『名大会話コーパス』中納言版・ひまわり版 公開データの形態論情報の修正

2022/8/30
言語資源ワークショップ2022

国立国語研究所　柏野　和佳子，　西川賢哉，　渡邊友香，　小磯花絵

名大会話コーパスの概要

◆『名大会話コーパス』は、日本語母語話者の100時間分の129件の会話（雑談）を収録して、文字化したコーパス（https://mmsrv.ninjal.ac.jp/nucc/）。

◆科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（研究代表者　大曾美恵子、平成13年度～15年度）により作成された。

- 名古屋近辺のデータが最も多いが、東京近辺、北海道、新潟で録音されたものもある。
- 共通語による会話が大半を占めるが、方言も使われている。
- 参加者の年代は様々で10代～90代までと幅広い。女性の方が多い。
- 日本語教育関係者、言語研究者が多いので、日本語のメタ言語的な使い方が多い。
- 親しい者同士の雑談が多いが、初対面同士、研究メンバー同士、先輩―後輩の会話もある。
- 話題を一切制限していない雑談であるが、参加者は録音していることを知らされていた。

◆国立国語研究所に移管後、形態素解析用辞書『UniDic』と形態素解析器『MeCab』を用いて形態論情報を自動付与し、人手修正を経て、オンライン検索システム『中納言』、全文検索システム『ひまわり』にて2016年12月より一般公開している。

はじめに

◆最初に行った形態素解析後の人手修正の内容は、柏野ほか(2016)「『名大会話コーパス』中納言版・ひまわり版公開データの作成」にて報告した（http://doi.org/10.15084/00001488）。

◆当時の形態素解析は話し言葉に対しての精度が低く、人手修正がしきれていない誤解析（境界誤り、形態論情報誤り）を多く残していた。現在、『日本語日常会話コーパス』（CEJC）の構築を経て、解析の精度を上げることができている。

◆そこで、『話し言葉UniDic』の3.0.1と3.1.0とで再解析し、現状と比較して出した差分を中心に、全体の形態素解析の修正を行った。その内容を報告する。

◆『名大会話コーパス』の本文はCEJCと転記の基準が異なり、CJECではタグ（https://www2.ninjal.ac.jp/conversation/cejc/transcript.html）を用い、「発音形」として別に記述するようなものが本文にそのままある。しかしながら、本文の変更・修正はしない、という方針は変えずに、境界と形態論情報の更新をした。
例：CEJCでは「すごーい」は発音形に書き、本文には「すご:い」で転記。

修正方法と修正結果

1. 差分より、一括処理にて修正できるものを特定し、修正した。
例：そうそう（感動詞-一般）→そう（副詞）/そう（副詞）　※CEJCにあわせた
2. 差分のうち、現状と比較して発音形や境界に差のあったもの（感動詞、方言や、名詞-数詞、代名詞）を中心に人手で確認し、修正した。
例：1（じゅう）/9（ここの→く）/日（にち）、何（なん→なに）/が
3. 全体に対し、誤解析の可能性の高いもの（伸ばす発音のある語や、平仮名や片仮名の小文字、算用数字のある語）を中心に人手で確認し、修正した。
例：うっ（感動詞-一般）/そ（副詞）→うっそ（「嘘」名詞-普通名詞-一般）

前回より修正したデータの形態素数は、43,482。全体（記号・補助記号・空白を除いた形態素数：1,131,971）の3.8%にあたる。43,482の形態素について更新した箇所の内容別の件数と合計件数は下記のとおり。

境界更新	発音形更新	語彙素読み更新	語彙素更新	品詞更新	活用型更新	活用形更新	更新合計
8,597	16,225	17,102	21,893	21,996	10,902	18,392	137,103

修正例：感動詞

◆「感動詞-一般」か「感動詞-フィラー」か「連体詞」か

例：「あー」（一般は語彙素：あー、フィラーは語彙素：ああ）
「あ」（一般は語彙素：あっ、フィラーは語彙素：あー）
「あの」（感動詞は語彙素：あの、連体詞は語彙素：彼の）

● 「感動詞-フィラー」を「感動詞-一般」に修正した例　※「mk_」は会話ID。

mk_data109	|それで、しかもお弁当を自分で、昔は支給だったのに、買わなきゃいけない。|【あ】|そうそう、何かね|
mk_data017	|でもそんなに残業しないじゃん。|【あーあーあー】|

● 「感動詞-一般」を「感動詞-フィラー」に修正した例

mk_data089	|あのー、なんだっけ、【えー】、なんだっけ、あの島。|
mk_data009_S003	|例えばだから、【ん】、ノーって言えないためにいろんな失敗したのよね。|

● 「連体詞」を「感動詞-フィラー」に修正した例

mk_data024_S001	|と、B先生って、でも、【あの】、あれですよね、あの、今、H大学。|
mk_data087_S002	|なんていうの、【あの】ねー、お財布取られたときみたいな感じ。|
mk_data074_S001	|千利休ではなくても、なんか、【そのー】千利休ではなくても、なんか、|

◆笑い声の認定

CEJCにあわせ、笑い声に相当する感動詞は原則として3拍の形が「語彙素」。
例：アハ／アハハ／アハハハハ→アハハ【あはは】

● 誤り（赤字）を右側（青字）のように3拍の語彙素に修正した例

mk_data101	ぎゃはは（感動詞-一般）|はは（普通名詞「母」）	ぎゃはははは（感動詞-一般「ぎゃはは」）
mk_data025_S003	あー（感動詞-一般）|はっ（感動詞-一般）|はっ（感動詞-一般）|はっ（感動詞-一般）	あーはっはっはっ（感動詞-一般「あはは」）

修正例：方言

◆ 方言の認定

● 誤り（赤字）を右側（青字）のように修正した例

mk_data070	|あっ、あとねー、とっぽい兄ちゃんや、とっぽい姉ちゃんや、ぼこに送っ【てくりょうし】。|	
	てく（助動詞）/りょうし（「猟師」名詞-普通名詞-一般）	て（助詞-接続助詞）/くりょう（「呉れる」動詞-非自立可能）/し（助詞-終助詞）
mk_data101	|でもうちのおじさんこの間さー、来てさー、その【あきゃー】のって、すごいさ、なんか、わかるんだけどねー。|	
	あ（感動詞-フィラー）/きゃー（感動詞-一般）	あかい（「赤い」形容詞-一般）/ー（補助記号-一般） ※語形「アキャイ」を登録。
mk_data076_S002	|あんなのはそれこそ食べる気が【しなんだ】。|	
	しなん（「至難」名詞-普通名詞-形状詞可能）/だ（助動詞）	し（「為る」動詞-非自立可能）/なんだ（助動詞）

修正例：伸ばす発音のある語

◆長音「ー」or 母音表記が末尾にくる場合

1. UniDicに既存の場合　→　それをあてる　例：ジャアー　→　語彙素「じゃあ」

2. UniDicに未登録の場合

①「ー」以外の部分が語幹などで処理可能の場合　⇒　分割処理
例：おっかしー　・・・　|おっかし[形容詞-語幹]　|ー[補助記号]　|
やめよー　・・・　|やめよ[動詞-意志推量形]　|ー[補助記号]　|

②対応する語はわかるが登録は避けたもの　⇒　品詞「未知語」
例：すっげー　・・・　|すっげー[未知語]|
ひでー　・・・　|ひでー[未知語]|
歩きゃあ　・・・　|歩きゃあ[未知語]|
やらなああかん　・・・　やら|なあ[未知語]|あかん

③解釈困難なもの　⇒　品詞「形態論情報付与対象外」
例：あ、ちゃー　・・・　あ|、|ちゃー[形態論情報付与対象外]|

◆長音「ー」or 母音表記が語中にくる場合

1. UniDicに既存の場合　→　それをあてる　例：しーまっ　→　語彙素「仕舞う」

2. UniDicに未登録の場合

①対応する語はわかるが登録は避けたもの　⇒　品詞「未知語」
例：ましょーう　・・・　|ましょーう[未知語]|　※「ましょー」|ましょ[助動詞]|ー[補助記号]　|
むーり　・・・　|むーり[未知語]|
食べてなあい　・・・　食べ|て|なあい[未知語]|
へえんね　・・・　|へえん[未知語]|ね

②解釈困難なもの　⇒　品詞「形態論情報付与対象外」
例：しびーれ　・・・　|しびーれ[形態論情報付与対象外]|
てーしょん　・・・　|てーしょん[形態論情報付与対象外]|

※新規登録の場合、多くは今後の解析精度に影響を与えないよう、UniDic上ではZ扱い（解析時参照外）で登録した。
例：ちょーっとー　語彙素「一寸」の下に語形「チョオットオ」にし、Z扱いで登録

まとめ

◆『中納言』や『ひまわり』で公開中の『名大会話コーパス』には、誤解析（境界誤り、品詞情報誤り）が多く残っていたため、新しい『話し言葉UniDic』の再解析結果を利用し、修正した。前回より修正したデータの形態素数は、43,482。

◆感動詞、方言、伸ばす発音のある語や、名詞-数詞、代名詞、平仮名や片仮名の小文字、算用数字のある語に誤解析が多くあり、それらを中心に修正した。

◆その結果、「19日」を「ジューココノニチ」と誤っていたようなものを多く修正できた。

◆なお、『名大会話コーパス』の本文はCEJCと転記の基準が異なり、CJECではタグに置き換えたり、「発音形」のところに記述したりするものが本文にある。書き言葉としても出現する可能性が高いと認めたものは新規登録したが、その数は少ない。原則は、UniDic の今後の解析精度に影響を与えないよう登録はせず、「未知語」や「形態論情報付与対象外」とした。

◆そのほか、できるだけ見直しをした。たとえば、「めん/たつ」が「麺/立つ」となっていたが、今回「面接の達人」の略語であることがわかり、「面達」と修正するなども行った。

謝辞　本研究は国立国語研究所の機関拠点型基幹研究プロジェクト「多世代会話コーパスに基づく話し言葉の総合的研究」（プロジェクトリーダー：小磯花絵）の研究成果です。また、本研究の一部にJSPS科研費20K00655の助成を得ています。