『日本語日常会話コーパス』での形態素解析：誤解析箇所の分析

『日本語日常会話コーパス』での形態素解析：誤解析箇所の分析

渡邊 友香（国立国語研究所 音声言語研究領域）
西川 賢哉（国立国語研究所 コーパス開発センター）

Morphological Analysis of *the Corpus of Everyday Japanese Conversation*: An error analysis

Yuka Watanabe (National Institute for Japanese Language and Linguistics)
Ken'ya Nishikawa (National Institute for Japanese Language and Linguistics)

要旨

『日本語日常会話コーパス』(CEJC) の短単位情報付与作業では、次の 4 段階の作業工程、(i) 転記を MeCab（解析器）+ UniDic（解析辞書）で自動解析、(ii) 音声を聴取しながら、付加情報の一つである「発音形」のみを人手修正、(iii) 人手修正された発音形を尊重しつつ再び自動解析、(iv) 短単位情報（境界情報、発音形以外の付加情報）を人手修正、を踏んでいる。今後の(iv) 人手修正作業の参考とするため、人手修正済みデータを対象に、複数の版の現代話し言葉 UniDic(Ver2.2.0, 2.3.0, 3.0.1, 3.1.0) を用いて (i)-(iii) を自動で実施し、その出力と人手修正結果とを比較した。その結果、UniDic の版が新しくなるにつれて誤解析の頻度が低下し、向上が見られたものの、誤りやすい個所がなお残っていることがわかった。特に、品詞が「記号」「代名詞」「接続詞」「名詞-助動詞語幹」「名詞-固有名詞-人名-一般」「名詞-固有名詞-一般」となるべき語は、UniDic の版が新しくなっても別の品詞として解析される、短単位境界を誤るなど、誤解析が起こりやすい。

1. はじめに

国立国語研究所で構築が進められている『日本語日常会話コーパス』(以下 CEJC と呼ぶ) では、短単位情報を付与する作業において、次の 4 段階からなる作業工程を踏んでいる。

(i) 転記を MeCab（解析器）+ UniDic（解析辞書）で自動解析
(ii) 音声を聴取しながら、付加情報の一つである「発音形」のみを人手修正
(iii) 人手修正された発音形を尊重しつつ再び自動解析
(iv) 短単位情報（境界情報、発音形以外の付加情報）を人手修正

上記の作業工程の詳細、ならびに作業工程の妥当性についてはすでに報告を行った（西川・渡邊 2019, 2020）。本稿では、前稿では行わなかった、自動解析結果の誤解析箇所の分析を行う。その分析をもとに、上記 (iv) の人手修正作業において、どのような個所に人的資源を重点的に割くべきかの見通しを立てる。

2. 分析方法

形態論情報誤解析箇所の分析は、短単位人手修正済みデータ（これを正解データとみなす）に対応する転記を対象に、複数の版の現代話し言葉 UniDic を使って、1 節の作業工程 (i)-(iii) を自動で行い、その出力と人手修正の結果とを比較することで実施した。

人手修正済みのデータは第 3 期内部公開用データ（約 61 万短単位）を使用した。CEJC では現在、本公開へ向けてモニター公開版のデータも含む全データでの短単位情報の整備が行われており、ほかのデータと比べて短単位情報の整合性が取れていると判断したため、こ

れを正解データとして選択した。CEJC 全体の約 4 分の 1 に相当するサイズである。

解析に用いる辞書は、以下の 4 種類の UniDic とする（括弧内は公開年月日[1]）：

- 現代話し言葉 UniDic Ver2.2.0（2017 年 9 月 5 日）
- 現代話し言葉 UniDic Ver2.3.0（2018 年 4 月 10 日）
- 現代話し言葉 UniDic Ver3.0.1（2019 年 12 月 17 日）
- 現代話し言葉 UniDic Ver3.1.0（2021 年 4 月 1 日）

これら複数の版の UniDic は、最新版の Ver3.1.0 を除き、実際に CEJC の短単位解析作業で使用されているものである（UniDic Ver3.1.0 は、CEJC での自動形態素解析がすべて終了した後にリリースされたため、実際の作業では使用されていない）。

3. 分析結果

以下に分析結果を示す。

3.1 全体の評価

まず、全体の傾向を見る。小木曽 (2014) に従い「境界」「品詞」「語彙素」の 3 段階の評価基準を設けた。

- 境界：短単位境界の正否
- 品詞：境界 + 短単位の品詞・活用型・活用形が正しく選べたか
- 語彙素：境界・品詞 + 語彙素と語彙素読みが正しいか

ここで、品詞の評価は境界が正しいことを前提としており、語彙素の評価は境界と品詞が正しいことを前提としている。そのため、必ずこの順に厳しい評価となる（小木曽 2014:103）。

UniDic における品詞情報は「名詞-固有名詞-人名-名」や「助詞-終助詞」のように、階層的に表現されるが（最大 4 層）、ここでは第 1 層の大分類（「名詞」「助詞」など）のみを用いた。

上記の評価基準から適合率・再現率・F 値を求めた[2]。図 1 より、最新版の Ver3.1.0 を除く、一番古い版の Ver2.2.0 から最新版の一つ前の版である Ver3.0.1 までで、境界・品詞・語彙素のいずれも増加傾向にあることがわかる。最新版の Ver3.1.0 だけは、境界・品詞・語彙素のいずれも、それまでの版の結果より低い数値が出た。

[1] これらの版の UniDic は、本稿執筆時（2021 年 8 月 7 日時点）で、すべて UniDic のサイトからダウンロードが可能である。

[2] F 値は情報検索システムの性能評価でしばしば用いられる概念で，「一致語数／出力語数」にあたる適合率 (precision) と，「一致語数／正解語数」にあたる再現率 (recall) の調和平均であり（中略）形態素解析の場合、適合率とは解析機器が出力した語数を分母として、出力結果と正解データとが一致した語数を分子としたものであり、再現率とは正解データの語数を分母として、出力結果と正解データとが一致した語数を分子としたものである（小木曽 2014:104）。

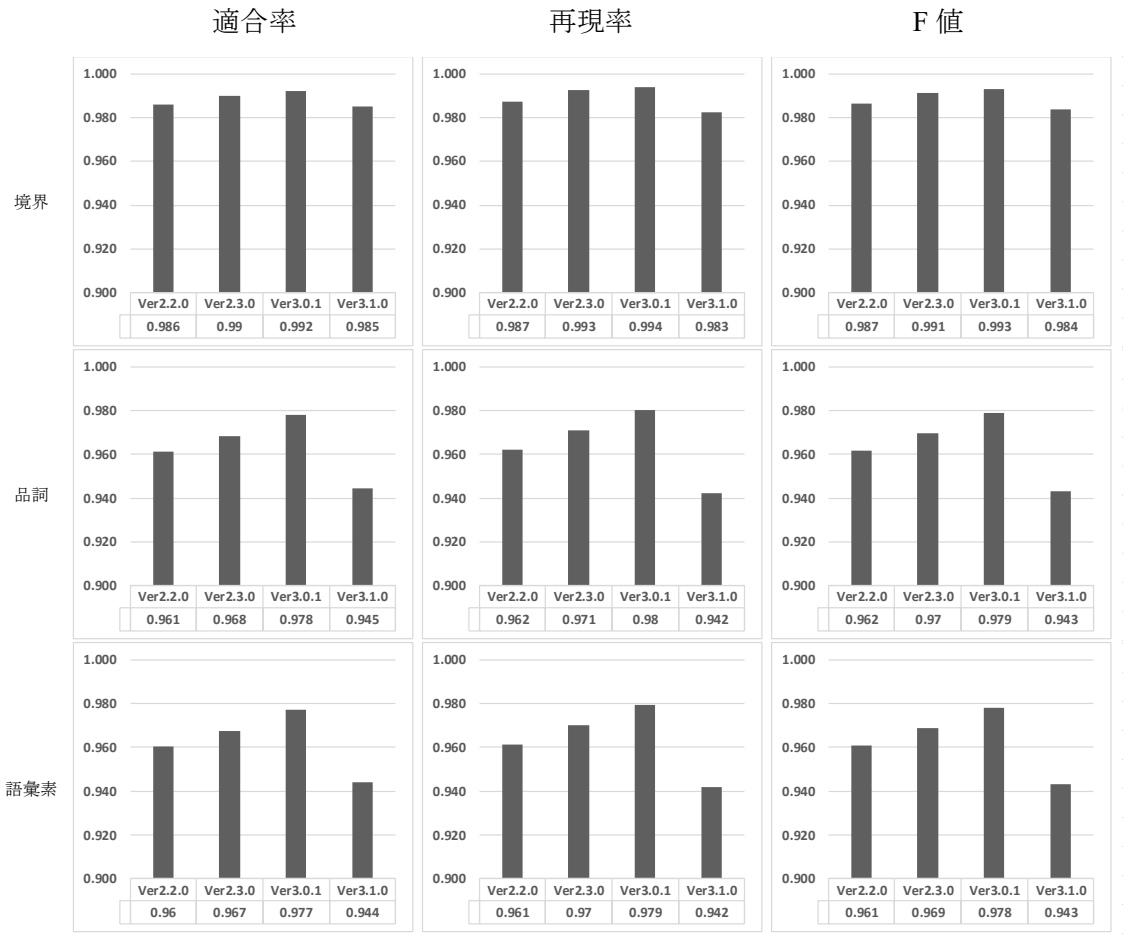


図 1　各版の UniDic での評価基準別による適合率・再現率・F 値

3.2　品詞ごとの評価

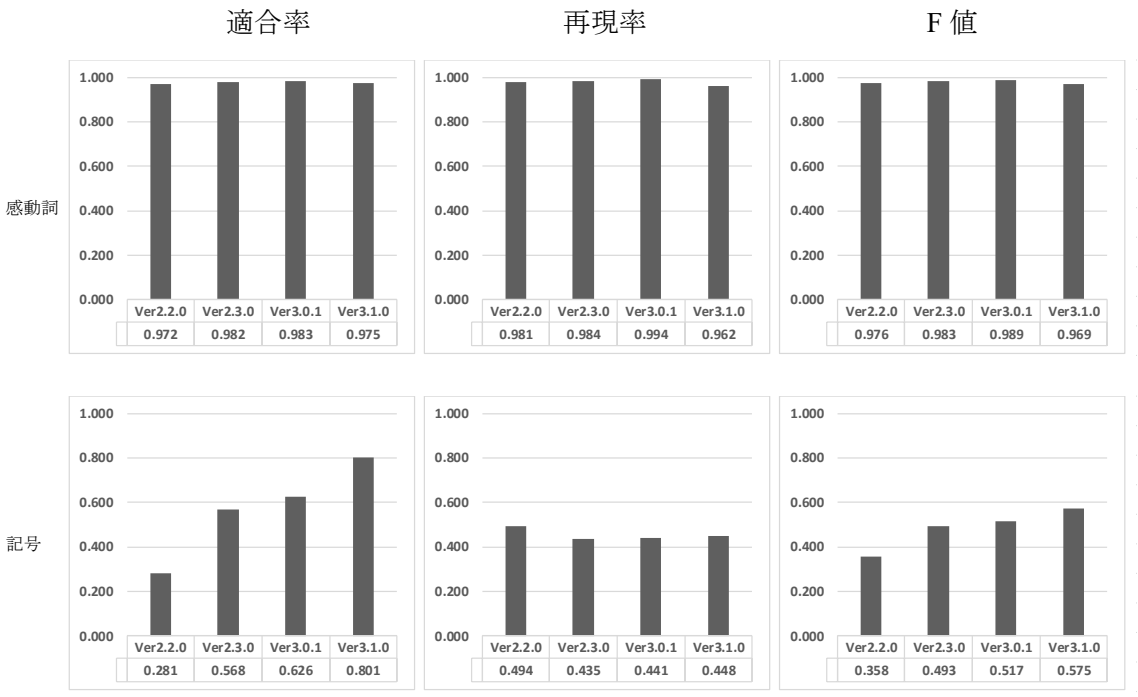
次に、品詞別による適合率・再現率・F 値を示す。ここでは、解析器が出力しない、CEJC 独自の品詞（「言いよどみ」「形態論情報付与対象外」「喃語」「伏せ字」「歌」）は、分析から除外した[3]。

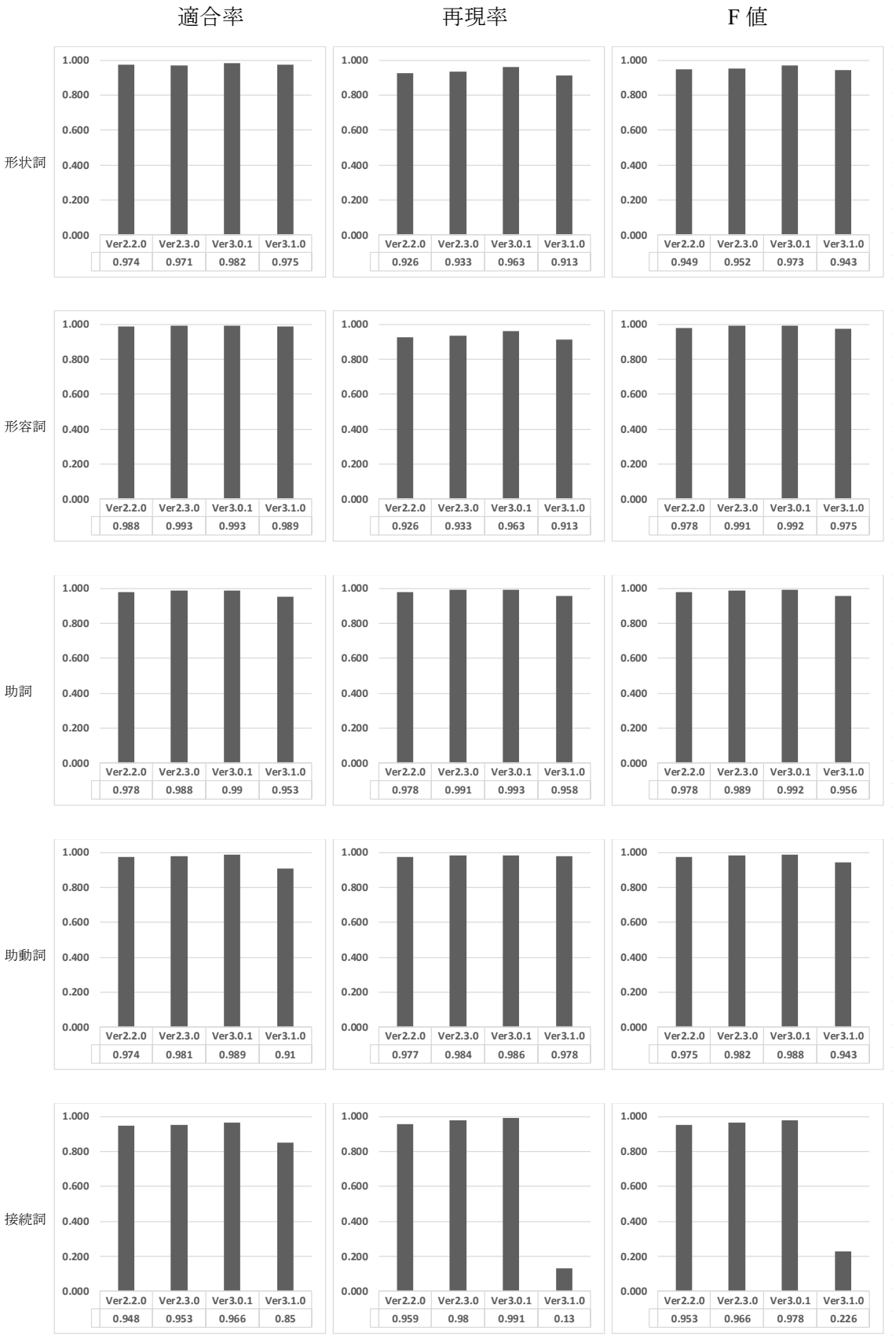
[3] CEJC 独自の品詞が付与される語は、CEJC 転記に与えられる転記タグ（臼田他 2018）から、ほぼ一意に特定できる。例えば、タグ(D)［言いよどみタグ］が付与された要素の品詞は「言いよどみ」となる、といった具合である。

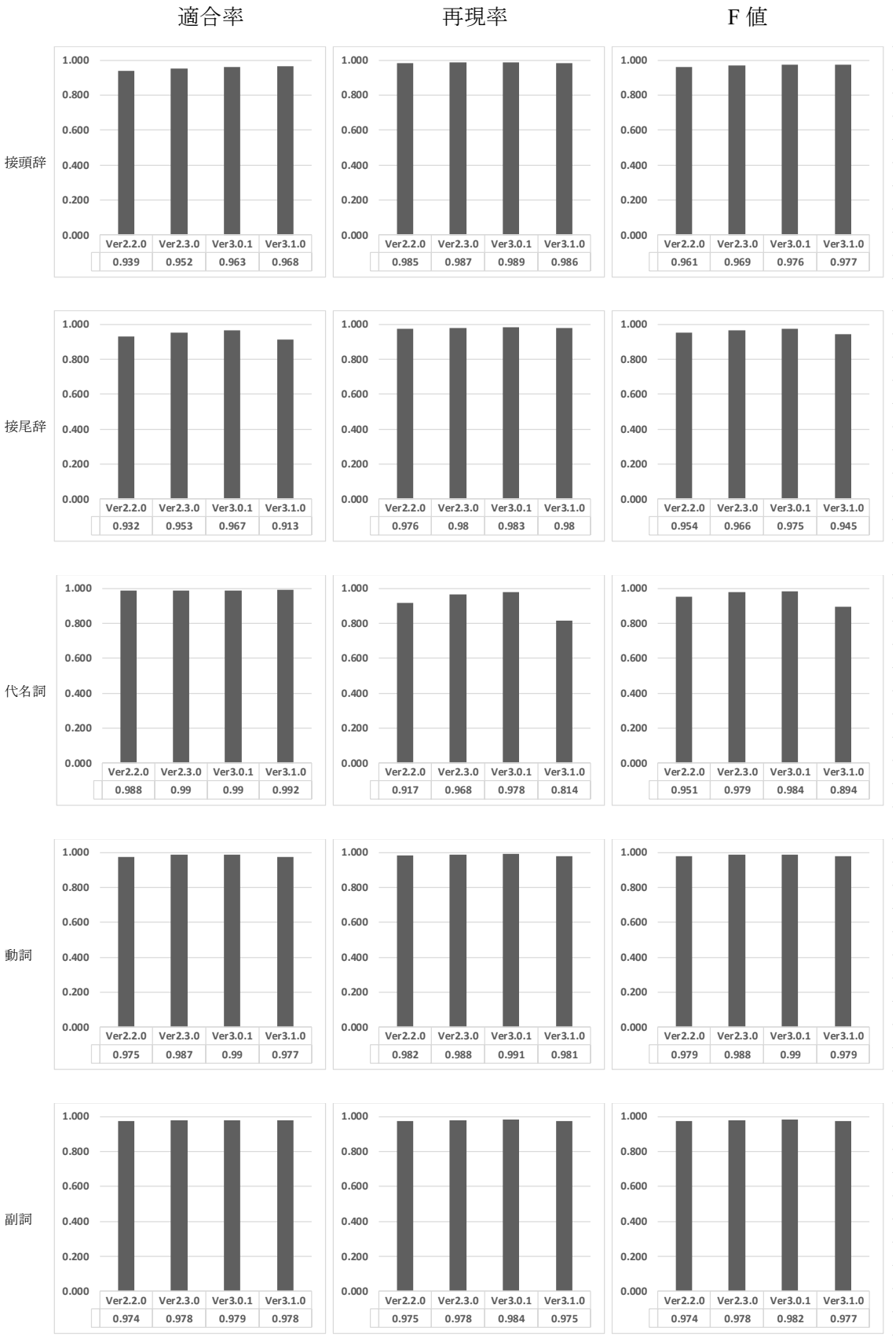
表 1　各版の UniDic での品詞別による出力語数・正解語数・一致語数

		出力語数	正解語数	一致語数
感動詞	Ver2.2.0	63327	62719	61535
	Ver2.3.0	62853	62719	61691
	Ver3.0.1	63389	62719	62336
	Ver3.1.0	61891	62719	60352
記号	Ver2.2.0	693	395	195
	Ver2.3.0	303	395	172
	Ver3.0.1	278	395	174
	Ver3.1.0	221	395	177
形状詞	Ver2.2.0	6405	6741	6239
	Ver2.3.0	6480	6741	6291
	Ver3.0.1	6611	6741	6494
	Ver3.1.0	6314	6741	6157
形容詞	Ver2.2.0	17631	18002	17417
	Ver2.3.0	17948	18002	17817
	Ver3.0.1	17979	18002	17851
	Ver3.1.0	17505	18002	17318
助詞	Ver2.2.0	177002	177054	173175
	Ver2.3.0	177593	177054	175445
	Ver3.0.1	177506	177054	175794
	Ver3.1.0	177898	177054	169589
助動詞	Ver2.2.0	76694	76452	74695
	Ver2.3.0	76762	76452	75266
	Ver3.0.1	76177	76452	75361
	Ver3.1.0	82147	76452	74784
接続詞	Ver2.2.0	4691	4638	4446
	Ver2.3.0	4771	4638	4546
	Ver3.0.1	4759	4638	4596
	Ver3.1.0	712	4638	605

		出力語数	正解語数	一致語数
接頭辞	Ver2.2.0	3742	3566	3513
	Ver2.3.0	3698	3566	3521
	Ver3.0.1	3663	3566	3527
	Ver3.1.0	3632	3566	3516
接尾辞	Ver2.2.0	10422	9952	9718
	Ver2.3.0	10229	9952	9751
	Ver3.0.1	10115	9952	9781
	Ver3.1.0	10680	9952	9753
代名詞	Ver2.2.0	21037	22681	20795
	Ver2.3.0	22165	22681	21951
	Ver3.0.1	22425	22681	22191
	Ver3.1.0	18610	22681	18456
動詞	Ver2.2.0	66328	65876	64688
	Ver2.3.0	65989	65876	65110
	Ver3.0.1	65940	65876	65277
	Ver3.1.0	66134	65876	64596
副詞	Ver2.2.0	36206	36175	35253
	Ver2.3.0	36183	36175	35375
	Ver3.0.1	36351	36175	35601
	Ver3.1.0	36063	36175	35278
名詞	Ver2.2.0	100477	101112	98031
	Ver2.3.0	100798	101112	98544
	Ver3.0.1	100573	101112	98772
	Ver3.1.0	100964	101112	99108
連体詞	Ver2.2.0	5100	4875	4816
	Ver2.3.0	5082	4875	4826
	Ver3.0.1	4908	4875	4792
	Ver3.1.0	5042	4875	4758




適合率
再現率
F 値
形状詞
形容詞
助詞
助動詞
接続詞
1.000
0.800
0.600
0.400
0.200
0.000
Ver2.2.0
Ver2.3.0
Ver3.0.1
Ver3.1.0
0.974 0.971 0.982 0.975
0.926 0.933 0.963 0.913
0.949 0.952 0.973 0.943
0.988 0.993 0.993 0.989
0.926 0.933 0.963 0.913
0.978 0.991 0.992 0.975
0.978 0.988 0.99 0.953
0.978 0.991 0.993 0.958
0.978 0.989 0.992 0.956
0.974 0.981 0.989 0.91
0.977 0.984 0.986 0.978
0.975 0.982 0.988 0.943
0.948 0.953 0.966 0.85
0.959 0.98 0.991 0.13
0.953 0.966 0.978 0.226


適合率
再現率
F 値
接頭辞
接尾辞
代名詞
動詞
副詞
1.000
0.800
0.600
0.400
0.200
0.000
Ver2.2.0
Ver2.3.0
Ver3.0.1
Ver3.1.0
0.939 0.952 0.963 0.968
0.985 0.987 0.989 0.986
0.961 0.969 0.976 0.977
0.932 0.953 0.967 0.913
0.976 0.98 0.983 0.98
0.954 0.966 0.975 0.945
0.988 0.99 0.99 0.992
0.917 0.968 0.978 0.814
0.951 0.979 0.984 0.894
0.975 0.987 0.99 0.977
0.982 0.988 0.991 0.981
0.979 0.988 0.99 0.979
0.974 0.978 0.979 0.978
0.975 0.978 0.984 0.975
0.974 0.978 0.982 0.977

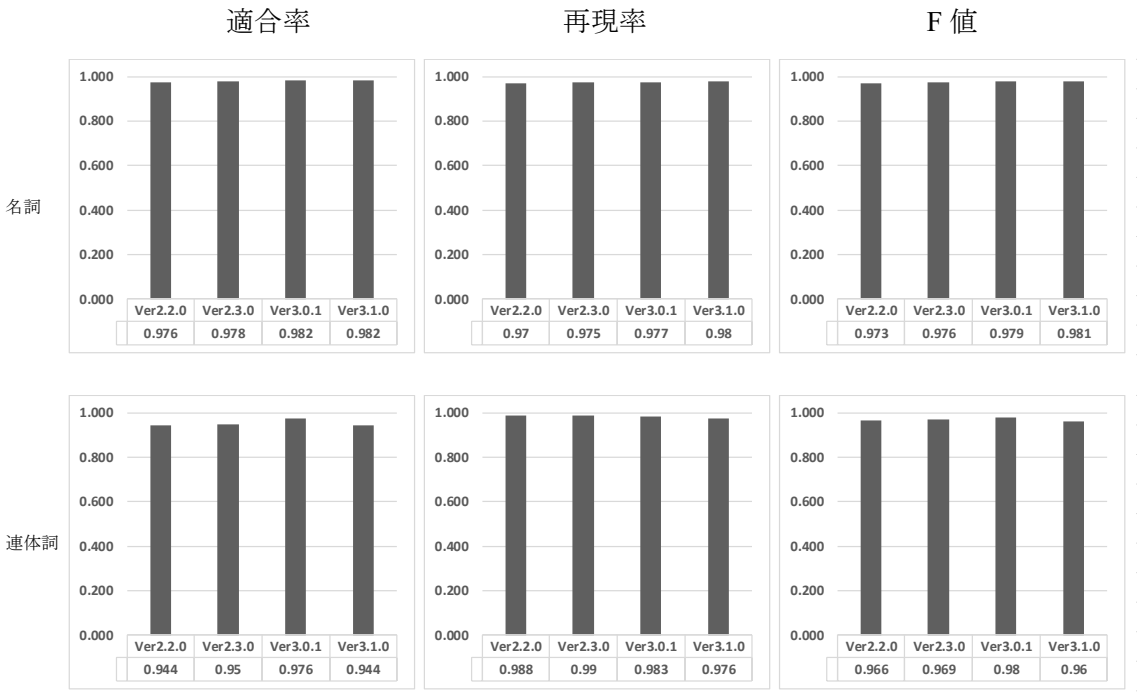


図 2　品詞別による各版の UniDic での適合率・再現率・F 値

図 2 より、最新版の Ver3.1.0 を除き、一番古い版の Ver2.2.0 から最新版の一つ前の版の Ver3.0.1 にかけて、どの品詞においても F 値の値が増加傾向にあることがわかった。中でも「記号」は、そもそもの出力語数が、一番古い版の Ver2.2.0 では 693 個、中間の版の Ver2.3.0 では 303 個、最新版の一つ前の版の Ver3.0.1 では 278 個、最新版の Ver3.1.0 では 221 個と、ほかの品詞と比べると極めて少ない（表 1 参照）。正解語数や一致語数も、正解語数は 395 個、一致語数は一番古い版から最新版にかけて 195 個、172 個、174 個、177 個と少なさが際立つ結果となっている。しかし、一番古い版から最新版の一つ前の版までの間で 0.358→0.517 と 0.159 ポイント増加しており、14 品詞の中で最も増加傾向が強く表れている。その次に増加傾向が高かったのが、「代名詞」の 0.033 ポイント (0.951→0.984)、「接続詞」の 0.025 ポイント (0.953→0.978) だった。UniDic の版が更新されるにつれて増加傾向が顕著に表れたということは、これら 3 品詞は誤解析が起きやすいことを意味している。増加の幅はどれも 1 ポイント未満ではあるものの、人的資源をかけて修正を行うべき語は、品詞が「記号」「代名詞」「接続詞」となる語であるといえる。

その一方で、最新版の Ver3.1.0 を除いて、一番古い版の Ver2.2.0 から最新版の一つ前の版である Ver3.0.1 までの間で、F 値の増加は見られたものの、先に述べた 3 品詞よりその傾向が低いものがあることもわかった。先に述べた 3 品詞を除く 11 品詞の中で、最も傾向が低かったのが「名詞」で、一番古い版から最新版の一つ前の版までで 0.973→0.979 と、0.006 ポイントしか増加していなかった。「副詞」や「動詞」についても、「副詞」は 0.008 ポイント (0.974→0.982)、「動詞」は 0.011 ポイント (0.979→0.99) しか増えていなかった。これら 3 品詞は、増加傾向が顕著に表れた「記号」「代名詞」「接続詞」と比べると、最新版以外の版の UniDic にて、新古を問わず F 値が高い傾向にある。もともと誤解析を起こしにくいものと推定され、すなわち品詞が「名詞」「副詞」「動詞」となる語は、人的資源をそこまでか

けずともよい語であるといえる。そのほかの品詞についても、増加傾向は「記号」「代名詞」「接続詞」ほど強く出ておらず、この結果からも人的資源を重点的にかけるべき語は、「記号」「代名詞」「接続詞」の 3 品詞であることがわかる。

なお、最新版の Ver3.1.0 の F 値については、適合率・再現率の値が低かっただけに、それまでの版より値が低い結果が出た。出力語数や一致語数の差が関係しているとは考えづらく、最新版だけ結果が低く出た原因を特定できていない。今後の課題である。

3.3 品詞「名詞」での精密評価

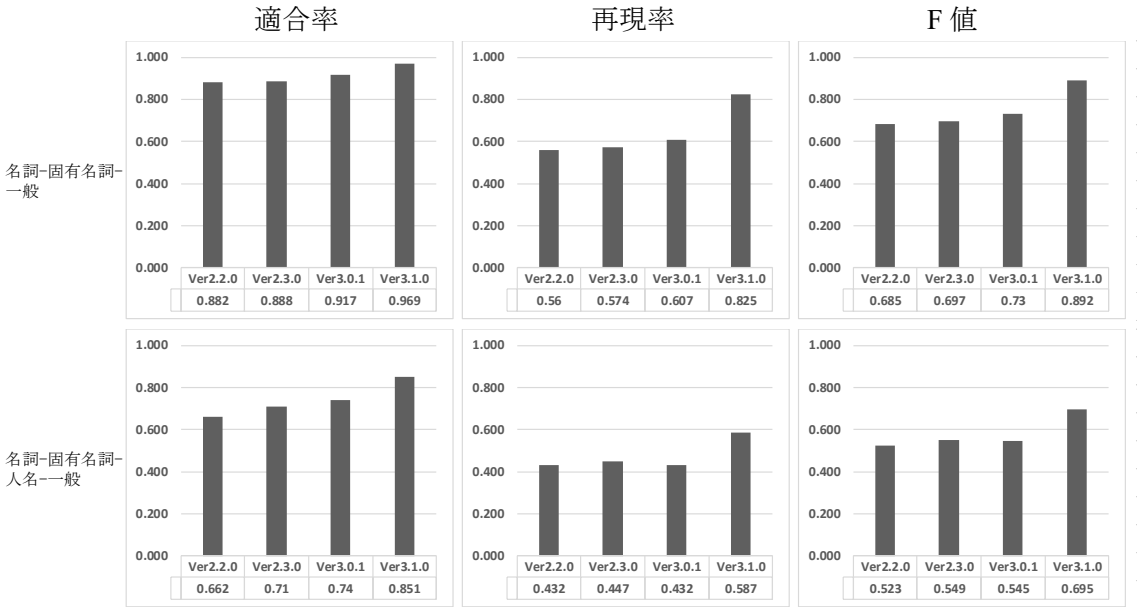
3.2 節にて、人的資源を重点的にかけるべき語は、品詞が「記号」「代名詞」「接続詞」となる語であり、反対にあまり必要としない語は、品詞が「名詞」「副詞」「動詞」となる語であることが明らかとなった。殊に「名詞」は、ほかの品詞より誤解析を最も起こしにくいと述べたが、これは果たして妥当なのだろうか。「名詞」と一口にいっても、その中には固有名詞や普通名詞、数詞が含まれており、固有名詞の中には人名や地名が、普通名詞の中にはサ変可能や副詞可能などの細かな分類がある。品詞が「名詞」となる語は本当に人的資源を要さない語であるのか、単に「名詞」と一括りにするのではなく、「名詞」の中身をより詳細に確認する必要がある。そのため本節では、品詞が「名詞」であるものについて、小椋・小磯他 (2011:53) をもとに品詞情報の第 4 層である小分類（以下の 14 個）を用いて、それぞれで適合率・再現率・F 値を割り出し、「名詞」が人的資源を必要としない語であるのか考察する。

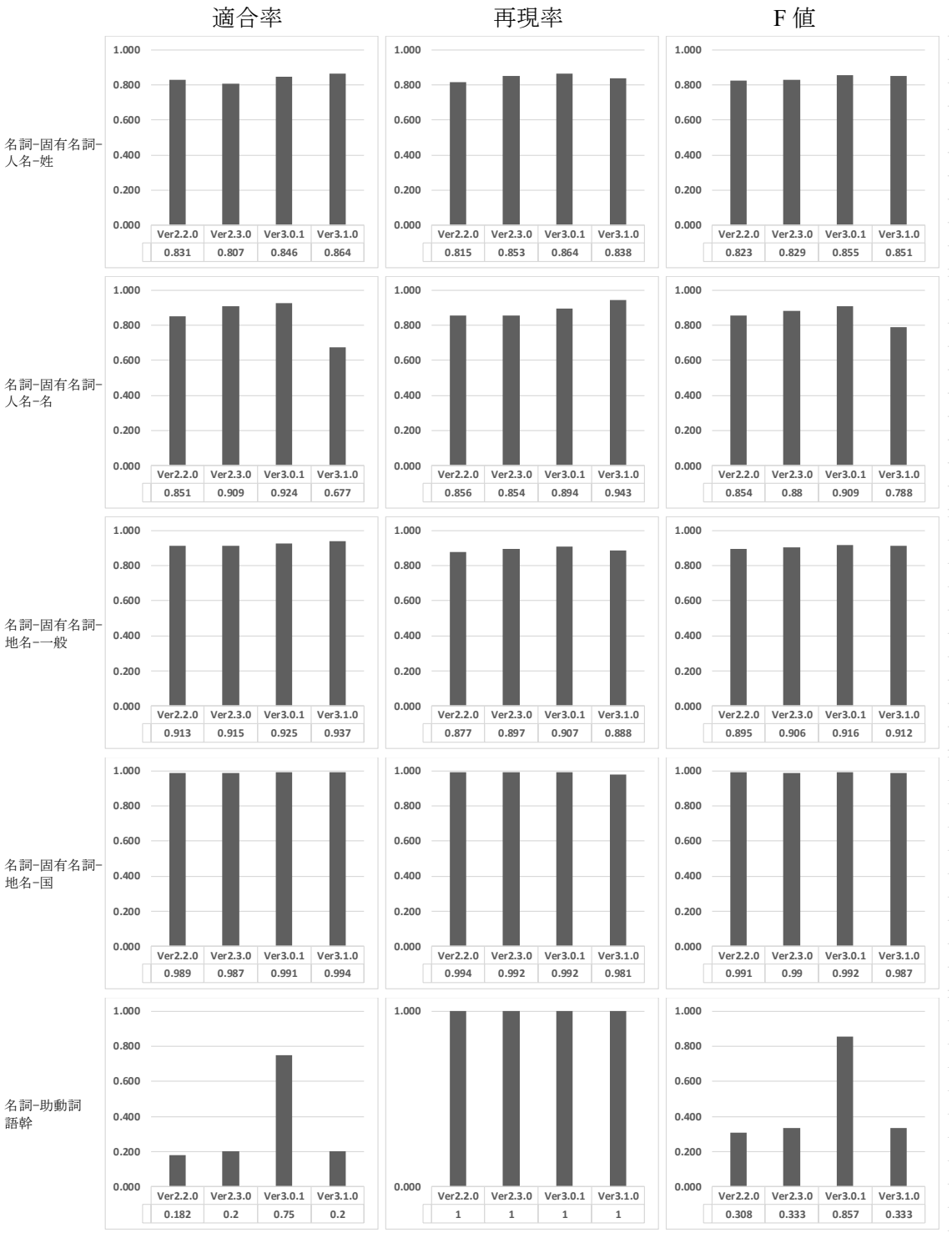
名詞-固有名詞-一般、名詞-固有名詞-人名-一般、名詞-固有名詞-人名-姓、
名詞-固有名詞-人名-名、名詞-固有名詞-地名-一般、名詞-固有名詞-地名-国、
名詞-助動詞語幹、名詞-数詞、名詞-普通名詞-サ変可能、
名詞-普通名詞-サ変形状詞可能、名詞-普通名詞-一般、名詞-普通名詞-形状詞可能、
名詞-普通名詞-助数詞可能、名詞-普通名詞-副詞可能

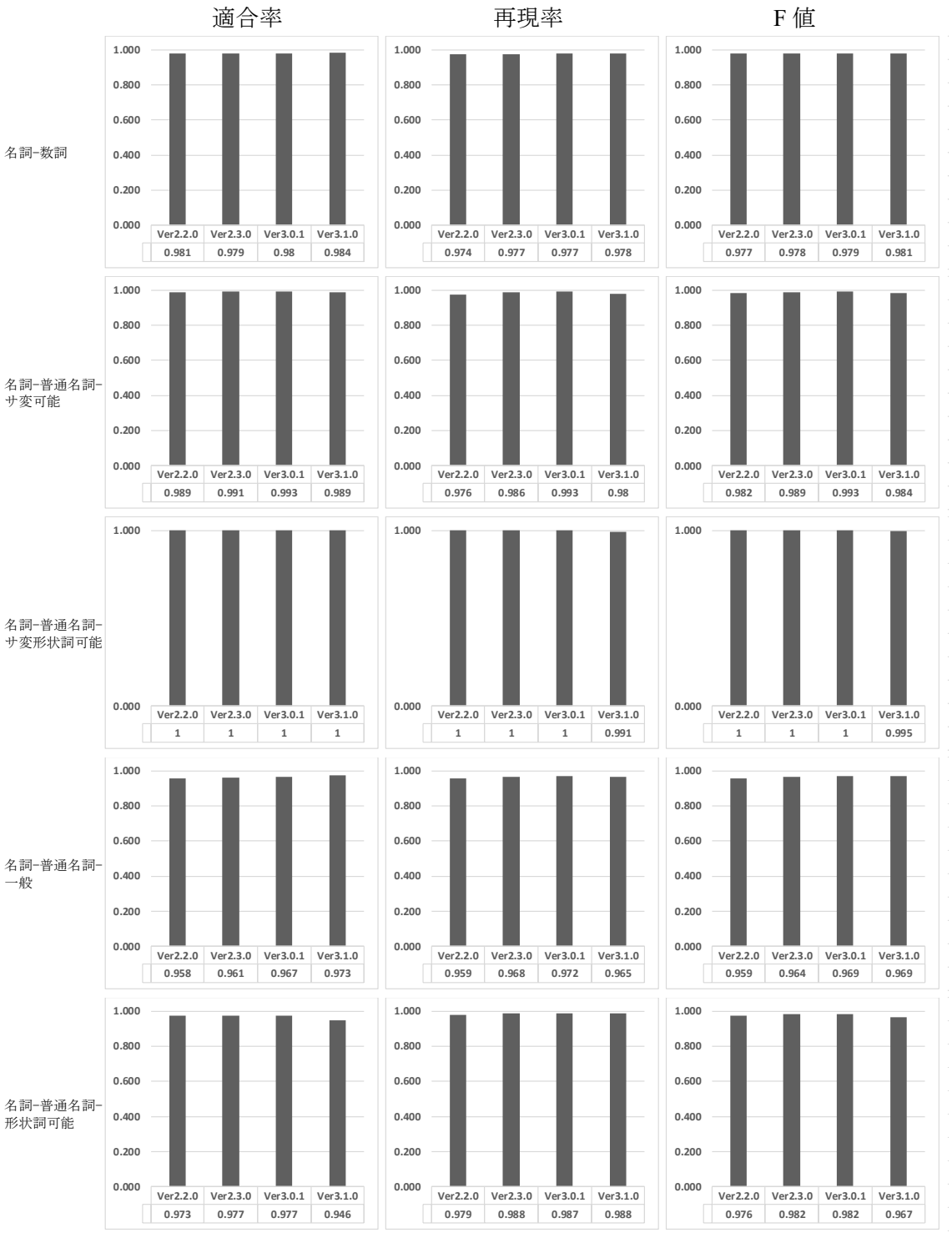
表 2　品詞「名詞」の内訳（出力語数・正解語数・一致語数）

		出力語数	正解語数	一致語数
名詞-固有名詞-一般	Ver2.2.0	654	1031	577
	Ver2.3.0	667	1031	592
	Ver3.0.1	683	1031	626
	Ver3.1.0	878	1031	851
名詞-固有名詞-人名-一般	Ver2.2.0	627	961	415
	Ver2.3.0	606	961	430
	Ver3.0.1	561	961	415
	Ver3.1.0	663	961	564
名詞-固有名詞-人名-姓	Ver2.2.0	1073	1094	892
	Ver2.3.0	1156	1094	933
	Ver3.0.1	1117	1094	945
	Ver3.1.0	1061	1094	917
名詞-固有名詞-人名-名	Ver2.2.0	1360	1353	1158
	Ver2.3.0	1271	1353	1155
	Ver3.0.1	1308	1353	1209
	Ver3.1.0	1886	1353	1276
名詞-固有名詞-地名-一般	Ver2.2.0	2415	2515	2205
	Ver2.3.0	2468	2515	2257
	Ver3.0.1	2464	2515	2280
	Ver3.1.0	2382	2515	2233
名詞-固有名詞-地名-国	Ver2.2.0	799	795	790
	Ver2.3.0	799	795	789
	Ver3.0.1	796	795	789
	Ver3.1.0	785	795	780
名詞-助動詞語幹	Ver2.2.0	33	6	6
	Ver2.3.0	30	6	6
	Ver3.0.1	8	6	6
	Ver3.1.0	30	6	6

		出力語数	正解語数	一致語数
名詞-数詞	Ver2.2.0	8783	8845	8614
	Ver2.3.0	8830	8845	8641
	Ver3.0.1	8819	8845	8644
	Ver3.1.0	8792	8845	8653
名詞-普通名詞-サ変可能	Ver2.2.0	10125	10259	10013
	Ver2.3.0	10205	10259	10118
	Ver3.0.1	10258	10259	10185
	Ver3.1.0	10163	10259	10050
名詞-普通名詞-サ変形状詞可能	Ver2.2.0	433	433	433
	Ver2.3.0	433	433	433
	Ver3.0.1	433	433	433
	Ver3.1.0	429	433	429
名詞-普通名詞-一般	Ver2.2.0	51986	51910	49796
	Ver2.3.0	52321	51910	50265
	Ver3.0.1	52170	51910	50435
	Ver3.1.0	51475	51910	50087
名詞-普通名詞-形状詞可能	Ver2.2.0	3572	3551	3476
	Ver2.3.0	3591	3551	3507
	Ver3.0.1	3590	3551	3506
	Ver3.1.0	3709	3551	3510
名詞-普通名詞-助数詞可能	Ver2.2.0	4875	5028	4793
	Ver2.3.0	4925	5028	4818
	Ver3.0.1	4908	5028	4830
	Ver3.1.0	4944	5028	4868
名詞-普通名詞-副詞可能	Ver2.2.0	13742	13331	13227
	Ver2.3.0	13496	13331	13216
	Ver3.0.1	13458	13331	13218
	Ver3.1.0	13767	13331	13230




適合率
再現率
F 値
名詞-固有名詞-人名-姓
名詞-固有名詞-人名-名
名詞-固有名詞-地名-一般
名詞-固有名詞-地名-国
名詞-助動詞
語幹
1.000
0.800
0.600
0.400
0.200
0.000
Ver2.2.0 Ver2.3.0 Ver3.0.1 Ver3.1.0
0.831 0.807 0.846 0.864
0.815 0.853 0.864 0.838
0.823 0.829 0.855 0.851
0.851 0.909 0.924 0.677
0.856 0.854 0.894 0.943
0.854 0.88 0.909 0.788
0.913 0.915 0.925 0.937
0.877 0.897 0.907 0.888
0.895 0.906 0.916 0.912
0.989 0.987 0.991 0.994
0.994 0.992 0.992 0.981
0.991 0.99 0.992 0.987
0.182 0.2 0.75 0.2
1 1 1 1
0.308 0.333 0.857 0.333


適合率
再現率
F 値
名詞-数詞
名詞-普通名詞-サ変可能
名詞-普通名詞-サ変形状詞可能
名詞-普通名詞-一般
名詞-普通名詞-形状詞可能
1.000
0.800
0.600
0.400
0.200
0.000
Ver2.2.0 Ver2.3.0 Ver3.0.1 Ver3.1.0
0.981 0.979 0.98 0.984
0.974 0.977 0.977 0.978
0.977 0.978 0.979 0.981
0.989 0.991 0.993 0.989
0.976 0.986 0.993 0.98
0.982 0.989 0.993 0.984
1 1 1 1
1 1 1 0.991
1 1 1 0.995
0.958 0.961 0.967 0.973
0.959 0.968 0.972 0.965
0.959 0.964 0.969 0.969
0.973 0.977 0.977 0.946
0.979 0.988 0.987 0.988
0.976 0.982 0.982 0.967

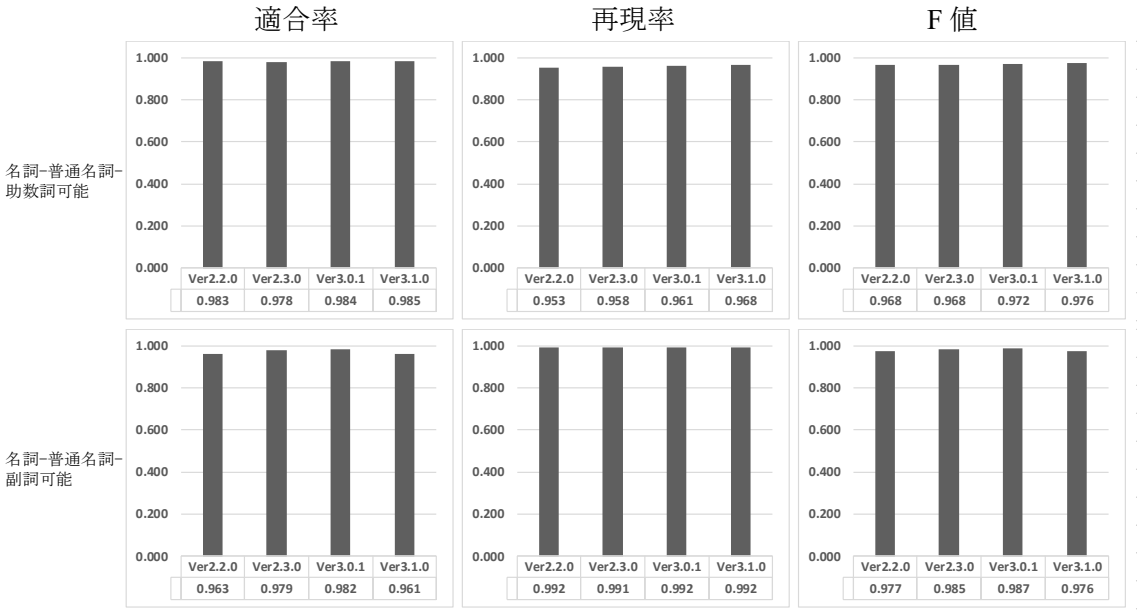


図 3　品詞「名詞」の内訳（適合率・再現率・F 値）

表 2 は、品詞「名詞」の細分類を出力語数・正解語数・一致語数で比較したもので、図 3 は、表 2 の結果を受けて、細分類別で適合率・再現率・F 値を割り出したものである。

表 2 ならびに図 3 より、3.2 節で人的資源を必要としないと述べた品詞「名詞」には、一番古い版の Ver2.2.0 から最新版の Ver3.1.0 にかけて、F 値の値が全体的に低く出たものと、そうでないものの 2 種類に分かれていることがわかった。

F 値の値が低く出たのは、最も値が低かったものから順に、「名詞-助動詞語幹」「名詞-固有名詞-人名-一般」「名詞-固有名詞-一般」の 3 品詞だった。「名詞-助動詞語幹」は、以下の例 1・例 2：

例 1：安室ちゃんがなんか全部撮り直したそうなんですよね。(K008_004b-IC02)
例 2：浜名湖のほうにあるそうですよ。(T020_008-IC02)[4]

のように、「そう（伝聞）」が使われた際に解析される品詞だが、表 2 の正解語数を見てもわかるとおり、そもそもの数値が極めて少ない。CEJC では、同じ書字形で品詞が「副詞」となる「そう」が多数出現し、それと混同しやすいことから、F 値の値が極めて低くなったものと推察される。

「名詞-固有名詞-人名-一般」と「名詞-固有名詞-一般」については、前者には外国人名のほかに日本名由来の愛称（「しーちゃん」の「しー」、「やっくん」の「やっ」など）が、後者には企業名や商品名、ペット名などが含まれている。これらは CEJC に収録されている日常会話では頻出する（もしくは多用される傾向にある）語と考えられ、そうであれば、UniDic にとって未知語である可能性が高い。機械による自動形態素解析を行った際に、未知語で解析される、もしくは「名詞-固有名詞-人名-一般」や「名詞-固有名詞-一般」以外の品詞に誤

[4] CEJC からの引用にあたっては、それが出現した会話 ID と話者 ID を（ハイフンで結んで）末尾に示す。また、着目している個所を示すため、適宜下線を施す。

解析されやすく、そのため F 値の値が低く出たと推測される。

その一方で、一番古い版の Ver2.2.0 から最新版の Ver3.1.0 にかけて F 値の結果が全体的に高く出たもののうち、UniDic の版が新しくなるにつれて値が増加傾向にあるのは、「名詞-数詞」「名詞-普通名詞-一般」「名詞-普通名詞-助数詞可能」の 3 品詞であることがわかった。「名詞-数詞」は 0.004 ポイント (0.977→0.981)、「名詞-普通名詞-助数詞可能」については 0.008 ポイント (0.968→0.976)、「名詞-普通名詞-一般」は 0.01 ポイント (0.959→0.969) 増加した。増加の幅は微小だが、どの版でも高い数値を保持していることから、これら 3 品詞は誤解析を起こしにくい語であるといえる。

またそれら 3 品詞とは別に、「名詞-普通名詞-サ変形状詞可能」についても、最新版の Ver3.1.0 を除いて、表 2 の出力語数・正解語数・一致語数が同値となり、かつ図 3 の F 値の結果も 1 ポイントと、ほかの品詞と比べて誤解析を起こしていないことがわかる。このことから、人的資源を要さない語は「名詞」全般ではなく、さらにその中の「名詞-数詞」「名詞-普通名詞-一般」「名詞-普通名詞-サ変形状詞可能」「名詞-普通名詞-助数詞可能」の 4 品詞であると断定することができる。

4. まとめ

本稿では、今後の人手修正作業の参考とするため、人手修正済みデータを用いて複数の版の現代話し言葉 UniDic にてどのような語（品詞）で誤解析が起きているのか、分析を行った。その結果、3.2 節で述べたとおり、品詞が「記号」「代名詞」「接続詞」になる語は、UniDic の版が新しいものになるにつれて誤解析の頻度は減少しているものの、いまだに誤解析を起こしやすいことが明らかとなった。また、3.2 節の結果から人的資源をかけなくともよい語は「名詞」であると思われたが、3.3 節においてそれは、「名詞」の中でも「名詞-数詞」「名詞-普通名詞-一般」「名詞-普通名詞-サ変形状詞可能」「名詞-普通名詞-助数詞可能」の 4 品詞に絞られること、「名詞-助動詞語幹」「名詞-固有名詞-人名-一般」「名詞-固有名詞-一般」の 3 品詞については、人の目を介した確認・修正作業がある程度必要であることも明らかとなった。

これら「名詞」のうちの 3 品詞と、先に述べた品詞が「記号」「代名詞」「接続詞」となる語については、1 節で示した CEJC の作業工程のうち (iv) 人手修正作業にて重点的に確認・修正をする必要がある。そして、その作業で修正されたものと、機械による自動形態素解析で正しく解析されたものとを合わせることで、コーパスを一般公開した際に、より高度で正確な情報を提供することができると期待される。なお、最新版の Ver3.1.0 でほかの版と異なる結果が出た原因については特定できなかった。今後の課題である。

謝　辞

本研究は国立国語研究所共同プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「大規模日常会話コーパスに基づく話し言葉の多角的研究」の成果である。

文　献

臼田泰如・川端良子・西川賢哉・石本祐一・小磯花絵 (2018) 「『日本語日常会話コーパス』における転記の基準と作成手法」『国立国語研究所論集』15, pp. 177–193.（http://doi.org/10.15084/00001602 よりダウンロード可能）

小木曽智信 (2014) 「形態素解析」山崎誠（編）『書き言葉コーパス：設計と構築』（講座日本語コーパス 2）. 朝倉書店, pp.89-115

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011) 「『現代書き言葉均衡コーパス』形態論情報規定集第 4 版（下）」特定領域研究「日本語コーパス」平成 22 年度研究成果報告書 (JC-D-10-05-02)（https://repository.ninjal.ac.jp/?action=repository_uri&item_id=2872&file_id=22&file_no=1 よりダウンロード可能）

西川賢哉・渡邊友香 (2019) 「『日本語日常会話コーパス』の短単位解析：作業工程を中心に」『言語資源活用ワークショップ発表論文集』4, pp.238-250.（http://doi.org/10.15084/00002575 よりダウンロード可能）

西川賢哉・渡邊友香(2020)「『日本語日常会話コーパス』に対する短単位情報付与：作業工程と評価」『言語資源活用ワークショップ発表論文集』5, pp.324-330.（http://doi.org/10.15084/00003172 よりダウンロード可能）

関連 URL

大規模日常会話コーパスに基づく多角的研究	https://www2.ninjal.ac.jp/conversation/
UniDic	https://unidic.ninjal.ac.jp/
MeCab	https://taku910.github.io/mecab/