コーパスコンコーダンサ『ChaKi.NET』の連続値データ型

浅原 正幸 (国立国語研究所コーパス開発センター) *
森田 敏生 (総和技研)

Double Type Data on 'ChaKi.NET'

Masayuki Asahara (Center for Corpus Development, NINJAL)
Toshio Morita (Sowa Research Co., Ltd.)

1. はじめに

ChaKi.NET (Matsumoto et al. (2005)) は、コーパスに付与されたメタデータ・形態論情報・係り受け情報を用いて文単位の検索を行ったり、形態論情報・係り受け情報に基づく頻度統計情報を取得したり、自動解析により付与された形態論情報・係り受け情報を修正したりすることができるコーパス管理システムである。書字テキストからなるコーパスを前提とし、内部におけるデータ型はテキストやアノテーションを格納する文字列型やアノテーションを抽象化した整数型が用いられてきた。このため、音声コーパスを格納するためには書き起こしテキストのみを格納せざるを得なかった。また、書字テキストであってもコーパスを読むときの読み時間など連続値型を格納することはできなかった。

本稿では、コーパス管理システム ChaKi.NET の新しいデータ型について紹介する。ChaKi.NET に新しいデータ型として形態素単位に時刻・時間情報を格納する連続値データ型を設けることで「日本語話し言葉コーパス」(CSJ) などの音声コーパスを時刻情報に対するスタンドオフ形式で形態論情報・係り受け情報を格納することができる。

具体的には既存のデータベースの形態素に対応する単位で、開始時刻 (`start_time`)・終了時刻 (`end_time`)・継続時間 (`duration = start_time - end_time`) の三カラムからなるテーブルを追加する。時間情報を含まない通常のコーパスは各カラムの値を `null` で初期化されている。時間情報が付与されているコーパスの場合、通常のデータベース作成手続きのあと、コマンドラインから実行する `timings.exe` を用いてデータベースに格納することができる。

以下では、発話時刻が収録されている「日本語話し言葉コーパス」(Corpus of Spontaneous Japanese; CSJ) を用いた連続値データ型の活用事例を紹介する。

2. 「日本語話し言葉コーパス」(CSJ) の活用事例

以下では「日本語話し言葉コーパス」(CSJ) を利用した活用事例について示す。利用するのは「日本語話し言葉コーパス」コア RDB 版 (version 1.0) の統語情報サブセットデータベース `csj_syn.db` である。この SQL データベース形式を ChaKi.NET インポート用拡張 CaboCha フォーマットや時刻情報用 TSV ファイルを変換するために必要なプログラム `csj2cab.rb` は

* masayu-a@ninjal.ac.jp

https://github.com/masayu-a/ChaKi-CSJDB2DB から入手することができる。

2.1 初期設定

まず、CSJ のメタデータ・形態論情報・係り受け情報を ChaKi.NET のデータベースとして格納するために拡張 CaboCha フォーマットに変換する。

ChaKi.NET インポート用拡張 CaboCha フォーマットファイル csj.cabocha の生成

```
> ls csj_syn
csj_syn.db
> ruby csj2cab.rb > csj.cabocha
```

以下に生成された拡張 CaboCha フォーマットの例を示す。

ChaKi.NET インポート用拡張 CaboCha フォーマットの例

```
#! DOCID 1 <TalkID>A01F0055</TalkID><Channel>L</Channel>...(省略)
#! DOC 1
...
...(省略)
...
と 助詞,格助詞,,,,,ト,と,ト
いう 動詞,,,,ワア行五段,連体形,イウ,言う,ユー
* 7 10D 0/0 0
こと 名詞,,,,,,コト,事,コト
で 助詞,格助詞,,,,,デ,で,デ
* 8 -1ROOT 0/0 0
えー 感動詞,,,,,,エー,えー,(F エー)
* 9 -1ROOT 0/0 0
す 言いよどみ,,,,,,,,(D ス)
* 10 -1ROOT 0/0 0
発表 名詞,,,,,,ハッピョウ,発表,ハッピョー
し 動詞,,,,サ行変格,連用形,スル,為る,シ
ます 助動詞,,,,,終止形,マス,ます,マス
EOS
```

各行の形式について説明する。

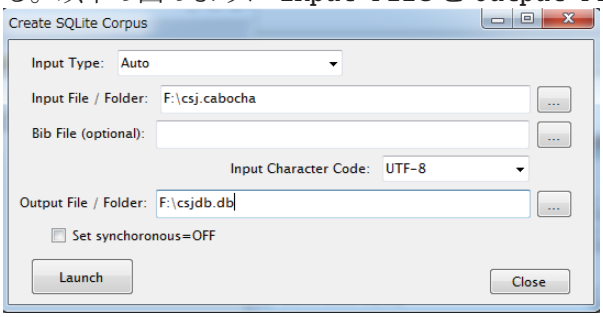
- #! DOCID で始まる行にはメタデータに相当する情報を格納する。CSJ のデータ格納においては、「談話基本情報」「話者基本情報」「対話情報」「再朗読情報」「単独印象評定情報」「集合印象評定情報」を格納する。「集合印象評定情報」については、一つのデータに対し複数の評定者による評定結果があるために、平均値を四捨五入した値を格納する。
- #! DOC で始まる行は、#! DOCID で定義したメタデータに対応するデータが以下の行から開始することを表す。
- * で始まる行は文節 ID と係り先の文節 ID、係り受けラベルが含まれている。それぞれ「文節係り受け」linkDepBunsetsu テーブル の「係り文節 ID」BunsetsuID・「受け文節 ID」ModifeeBunsetsuID・「係り受けラベル」Dep_Label を文ごとの 0-origin の値に変換して格納する。また、格納する制約上、係り受けラベルに対して表 1 のような変更を行った。

表 1　係り受けラベルの修正

元の係り受けラベル	変更したラベル
無印	‘‘D’’
‘‘A2’’	‘‘AA’’
‘‘D’’	‘‘DD’’
‘‘D_X’’	‘‘DX’’
‘‘R_P’’	‘‘RP’’
‘‘S:複数文節言い直し:S1’’	‘‘SS’’
‘‘S:複数文節言い直し:E1’’	‘‘SE’’

- EOS で始まる行は CSJ で規定されている節境界を表現する。
- 上記以外の形態素表層形で始まる行は形態素解析 MeCab の出力形式に変換したものに相当する。

生成された csj.cabocha を [Tools] → [Create SQLite Corpus] から SQLite 形式のデータベースに変換する。以下の図のように Input File と Output File を指定する。



この SQLite 形式のデータに対し、各形態素の開始時刻・終了時刻を格納するために、拡張 CaboCha フォーマットの形態論情報に対応する行（‘‘*’’, ‘‘#’’, ‘‘EOS’’ で始まらない行）に一対一対応する以下のような三列からなる TSV ファイルを作成する(1)。以下が時刻情報保存用の TSV ファイルの例である。

時刻情報保存用 TSV ファイルの例

```
...
...(省略)
...
と          4.00558         4.140573
いう        4.140573        4.196929
こと        4.196929        4.419734
で          4.419734        4.570856
えー        4.696272        4.805055
す          4.805055        4.883036
発表        5.551385        5.978241
し          5.978241        6.078064
ます        6.078064        6.532801
```

一列目は「短単位」subsegSUW の「タグ無し出現形」PlainOrthographicTranscription、二列目は「短単位」segSUW の「開始時間」StartTime、三列目は「短単位」segSUW の「終了

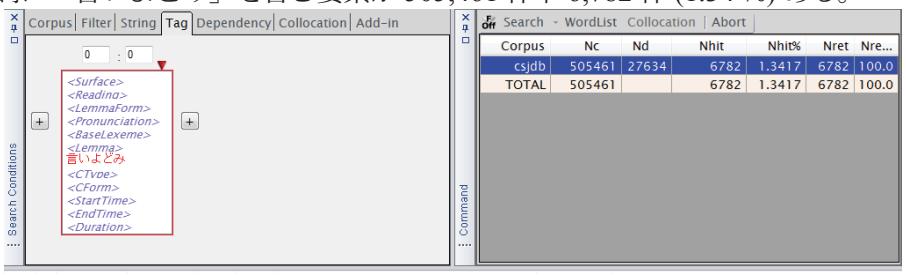
時間」EndTime に対応する。

2.2　検索事例

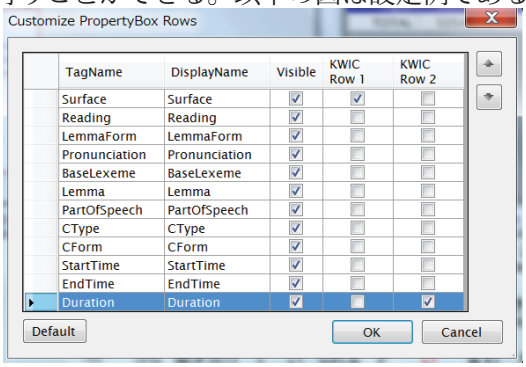
以下では、発話時間 (duration) と形態論・係り受け情報を組み合わせた検索事例について紹介する。

2.2.1　言いよどみと発話時間

以下の図は品詞に「言いよどみ」を含む要素数を [Tag Search] 機能を用いて検索した例である。品詞に「言いよどみ」を含む要素が 505,461 件中 6,782 件 (1.34 %) ある。



[Options] → [Property Box Settings] から KWIC 表示の 2 列目 (KWIC Row 2) に発話時間を表示する設定を行うことができる。以下の図は設定例である。

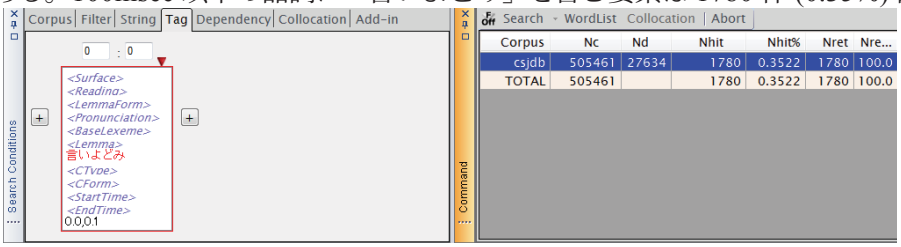


設定を行うと下図のように単語単位の KWIC 表示の下に、発話時間が表示される。発話時間上にある丸の大きさが発話時間の長さを表す。

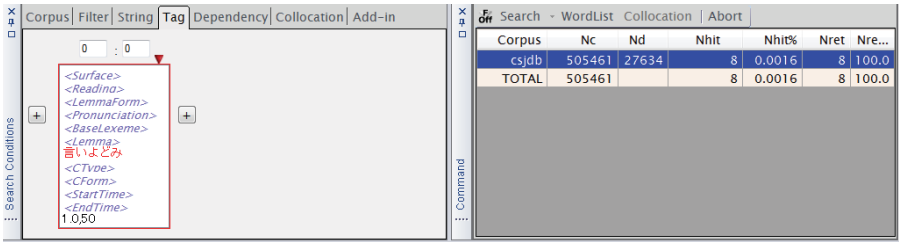
この品詞に「言いよどみ」を含む要素を発話時間を用いて絞込検索をすることができる。具

体的は [Tag Search] の Duration の列に最小値と最大値をコンマ区切りで表現することにより行う。

以下の例は 100msec 以下 (0.0 sec 以上 1.0sec 以下) の品詞に「言いよどみ」を含む要素の検索方法である。100msec 以下の品詞に「言いよどみ」を含む要素は 1780 件 (0.35%) 出現する。

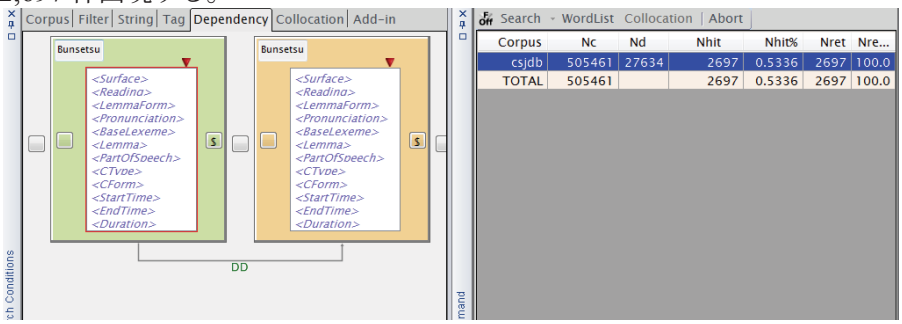


以下の例は 1000msec 以上 (1.0 sec 以上 50.0 sec 以下) の品詞に「言いよどみ」を含む要素の検索方法である。1000msec 以上の品詞に「言いよどみ」を含む要素は 8 件出現する。

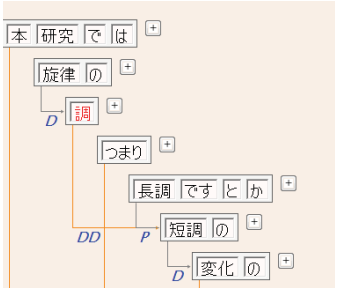


2.2.2　言い直しと発話時間

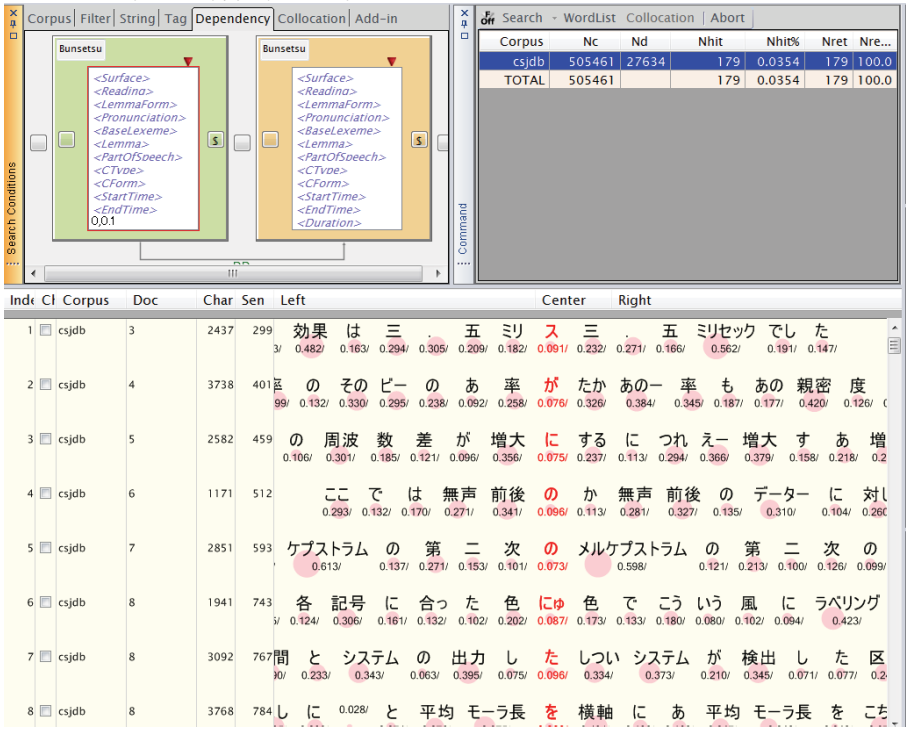
以下の図は「言い直し」を [Dependency Search] 機能を用いて検索した例である。「言い直し」は 2,697 件出現する。



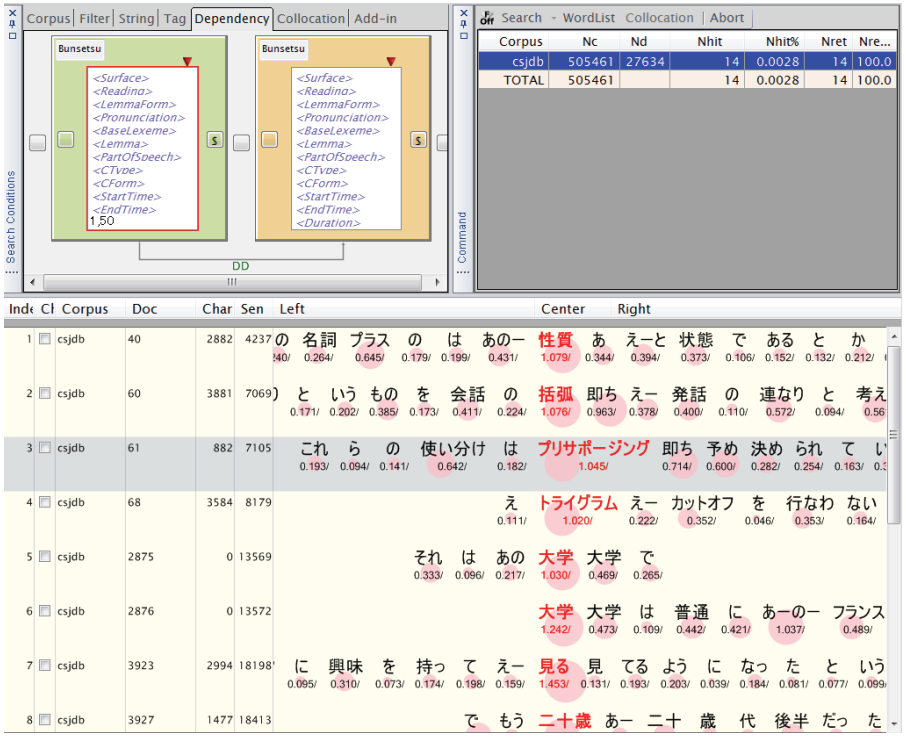
CSJ において「言い直し」は以下のような係り受け関係として表現されている。元のデータベースでは係り受けラベル ‘‘D’’ として表現されているが、他の係り受けアノテーションつきコーパスの多くが通常の係り受け関係をラベル ‘‘D’’ として表現するために、明確に区別するために ‘‘DD’’ として格納している。‘‘DD’’ の係り元が言い直す前の表現、‘‘DD’’ の係り先が言い直したあとの表現の主辞を含む文節を表す。



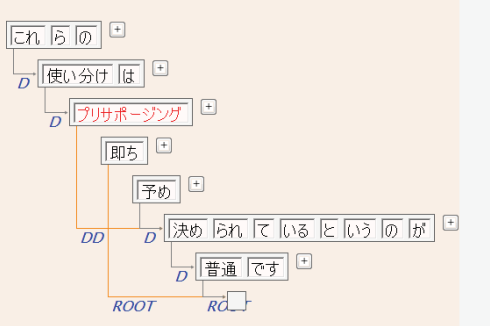
以下の検索事例では、言い直す前の表現が形態素単位で 500msec 以下である「言い直し」を検索したものである。179 件出現する。



以下の検索事例では、言い直す前の表現が形態素単位で 1000msec 以上である「言い直し」を検索したものである。14 件出現する。

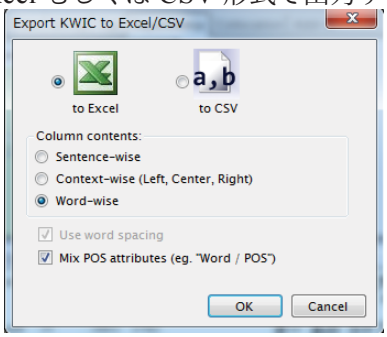


次の図は発話時間が長い形態素の言い直しの例である。長い表現の言い直しの場合、単純な単語の言い直しではなく、以下のような言い換えのようなものが含まれている。



2.3　検索結果の出力

格納されている時間情報は検索結果とともに出力することができる。[File] → [Send To Excel/CSV] より Microsoft Excel もしくは CSV 形式で出力することができる。



出力形式として、文単位 (Sentence-wise)・文脈単位 (Context-wise: 左文脈・KWIC 中央・右文脈)・単語単位 (Word-wise) の三種類を選択することができる。この際に Mix POS attributes を指定すると単語と品詞情報とともに、開始時刻・終了時刻・継続時間の三つの値を ‘‘/’’ 区切りで出力することができる。以下は [Tag Search] で「ちゃう」を検索した際に、単語単位に Excel 出力した例である。

O	P	Q
-1	0	1
崩れ/動詞/424.338043/424.642120/0.304062	ちゃう/助動詞/424.642120/424.845276/0.203146	場合/名詞/424.845276/425.126404/0.281127
違っ/動詞-促音便/446.386780/446.650543/0.263762	ちゃう/助動詞/446.650543/446.870422/0.219855	ん/助詞-準体助詞/446.870422/446.955505/0.085104
てっ/助動詞-促音便/814.408142/814.569031/0.160911	ちゃう/助動詞/814.569031/814.750854/0.181817	と/助詞-格助詞/814.750854/814.845886/0.095027
移っ/動詞-促音便/751.254944/751.534119/0.279165	ちゃう/助動詞/751.534119/751.887451/0.353369	ん/助詞-準体助詞/751.887451/751.987610/0.100138
似/動詞/105.125450/105.345215/0.219768	ちゃう/助動詞/105.345215/105.555038/0.209820	ん/助詞-準体助詞/105.555038/105.606956/0.051925
割っ/動詞-促音便/989.677856/989.826477/0.148596	ちゃう/助動詞/989.826477/990.036377/0.209935	と/助詞-接続助詞/990.036377/990.409424/0.373029
れ/助動詞/429.687408/429.772552/0.085138	ちゃう/助動詞/429.772552/429.964447/0.191878	と/助詞-格助詞/429.964447/430.040039/0.075589
なくなっ/動詞-促音便/430.345825/430.763397/0.417571	ちゃう/助動詞/430.763397/431.096466/0.333084	
れ/助動詞/466.494843/466.601105/0.106247	ちゃう/助動詞/466.601105/466.830414/0.229332	と/助詞-接続助詞/466.830414/467.372040/0.541617
狭まっ/動詞-促音便/476.929352/477.441620/0.512289	ちゃう/助動詞/477.441620/477.666962/0.225334	と/助詞-格助詞/477.666962/477.787842/0.120880
し/動詞/646.292908/646.367004/0.074088	ちゃう/助動詞/646.367004/646.499756/0.132802	言う/動詞/646.499756/646.609375/0.109579
出/動詞/647.586609/647.724670/0.138020	ちゃう/助動詞/647.724670/647.901306/0.176654	ん/助詞-準体助詞/647.901306/647.965637/0.064349
やっ/動詞-促音便/460.530823/460.676758/0.145939	ちゃう/助動詞/460.676758/460.838287/0.161542	って/助詞-副助詞/460.838287/460.984924/0.146635
焦げ/動詞/624.626648/624.800354/0.173747	ちゃう/助動詞/624.800354/625.041748/0.241395	ん/助詞-準体助詞/625.041748/625.158508/0.116716
し/動詞/663.762878/663.828552/0.065692	ちゃう/助動詞/663.828552/663.979797/0.151238	

3.　おわりに

本稿では、コーパス管理システム ChaKi.NET の新しいデータ型について紹介した。形態素単位に連続値型を三つ設けることで、開始時刻・終了時刻・経過時間の値を格納することができる。今後「日本語話し言葉コーパス」でよく利用される他の連続量について格納する方法について検討していきたい。

今回は「日本語話し言葉コーパス」を事例として活用方法について紹介した。他の利用方法

として、書字テキストの読み時間を格納することが考えられる。今後、書字テキストの読み時間を格納した場合の活用事例を紹介したい。

謝辞

本研究の一部は科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

Matsumoto, Yuji, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Ohtani, and Toshio Morita (2005). “Chaki: An annotated corpora management and search system.” *Proc. of the Corpus Linguistics Conference Series (Corpus Linguistics 2005)*.

関連 URL

- 「ChaKi.NET」 Web ページ：http://sourceforge.jp/projects/chaki/releases/
- 「日本語話し言葉コーパス」Web ページ:http://www.ninjal.ac.jp/corpus_center/csj/