学部コンピュータ理工学科 3年生後期
   - 2014, 2015
   - のべ履修者: 37人

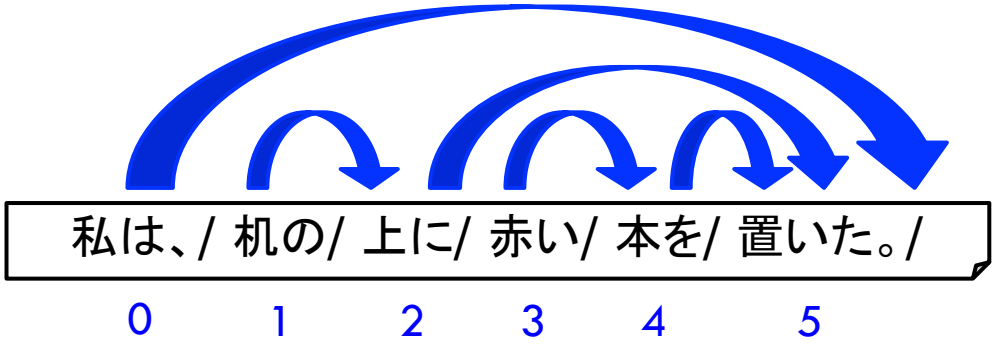
2. 発話内の語句が曖昧な時に人間らしく応答
   - 電気通信大学情報理工学部総合情報学科 3年生後期
   - 2016
   - のべ履修者: 49人

レポート課題(1/5)

- 実験目的
  本実験では、**読みやすさに影響しそうな素性**を少なくとも5つ選択し、それらの有効性を検証する。
  - レポートを書くときは、具体的な数値で下線を置き換える
- 原理・解説
  選択した少なくとも5つの素性の各々について、
  易しい文章や難しい文章で
  その素性値はどのような傾向があると考えられるか
  述べよ。(必要に応じて参考文献を引用する。)

読みやすさに影響しそうな素性の例

- 単純な出現回数
  - 文字、よみ、語、文節、文
- **1文あたりの平均**出現回数
  - 文字、よみ、語、文節、読点
- 語彙表における割合
  - 名詞、動詞、形容詞、格助詞、**「です」や「ます」**
- **複合名詞**（「山梨大学」や「甲府駅南口武田信玄公銅像」など）
  - 延べの出現回数、異なりの出現回数、平均構成形態素数
- **特定種類の係り受け**（述語 = 動詞 or 形容詞 or 形状詞）
  - 名詞->動詞、動詞->動詞、動詞->名詞、名詞->述語、述語->述語
- **係り受けの距離の平均**
  - 文節0と文節5の距離は5、文節2と文節5の距離は3



レポート課題(2/5)

- 実験方法
  使用したデータとプログラムについて説明せよ。
  (同じ実験が再現できるように不足なく書く。)
  - 少なくとも9つのテキストを使用すること
    - 「Yahoo!きっず」が紹介するWebページから少なくとも3つ
    - 「ニコニコ大百科」内の掲示板のスレから少なくとも3つ
    - 上記以外の青空文庫などから少なくとも3つ
  - **各テキストに対して判定した難易度を明記する**
    - 個人の主観で判断してよい
    - 難易度の段階も、2段階以上(易-難、易-並-難)で自由に決めてよい
  - 各素性を抽出するプログラムについて概説する

アノテーション

レポート課題(3/5)

- 実験結果
  各素性に関して、
  すべてのテキストの素性値を統計的に分析せよ。
  - この章では、表やグラフを使って見やすく表現すること

レポート課題(4/5)

- 実験結果に関する考察
  前ページの結果について、有効な素性があったと
  言えるかどうか議論せよ。
  - 可能ならば、一番有効な素性はどれであるか考察する
  - 実験前に想定していたとおりであったかどうか述べる。
    想定と異なる結果が出た場合、どのように仮説を更新すれば
    よいか考察する
- 考察課題
  「ニコニコ大百科」内の掲示板のスレに関して、
  読みやすさの難易度に関わる要因を考察せよ。
  - 読み手の知識や書かれたテキストの複雑さなどの観点から

レポート課題(5/5)

- まとめ
  数行でまとめる。
- 参考文献
  参照した文献やウェブページを列挙せよ。
  本授業の配布資料はここに列挙しなくてよい。
  (他の授業の配布資料は、必要に応じてここに列挙すること。)
- 同時に提出するファイル
  すべてのプログラムとすべてのデータも合わせて
  提出する

15回の授業構成

- 計算言語学は実質4回

回	演習内容
1～11	機械学習(強化学習)
12	BCCWJ短単位語彙表データを利用した文字カウント
13	「こころ」の語彙表を作る
14	HTML文書からテキストを取り出す
15	読みやすさ指標を計算する

学生のアノテーション

- 「Yahoo!きっず」
  - みな、“易”を付ける
- 「青空文庫」
- 「ニコニコ大百科内の掲示板」
  - 学生による
  - 文書による

実習授業

1. 読みやすさに影響する要因の検証
   - 山梨大学工学部コンピュータ理工学科 3年生後期
   - 2014, 2015
   - のべ履修者: 37人

2. 発話内の語句が曖昧な時に人間らしく応答
   - 電気通信大学情報理工学部総合情報学科 3年生後期
   - 2016
   - のべ履修者: 49人
   - 残念ながら、設備の都合で説明動画は提供できず

本実験で利用するプログラム

- Python 3で記述
- ユーザーとシステムの1対1の対話
- ユーザーからの1つの入力に対して、
  システムが応答文を1つ返す
- **人間らしい対話を続けることはできない**

- 複数の意味が推測できるキーワードに敏感
  - この特徴を活かして、限られた範囲ではあるが、
    人間らしい応答を返したい

実行例

```
システム: 会話を楽しみましょう。何も入力しないでEnterを押すと、会話を終了できます。
ユーザー: こんにちは
システム: こんにちは!
ユーザー: 私はとてもワクワクしています。
システム: なるほど。
ユーザー: 昨日、ケーキを食べました。
システム: ケーキが好きなんですね。
ユーザー: 好きなんです、ケーキ。
システム: ケーキが好きなんですね。
ユーザー: そういえば、発売したばかりの「ワンピース 」を買ってきました。
システム: 「ワンピース」は曖昧です。次のどの意味ですか?
1. 2つ以上のパーツで構成されうるものにおいて、全体が一つのパーツで構成されているもの。
2. ワンピース - 上着とスカートが一続きになった女性・子供服。
3. ワンピース（水着）- トップ（上半身）とボトム（下半身）が繋がっている水着。
4. ONE PIECE - 尾田栄一郎作の漫画。ならびに劇中に登場する「ひとつなぎの大秘宝」の呼び名。
5. ONE PIECE（アニメ）- 上記を原作としたテレビアニメ。
6. ONE PIECE（2000年の映画）- 上記テレビアニメの劇場版第1作。
7. ワンピース（短編映画）- 矢口史靖と鈴木卓爾が監督した短編映画作品のシリーズ。
8. ONE PIECE（S-WORDのアルバム）- S-WORDのオリジナル・アルバム。
ユーザー:
```

レポート課題

- 次の2種類を提出してください
  1. **曖昧な語句を見つけた時の対話例** (5点)
     - Excelファイルを提出
     - ファイル名: 学籍番号-dialogue.xls (.xlsxも可)
  2. 曖昧な語句が入力された時に人間らしい応答を返す対話システム (5点)
     - プログラム等をzipで固めて提出
     - ファイル名: 学籍番号-system.zip

レポート課題: 対話例

- テンプレートのExcelファイルをWebClassから入手する
  - 「レポート課題: Excelファイル」
- Excelファイルには、5つのシートがある
  - 3つの「対話」と2つの「入力例」
- 入力例を参考にして、**対話例を3つ考える**
  - **「曖昧な語句」とその意味**は、ambDicと同じものを**記入する**
  - 「説明」欄には、**理想的なシステムの内部動作を記入する**
- 素人でも対話が楽しめるように、適切な参考文献を挙げる
  - 残念ながら、担当教員はなんでも知っているわけではありません
- 公序良俗に反する題材は避けること
- 個人や団体等を誹謗中傷する題材は避けること

アノテーション

対話とアノテーションの例

	A	B	C	D	E	
1	<対話>		説明	曖昧な語句	意味1	意味2
2	システム	会話を楽しみましょう。何も入力しないでEnterを押すと、会話を終了できます。				
3	ユーザー	こんにちは				
4	システム	こんにちは!				
5	ユーザー	私はとてもワクワクしています。				
6	システム	なるほど。				
7	ユーザー	そういえば、発売したばかりの「ワンピース」を買ってきました。		ワンピース	上着とスカートが一続きになった服。	尾田栄一郎の漫画。主
8	システム	発売したね。	どちらか分からないので、無難に返答する。			
9	ユーザー	スカート部分が長め。	キーワードにより、どちらの意味か分かる。			
10	システム	私も、上着とスカートが一続きになった服が好きです。	辞書から文字列を取ってきて返答する。			
11	ユーザー	白色がいいよね。				
12	システム	白色が好きなんですね。				
13						
14	<参考文献>					
15	ワンピース	https://one-piece.com				
16						
17						

対話とアノテーションの例

	A	B	C	D	E	
1	<対話>		説明	曖昧な語句	意味1	意味2
2	システム	会話を楽しみましょう。何も入力しないでEnterを押すと、会話を終了できます。				
3	ユーザー	こんにちは				
4	システム	こんにちは!				
5	ユーザー	私はとてもワクワクしています。				
6	システム	なるほど。				
7	ユーザー	そういえば、発売したばかりの「ワンピース」を買ってきました。		ワンピース	上着とスカートが一続きになった服。	尾田栄一郎の漫画。
8	システム	発売したね。	どちらか分からないので、無難に返答する。			
9	ユーザー	スカート部分が長め。	キーワードにより、どちらの意味か分かる。			
10	システム	ワンピースが一続きになった服が好きなんですね。	辞書から文字列を取ってきて返答する。			
11	ユーザー					
12	システム	白色が好きなんですね。				
13						
14	<参考文献>					
15	ワンピース	https://one-piece.com				
16						
17						

曖昧な語句とその語義

実現したい対話

発話生成のためのアノテーション

加点項目

- 曖昧な語句がちゃんと曖昧である
  - 題材が適切である
- 応答が冴えている
- **オリジナリティーが高い**

- 各発話が自然な日本語である
  - 略称や顔文字などの利用可
- 自然な対話である
- 参考文献が適切である

- システムがかわいい or かっこいい

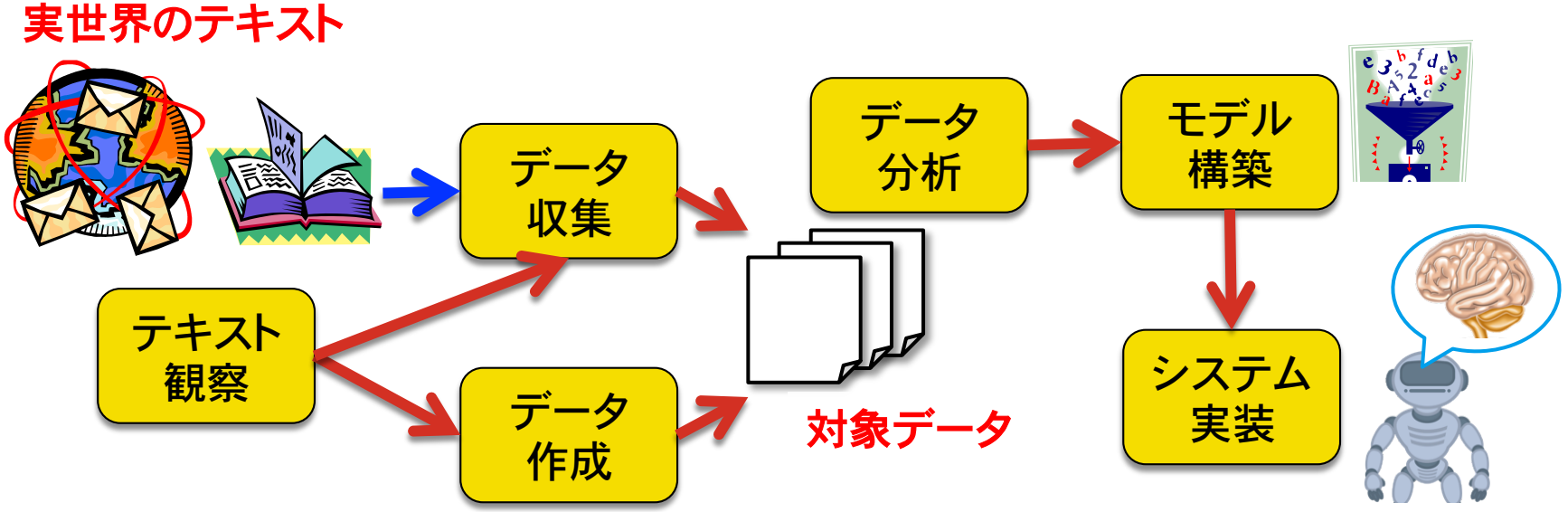
２日で4回の授業構成

- １日に2コマ連続

回	演習内容
1	演習環境を整える
2	対話システムで遊ぶ
3	対話データについて学ぶ
4	対話システムを改良する

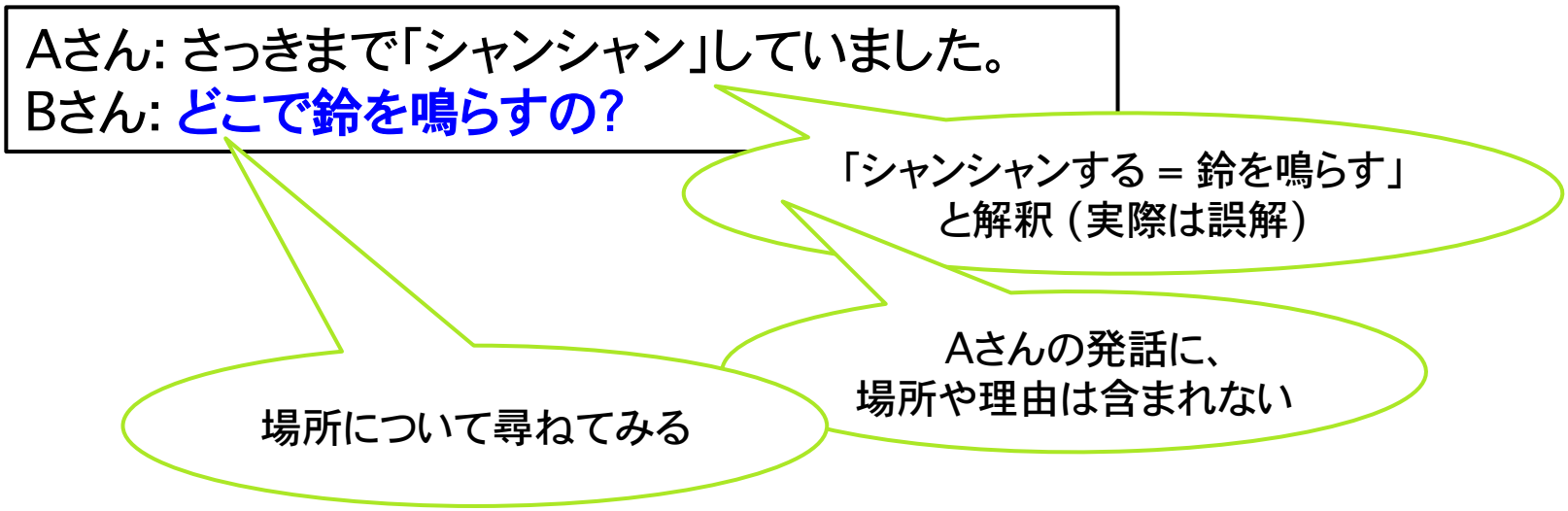
「計算言語学」の説明

- 言語学の1分野
- コンピューターを使って、人間が話す言語について研究する
- 言語学的視点から、文章を処理するための手法を提案する
- 研究の流れ:



テキストデータの分析

- 人間の頭の中で実行されていると思われる過程を明確にする
  - もっともらしい説明で記述することを目指す
  - もちろん、各人で、異なる過程が動いていることもある
- 分析の例:



学生らによる(擬似)対話コーパス

- 曖昧な語句に多様性
  - ほとんどの学生が独自の語義を定義していた
    - 印象的なもの:「元カレ」
  - 思いの外、人名や地名は少なかった
    - 日本語Wikipediaの「曖昧さ回避」ページを渡している影響
- 発話生成アノテーションは、多くの場合、サンプル例の表現をそのまま利用していた
- とてもくだけたスタイルと内容の対話が多く得られた
  - くだけた日本語の解析やくだけた発話の生成の研究に利用できると思われる

今後の課題

- 得られたデータの共有
  - 研究教育利用してもよいか、学生に書面で尋ねる
- 授業資料の共有?
- 同じテキストに対して複数人でラベル付けして
  その傾向や揺れを調査したい場合(難易度など)、
  異なる機関の間で対象データを統一できるか?
  - BCCWJの利用は、しないほうがよい
  - 梵天の検索結果を利用できる?

まとめ

- 言語資源活用に関して2つの話題を話した
  - 再設計の理由
  - 演習授業での活用

- ご静聴ありがとうございました

- ＜電気通信大学は2018年に創立100周年を迎えます＞