形態素解析器『Sudachi』のための大規模辞書開発

著者	坂本 美保, 川原 典子, 久本 空海, ?岡 一馬, 内田 佳孝
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	118-129
発行年	2018
URL	http://doi.org/10.15084/00001644

形態素解析器『Sudachi』のための大規模辞書開発

坂本 美保，川原 典子，久本 空海，髙岡 一馬，内田 佳孝
（株式会社ワークスアプリケーションズ　ワークス徳島人工知能 NLP 研究所）

Large Scale Dictionary Development for Sudachi

Miho Sakamoto, Noriko Kawahara, Sorami Hisamoto,
Kazuma Takaoka, Yoshitaka Uchida
(WAP Tokushima Laboratory of AI and NLP)

要旨

我々は，汎用的な日本語形態素解析器『Sudachi』とその辞書を開発した。本稿では，Sudachi の辞書開発内容について述べる。我々は，まず，UniDic をベースとして，見出し表記，品詞，各種パラメータ等，形態素解析をするための辞書情報を整えた。次に，実用上 UniDic に不足している語句を見出しとして追加した。これには，NEologd から取り込んだ膨大な固有名称も含まれる。さらに，登録見出しについて，アプリケーションが利用しやすい形態素単位の整備，表記のゆれを同一視するための正規化表記の整備等を行い，辞書内容を充実させた。また，形態素解析精度の向上のため，UniDic 由来の見出しについても，弊害となる見出しの抑制や間違いの修正，形態素単位の調整を行った。我々のこれまでの成果は，最新版の辞書ソースに反映し OSS として公開している。

１．はじめに

IT 技術の進展により，近年，産業界において日本語テキスト処理の利用機会はますます増えている。形態素解析はテキスト処理の重要な基盤技術であるが，自由に利用できて，かつ有用な形態素解析リソースは不足している[1]。

商用利用される形態素解析器としては，OSS として公開されている MeCab[2] (Kudo et al.,2004)，kuromoji[3]が大半を占めており，これらで利用可能な辞書としては，IPAdic (Asahara and Matsumoto, 2003)，NAIST Japanese Dictionary[4]，UniDic[5] (Den et al., 2007; Den et al., 2008)，NEologd[6] (Sato et al., 2016; Sato et al., 2017) などがある。しかし，IPAdic，NAIST Japanese Dictionary は，長年メンテナンスされていないため辞書内容が最新でない。

1 http://www.lrec-conf.org/proceedings/lrec2018/pdf/8884.pdf , pp. 1-2.
2 http://taku910.github.io/mecab/
3 https://www.atilika.com/ja/kuromoji/
4 https://ja.osdn.net/projects/naist-jdic/
5 http://unidic.ninjal.ac.jp/
6 https://github.com/neologd

また，UniDic，NEologd は，登録見出しの単位に特徴があり，用途によっては，そのままでは使いにくい。UniDic では，言語の形態論的側面に着目して規定された短単位[7]で見出し登録されている。そのため，たとえば語義を取り扱いたい場合や語彙調査をする場合にはそのままでは不足が生じる。一方，NEologd では，複数の短単位から成る固有表現が一塊で登録されているため，そのまま検索システムで利用すると再現率が低くなる等，支障がある[8]。

我々は，汎用的な辞書として使用できる大規模かつ高品質の辞書データの構築を目指す。

２．Sudachi 辞書の特長

２．１　豊富な語彙

『現代日本語書き言葉均衡コーパス』（BCCWJ）(Maekawa et al., 2014) の形態論情報をアノテーションするために開発された UniDic には，様々なジャンルの語句が斉一な単位で登録されている。見出し数は 75 万語を超える[9]。しかし，UniDic で規定するところの短単位で登録されているため，基本的だと感じる語句が見出し登録されていないことがある。たとえば，「小学校」や「自転車」など日常生活に密着した語句や，「太平洋」「東京都」などの地名，「集英社」「サランラップ」など認知度の高い固有名称も登録されていない。また，「ゆるキャラ」や「スマホ」など，新語への対応も不十分である。

そこで，我々は，新語や固有表現の収集を高頻度で行っている NEologd から，語句を大量に追加した。その際，NEologd の見出しの内部構造に含まれる複合語で，共通して他の見出しにも含まれる複合語（主に接辞付きの語句）についてもあわせて追加した[10]。
表 1 に例を示す。

表 1　NEologd からの見出し追加

NEologd の見出し	Sudachi に登録した見出し	
自転車シェアリング	自転車シェアリング	自転車
自転車通勤	自転車通勤	自転車
不動産登記	不動産登記	不動産
不動産鑑定士	不動産鑑定士	不動産，鑑定士
古民家鑑定士	古民家鑑定士	古民家，鑑定士
古民家再生協会	古民家再生協会	古民家

7　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf
8　http://www.anlp.jp/proceedings/annual_meeting/2017/pdf_dir/B6-1.pdf
9　我々が採用したバージョンは，unidic-mecab-2.1.2_src.zip
（http://unidic.ninjal.ac.jp/back_number#unidic_cwj）である。
10　接辞付きの語句については，国語辞典から登録したものもある。

これにより，UniDic 由来の短単位語句に加え，NEologd 由来の膨大な固有名称，およびその内部構造である接辞付き語句等を補強し，辞書見出しを充実させることができた（付録 A 参照）。

２．２　３種類の形態素単位

上述のように，登録見出しについては膨大な量を確保できたが，次に，これらを用いた形態素解析結果としてどの長さで形態素認定すべきかを検討しなくてはならない。最適な形態素の長さは，アプリケーションによって異なるからである。たとえば，NEologd 由来の長い単位の語句は，固有表現抽出やテキストマイニングには有利だが，検索システムで使う場合には，短い単位でもインデキシングしないと再現率が低下する等，不都合な面がある。

この問題に対処するため，Sudachi (Takaoka et al., 2018)には，３種類の形態素単位（A 単位（短単位），B 単位（中単位），C 単位（NE 単位））が用意されている。

「A 単位」とは，ほぼ UniDic の短単位規定[11]と同じであるが，次項で述べるように，一部のものについて，さらに短くしたものを「A 単位」としている。「B 単位」とは，「A 単位」に接辞および漢字１文字の名詞[12]が結合したもの，および，複合動詞[13]である。「C 単位」とは，さらに多くの語句が結合したもので，複合名詞や固有名称，慣用句などがこれに相当する。各アプリケーションは，解析時に，これらの中から形態素単位を選択することができる。

我々は，３種類の形態素単位を提供するため，分割情報のアノテーションを行った。分割情報とは，上記３種類の形態素単位の規定に基づき，登録見出しの内部構造を記述し辞書に格納したものである。表 2 に，分割情報とそれに基づいて認定される形態素を示す。

表 2　分割情報と３種類の形態素解析結果

Sudachi の登録見出し	A 単位	B 単位	C 単位
選挙／管理／委員会	選挙｜管理｜委員｜会	選挙｜管理｜委員会	選挙管理委員会
委員／会	委員｜会	委員会	委員会
カンヌ／国際／映画祭	カンヌ｜国際｜映画｜祭	カンヌ｜国際｜映画祭	カンヌ国際映画祭
映画／祭	映画｜祭	映画祭	映画祭

”／”…分割情報として保持している見出しの内部構造の境界，　”｜”…形態素解析結果における形態素境界（以下，同記号を用いる）

11　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf
12　「～案（アン）」「～法（ホウ）」のように様々な名詞の下につく漢字１文字の名詞を指す。
13　「動詞+動詞」で構成されるもののみを「B 単位」とした。

２．２．１　UniDic 由来の見出しの再分割

UniDic では，「和語・漢語は、２最小単位の１次結合体を１短単位とする。」「外来語は、１最小単位を１短単位とする。」[14]という短単位規定に基づいて，見出し登録されている。

しかし和語については，漢語と異なり最小単位[15]が自立語として使用されることが多いため，最小単位で形態素認定した方が実用上都合がよいものがある。たとえば，「夏頃」は，UniDic では規定通り「短単位」だが，検索での利用を想定した場合，「夏」単独でも検索キーワードとして使われる可能性が高い。

そこで，我々は，UniDic 由来の見出しについて，次のようなものに分割情報を付与した。

a)　複合動詞

「動詞＋動詞」で構成される複合動詞については，語彙的複合動詞，統語的複合動詞の別を問わず，基本的に分割情報を付与し単独動詞を「A 単位」，複合動詞を「B 単位」とした（表 3）。

検索システムで利用する場合，たとえば「錆び付く」から「錆びる」が検索できない（あるいはその逆）のは，重大な不具合だからである。ただし，「見積もる」「仕舞う」のように，構成要素である単独動詞の意味が全く継承されていないものは，分割情報を付与せず，この長さを「A 単位」とした。

表 3　複合動詞の分割情報の例

錆び／付く	反り／返る	取り／出す
塗り／分ける	売り／渡す	譲り／受ける

b)　複合名詞

最小単位を用いた別の表現に容易に言い換えられるものや，最小単位に接辞が付いたものは，分割情報を付与し，各構成語を「A 単位」とした（表 4）。

表 4　複合名詞の分割情報の例 1

猫／探し	ゴミ／拾い	湯／洗い
紙／おむつ	仮／住まい	右／ふくらはぎ

14　http://unidic.ninjal.ac.jp/glossary#suw
15　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf , pp. 1-27.

また，表5に挙げるような見出しは，UniDicの短単位規定に基づいて斉一にアノテーションされた結果であるが，コーパス色が強いため，登録されていない他の類型表現と「A単位」での形態素解析結果を合わせるため，分割情報を付与し，各構成語を「A単位」とした。

表5　複合名詞の分割情報の例２

家鴨／柄	二人／組	迷子／牛
ユリ／科	母／山羊	海老／問屋

UniDic由来の見出しの再分割（分割情報付与）はごく一部しか行っていないが，和語見出しについては，まだ再分割する余地があると考えている。
たしかに語構成としては複数の最小単位から成り立っていても，「気持ち」や「靴下」の類まで再分割する必要がないのは明らかであり，UniDicの短単位は，「基準の分かりやすさ，ゆれの少なさという条件を満たしつつ，用例を収集して分析を行うという利用目的にもかなう単位」[16]であるといえる。これに手を入れるということは，基準の不明瞭さやゆれを生み，外部から見て，辞書の開発方針がわかりにくくなる可能性はある。
しかし，我々が分割情報を付与した複合動詞や複合名詞のような語群が，これまでUniDicが検索システム等で使いにくかった一因であることは確かである。UniDic由来の見出しを再分割するにあたって，我々は，できる限りわかりやすい基準を追究している。

２．２．２　NEologd由来の語句の短単位化

NEologd由来の語句は長単位のものが多いため，基本的にすべて分割情報を付与することとしている。ただし，量が膨大なため，まず機械的に分割情報を付与し，それを人手でチェックするという手順をとった。すなわち，NEologd由来の語句を登録せずに作成したSudachi辞書を用いて，NEologd由来の語句を形態素解析し，その形態素解析結果を仮の分割情報とした。これを一つ一つ確認し，間違っていれば修正する，とした。

また，形態素解析結果としては間違っていなくても，我々が付与したい段階的な分割情報でない場合がある。たとえば，「医療費控除」の形態素解析結果が「医療｜費｜控除」の場合，これは正しい形態素解析結果であるが，段階的な形態素単位を提供するためには，「医療費／控除」という分割情報を付与しておく必要がある。さらに「医療／費」の見出し登録も必要である。
こうした確認作業をしながら，分割情報の精度を高めていっている（付録A参照）。

16　http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-01.pdf , p.10.

２．３　正規化表記

日本語には，ひらがな，カタカナ，漢字，ローマ字の文字種があり，さらに漢字については，送り仮名，異体字，代用漢字の選択があるなど，非常に複雑な表記体系となっている。たとえば，「アキカン」には，「空き缶，空缶，空き罐，空罐，空きカン，空きかん」等の表記が可能である。

また，外来語の表記については，カタカナで表記する場合，語末の長音のあるなしに加え，原音や原綴りになるべく近く書き表そうとする場合に生じるゆれや，「アフィリエイト（正）ーアフェリエイト（誤）」等の誤用によるゆれもある。また，カタカナで表記せず原綴りをそのまま使用する場合もあり，表記法に多くのゆれが存在する。

テキスト処理において，これら表記のゆれを同一視できるかどうかは，タスクの精度に影響する重要な要素である。我々は，全ての登録見出しに正規化表記の情報を付与し，表記ゆれを同一視できるようにしている。

表６　正規化表記の例

登録見出し	正規化表記
空き缶	空き缶
空缶	空き缶
空き罐	空き缶
空罐	空き缶
空きカン	空き缶
空きかん	空き缶
美術館	美術館
団体戦	団体戦

表6に示すように，それぞれの登録見出しについて，１つの正規化表記の情報を付与している。正規化表記が見出し表記と同じものは，それ自身が正規化表記である。

正規化表記の情報付与に際しては，次の２点が問題となる。一つは，何々を表記ゆれとみなすか，そしてもう一つは，どの表記を正規化表記とするかである。

a）表記ゆれの範囲

我々は，次のようなもの，およびその組み合わせパターンを表記ゆれとしている。

表 7　表記ゆれの範囲

パターン	例
文字種の違い	向日葵-ひまわり-ヒマワリ，燐酸-りん酸-リン酸
漢字の違い （異体字,代用表記,慣用表記）	芸術-藝術，驚歎-驚嘆，徳用-得用
送り仮名の違い	受け付け-受付け-受付
外来語の表記違い	コミュニティー-コミュニティ-コミニティー-community
誤用	シミュレーション-シュミレーション
くだけた言い方	～ちゃあ-～ては

UniDic 由来の見出しについては，同じ語彙素を持つ見出しグループを表記ゆれと見なした。ただし，UniDic では，やや広めに表記ゆれを吸収している部分があったため，表 8 に示すように，新聞等でも書き分けが見られるものは，同じ正規化表記を付与せず，別々の語句とした。

表 8　UniDic の語彙素と異なる正規化表記の採用例

登録見出し	UniDic の語彙素	Sudachi 正規化表記
炊く	焚く	炊く
焚く	焚く	焚く
卸す	下ろす	卸す
下ろす	下ろす	下ろす

また，ひらがな表記については，複数の語句の可能性があるものは，強引にどれかに正規化することはしていない（表 9）。

表 9　多義性のあるひらがな表記の例

登録見出し	UniDic の語彙素	Sudachi 正規化表記
そば	側	そば
そば	岨	そば
そば	蕎麦	そば

b) 正規化表記の選定

正規化表記は表記ゆれを同一視するための情報であり，いわゆる正書法的な観点から表記の選定は行っていない。そのため以下のケースがある。

① よく使われる表記が必ずしも正規化表記でない場合がある。
例）うどん → 饂飩（正規化表記）

② 同じ構成語を含む複合語について，統一的な正規化表記が付与されていない場合がある。
例）イヤフォン → イヤホン（正規化表記）
スマートホン → スマートフォン（正規化表記）

③ 分割情報によって短い単位に分割された場合，各構成語は，元の複合語と異なる正規化表記となる場合がある（表 10）。

表 10 形態素単位により異なる正規化表記の例

登録見出し	「C 単位」の正規化表記	「A 単位」or「B 単位」の正規化表記
取扱説明書	取扱説明書	取り扱い｜説明書
取り扱い説明書	取扱説明書	取り扱い｜説明書
ごまドレッシング	胡麻ドレッシング	ごま｜ドレッシング
胡麻ドレッシング	胡麻ドレッシング	胡麻｜ドレッシング

①，②については，表記ゆれの同一視という観点からは整合しているため，問題はないと認識している。また，③については，検索システムで使う場合，「C 単位」と「A 単位」を併用してもらうことにより精度は確保できると考えている。

３．辞書の精度

３．１ 弊害語の抑制

UniDic 由来の見出しは基本的にすべて採用しているが，形態素解析に弊害となる可能性のあるものは登録を保留とした（付録 A 参照）。たとえば，２文字のカタカナ・ひらがな表記で品詞が記号のものや，１文字の漢字表記で品詞が記号のもの，複合型の数詞等である。これらは，辞書に未登録の語句がブツ切れに形態素解析される弊害や，数詞処理の弊害となるおそれがあるため，登録を見合わせた。

例） アー,記号,一般,*,*,*,*,アー　　うら,記号,一般,*,*,*,*,ウラ
垈,記号,一般,*,*,*,*,ヤ　　六十,名詞,数詞,*,*,*,*,ロクジッ

また，間違いについては適宜修正を行った[17]。

例） 油染みる,動詞,一般,*,*,下一段-マ行,終止形-一般,アブラジミル

→ 「動詞,一般,*,*,上一段-マ行,終止形-一般」（品詞修正）

圧し遺る,動詞,一般,*,*,五段-ラ行,終止形-一般,オシヤル

→ 「圧し遣る」（表記修正）

３．２ 単語コスト

Sudachi が形態素解析をするためのパラメータは，UniDic の単語コストや連接コストを利用している。UniDic 由来の見出しはそのまま値を継承すればよいが，NEologd 由来の語句や内省で追加した語句については，なんらかの方法で単語コストを与える必要がある。そこで，新規追加見出しについては，次のような推定コスト値を付与するとしている。

- 分割情報により内部構造が記述されている見出しについては，それぞれの構成語の単語コスト（から一定の値を引いたもの[18]）および構成語間の連接コストの和を全体の複合語の単語コストとする。
- 分割情報を付与されていない見出しについては，字種や文字数により，特定の値を付与する[19]。

図 1 に例を示す。

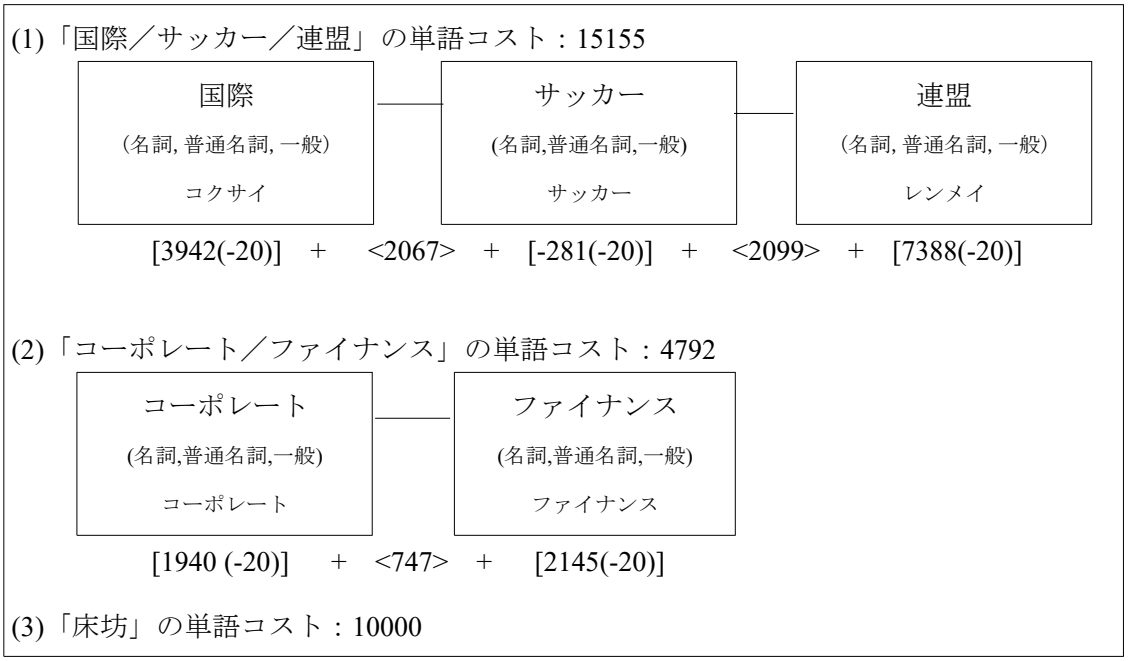

図 1 単語コストの推定 1 [ ] … 単語コスト値 < > … 連接コスト値

17 我々が採用したバージョンは，unidic-mecab-2.1.2_src.zip（http://unidic.ninjal.ac.jp/back_number#unidic_cwj）である。

18 現在は，-20 を適用している。

19 現在は，見出し表記が，アルファベットとスペースのみで構成される場合は"5000"，顔文字は"5000"，記号は"22000"，記号以外で３文字以下の表記は"10000"，記号以外で４文字以上の表記は"15000"，としている。

概ね，この推定コスト値で正しい形態素解析結果を得られているが，新規追加見出しに付与した推定コスト値より，他の登録語の方がコストが低いために，新規追加見出しが誤解析となる場合がある（図2）。

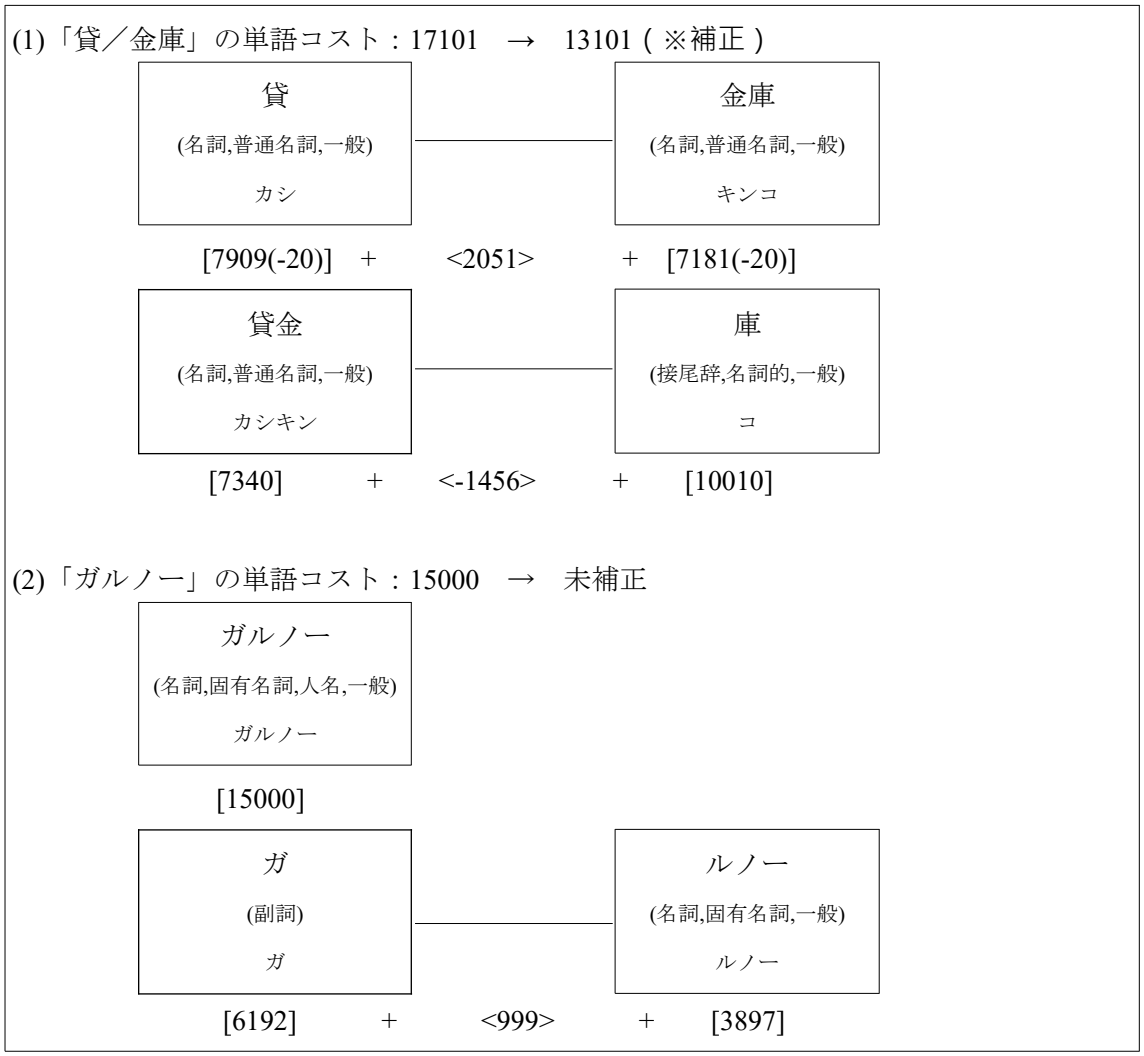

図2 単語コストの推定２ [ ] … 単語コスト値 <> … 連接コスト値

「貸金庫」の例では，「貸｜金庫」の合計コストより「貸金｜庫」の合計コストの方が低いため，見出しの内部構造に基づく推定コスト値 (17101) では，「貸金庫」は形態素解析の第一解にならず，「貸金（カシキン）｜庫（コ）」が第一解となる。これを改修するため，現在は，手動で「貸金庫」の単語コストを適当な値に補正している。
また，「ガルノー」の例では，文字数と文字種により特定の単語コスト値 (15000) を付与しているが，「ガ（副詞）｜ルノー（名詞,固有名詞,一般）」の合計コストの方が低いため，「ガルノー」は形態素解析第一解にならない。

このように，新規追加見出しに付与する単語コストの推定方法には，現状の手法では，根本的な問題があることを認識している。すなわち，追加しようとしている見出しが未登録の状態で，その見出しの内部構造として記述した短単位で正しい形態素解析結果が得られる，あるいは，内部構造によらず特定の値を付与した見出しについては，当該文字列について，他の登録語による形態素解析結果の方がコストが高い，という前提に立っているからである。
新規追加見出しに付与する単語コストの推定方法については，現在，抜本的な見直しを検討中である。

４．メンテナンスの継続

今後も，NEologd から定期的に新語を取り込み辞書の最新性を確保するとともに，機械的なチェック，人手によるチェックを併用しながら辞書内容を拡充，洗練していく。また，既登録語についても，正規化表記や分割情報の付与漏れ，品詞や読みの間違い等，点検が必要な見出しが半数近く存在する。これらの修正も継続して行う。その成果は，随時，OSS として公開していく[20]。”長期にわたる継続的なメンテナンス”[21]は，我々の開発方針の一つである。

５．おわりに

我々は，UniDic をベースに，NEologd から大量の固有名称を登録し，大規模な辞書データを構築した。280 万語を超える登録規模となり，付加情報の整備も着実に進んでいる。

「C 単位」を使えば，「調布の味の素スタジアム」という文字列は，「調布(名詞,固有名詞,地名,一般)｜の(格助詞)｜味の素スタジアム(名詞,固有名詞,一般)」と解析できるし，「ロミオとジュリエットを上演」は，「ロミオとジュリエット(名詞,固有名詞,一般)｜を(格助詞)｜上演(名詞,普通名詞,サ変可能)」と解析できる。形態素解析レベルで，「調布の → 味」や「ロミオと → 上演」のような間違った係り受けの可能性を排除できることが期待できる。しかし，文の構造を解析する上で重要な要素である「として」「にもかかわらず」のような複合辞については，未登録である。また，「役に立つ」「年をとる」のような成句についても，これらを一塊で認識できるデータはない。今後は，こうした連語のデータ構築も進めていきたい。

また，ベースとする UniDic を新しいバージョンに差し替えることも検討中である。さらに，同表記で読みが異なる語句の曖昧性解消や，新規に見出しを追加する際の単語コストの最適化についても課題である。

20 https://github.com/WorksApplications/Sudachi
21 http://www.lrec-conf.org/proceedings/lrec2018/pdf/8884.pdf , p.2.

謝　辞

Sudachi の辞書開発にあたっては，UniDic，NEologd から多くの研究成果を継承している。UniDic の開発に尽力された方々，そして，LINE 株式会社 佐藤敏紀氏をはじめ，NEologd の開発関係者の皆様に感謝申し上げる。また，奈良先端科学技術大学院大学 情報科学研究科教授の松本裕治氏には，Sudachi の形態素単位や正規化表記について，数回にわたる広範な議論の中で適切な助言をいただいた。深くお礼申し上げる。

文　献

工藤拓．Mecab: Yet another part-of-speech and morphological analyzer.

浅原正幸，松本裕治 (2003)『ipadic version 2.7.0 ユーザーズマニュアル』
奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座

伝康晴，小木曽智信，小椋秀樹，山田篤，峯松信明，内元清貴，小磯花絵 (2007)
『コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用』
日本語科学 22, pp. 101-123.

伝康晴，山田篤，小椋秀樹，小磯花絵，小木曽智信 (2008)
『UniDic version 1.3.9 ユーザーズマニュアル』

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, Yasuharu Den (2014)
"Balanced corpus of contemporary written Japanese." Language Resources and Evaluation, 48:345-371

小椋秀樹，小磯花絵，冨士池優美，宮内佐夜香，小西光，原裕 (2011)『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下）』』文部科学省科学研究費特定領域研究「日本語コーパス」データ班

佐藤敏紀，橋本泰一，奥村学 (2016)『単語分かち書き用辞書生成システム NEologd の運用 ―文書分類を例にして ―』情報処理学会 研究報告自然言語処理 2016-NL-229-15

佐藤敏紀，橋本泰一，奥村学 (2017)『単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討』言語処理学会 第 23 回年次大会 発表論文集, pp.875-878

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, Yuji Matsumoto (2018) “Sudachi: a Japanese Tokenizer for Business”

付　録　A

最新版辞書(2018 年 7 月版)の登録見出し数（うち，精査済み数）	2,801,739　(1,339,630)
分割情報を付与した見出し数（うち，UniDic 由来の見出し数）	1,676,735　(　284,335)
UniDic 由来の語句で登録保留としたもの	1,648