ChaKi.NET lite の開発: Universal Dependencies コーパスの利用を見据えた ChaKi.NET ユーザインターフェイスの改良

伊藤 薫（九州大学）†
森田 敏生（総和技研）

Development of ChaKi.NET lite: Improvement of the User Interface of ChaKi.NET for the Universal Dependencies Treebank

Kaoru ITO (Kyushu University)
Toshio MORITA (Sowa Research Co., Ltd.)

要旨・既発表の有無

ChaKi.NET はコンコーダンサやアノテーションツールを含む多機能なコーパス管理システムである。現行の (ChaKi Legacy でない) ChaKi.NET は 2009 年 (Ver. 1.1, 現在公式サイトで確認できる最古のバージョン) に公開されたものであり、10 年以上に渡って機能追加や品質の向上が行われてきた。しかし、可能な処理が増え多機能化が進む一方で、インターフェイスが複雑化しコンコーダンサとしての利用等、簡易な作業にも学習コストが嵩む状態になっている。本研究では、ChaKi.NET の機能のうちコンコーダンサ機能に焦点を絞り、複雑なファイル操作に馴染みのないユーザにも利用しやすいインターフェイスを備えた ChaKi.NET lite の開発について述べる。想定するユーザ層としては Corpus of Contemporary American English (COCA) 等のコーパスを使用したことのある言語学者であり、近年自然言語処理分野で開発が盛んである通言語コーパス群 Universal Dependencies (UD) コーパスを容易に使用可能にすることを目指している。今回追加した主な機能は、UD コーパス群読み込み操作の簡略化、複数 UD コーパスの一括検索機能追加、検索実行および結果表示インターフェイスの改良である。

既発表無。

1. はじめに

ChaKi.NET (Matsumoto et al. 2006) は多機能なコーパス管理システムであり、コンコーダンサやアノテーションツールなど、自然言語処理向けのアノテーションやデータ閲覧にとどまらず、言語学の研究にも有用なツールである。ChaKi.NET の持つ機能のうち、言語学研究において今後有用だと思われるのが様々なフォーマットのコーパス読み込み機能や、タグや係り受けなど、様々な統語情報を指定可能な検索機能が挙げられる。現在主に用いられている ChaKi.NET では、2009 年に Ver. 1.1 として公開されたものに様々な機能が追加されており、様々な処理が可能であるが、機能の豊富さや歴史の長さ故に新規ユーザにとっては学習コスト

† ito@flc.kyushu-u.ac.jp

が高く利用しづらい面も抱えている。

そこで本研究では、ChaKi.NET の機能をコンコーダンサ関連に特化し学習コストを下げることにより、気軽に利用可能なツールである ChaKi.NET lite を開発する。ChaKi.NET lite デザインの方針は下記の通りである。このうち、最後の Universal Dependencies (UD) の重要性については次節で述べる。

- インターフェイスの改良により直感的に利用可能であること
- ChaKi.NET のデータベースと互換性があること
- 言語学研究に適した機能を持つこと
- コーパスを利用して言語学研究を行うユーザを主な対象とすること
- UD ツリーバンクの操作に適していること

なお、本論文のスクリーンショットは全て開発中のものであり、実際に公開するツールとは仕様が異なる場合がある。以下、本論文では UD を重視する理由について述べた後、主要な改良点を紹介する。

2. Universal Dependencies (UD)

UD は通言語的に統一された方法で依存構造や品詞などについてアノテーションが付与されたコーパス群を開発する国際プロジェクトである (Björkelund et al. 2017)。UD は依存構造が内容語主辞で付与されていること、Universal POS (UPOS) と呼ばれる品詞体系や依存関係タグが定義され、コーパスに付与されていることなどを特徴とする。また、人手アノテーションのしやすさやパージング精度の高さなど、言語学的妥当性以外も考慮して設計されている (浅原ほか 2019)。UD ツリーバンクの各コーパスは CoNLL-U というフォーマットで書かれているが、この形式は人間が直接依存構造木を読むことや、例文を検索することには適していない。本研究では、CoNLL-U の読み込みやタグ、係り受け情報の検索、依存構造木表示など、UD コーパスの利用に適した ChaKi.NET を改良することで、言語学や関連分野の研究を促進するツールの開発を目指す。

言語学研究における UD ツリーバンクの利点は、主にデータの多様さにある。最新バージョンの Ver. 2.10 では 130 言語、228 ツリーバンクを収録している。これにはチュクチ語やワルピリ語などの少数言語、アッカド語やコプト語、古代中国語（漢文）等の古代語、ヒンディー語—英語やトルコ語—ドイツ語などのコードスイッチング、学習者コーパス、手話（スウェーデン手話）などが含まれ、個々のデータ量は少ない場合もあるが時空間的、様態的に非常に幅広い言語が収録されている。また、先に述べたように UD は同じ枠組みでデータが記述されているため横断的な検索が容易である。つまり、品詞を例に挙げると UD では UPOS と呼ばれる統一された品詞体系でアノテーションされているため、一般名詞と接置詞（前置詞・後置詞）の係り受けなどを検索したい場合は全ての言語で `NOUN` タグと `ADP` タグの係り受けを検索すればよく、個別の言語に合わせてクエリを記述する煩雑さが軽減される。また、多くのコーパスは XPOS と呼ばれる個別言語に特化した（コンバージョン元コーパスの情報であることが多い）品詞情報を保持しており、個別言語に特化した検索も可能である。また、多くのデータがフリーで利用できることも、多言語データへの気軽なアクセスを可能にしている。このような

特徴から、UD ツリーバンク利用促進は類型論研究に役立つことが期待される。

既存ツールとしては UD 公式サイトで SETS treebank search, PML Tree Query, Kontext, Grew-match, INESS の 5 つが紹介されているが、これらはいずれもそれぞれ強みを持つものの、検索クエリ入力の記述方法が複雑であり、記法を習得するのに学習コストがかかる。本研究では、もともと GUI による検索クエリ記述が可能である ChaKi.NET を UD 向けに改良し、より簡便な検索を可能にする。

3. インターフェイスの改良

本節では、ChaKi.NET からの主な改良点について述べる。主な改良点は画面全体に表示される要素とその配置、ドラッグ&ドロップによるツリーバンク読み込み機能の追加、複数コーパスの表示改良および選択機能、操作フローの改善であり、これらを順に紹介する。

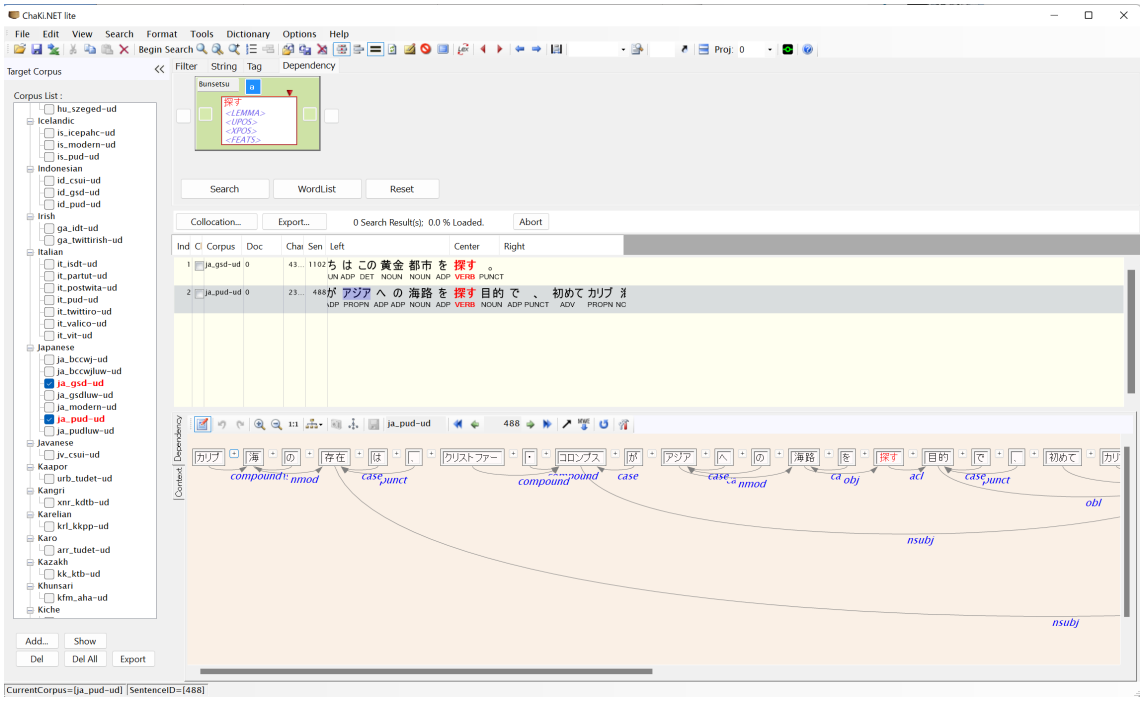

図 1　ChaKi.NET lite（開発中）の画面

3.1 基本設計

ChaKi.NET にはアノテーション機能が含まれるが、言語学の研究においてはアノテーション機能を用いる機会は少ない。実際、主要なコーパスである Corpus of Contemporary American English (COCA), British National Corpus (BNC), 『現代日本語書き言葉均衡コーパス』(BCCWJ) にも主要なインターフェイスにアノテーション機能はなく、需要も少ないと思われる。そこで、ChaKi.NET lite では需要の高いコンコーダンサ機能に特化し、1 画面で操作が完結することを目標に設計した。具体的にはコーパス読み込み、係り受け検索、結果表示、係り受けツリー可視化という一連の操作を直感的に行えることを目指し、ChaKi.NET

から表示する機能やペインの絞り込みを行った。これを図 1 に示す。図 1 の左列でコーパスリストの読み込みやコーパス選択、右列上段で係り受け検索、右列中段で検索結果表示とツリー表示対象文の選択、右列下段で選択した文のツリー表示を行うことができる。

3.2 ドラッグ&ドロップによる UD ツリーバンク一括読み込み

ChaKi.NET では.conllu ファイルを読み込む際、アプリケーション上の機能を用いて.db ファイルに変換する操作が必要であった。この操作は『中納言』を介した BCCWJ の利用、あるいは、ウェブ上のインターフェイスが提供されている BNC, COCA を始めとしたコーパスを中心に利用している研究者にとっては使用上の障壁になる。そこで、UD ツリーバンクデータから事前に変換・提供される.db ファイルの格納されたフォルダを画面上にドラッグ&ドロップすることで一括読み込みする機能を追加した。これにより、複雑な前処理なしで大規模な通言語コーパスの利用を可能にした。

3.3 複数コーパス選択機能

ChaKi.NET lite では 3.2 で述べたツリーバンク一括読み込みの後、収録コーパスを複数選択し、横断的な検索が可能である。コーパスは UD ツリーバンク内のフォルダ構造を反映し、日本語、英語など言語ごとにまとめて表示され、クリックにより折りたたみや展開が可能である。また、複数選択に関しては各コーパス横のチェックボックスをクリックすることで検索対象とするか否かを選択でき、選択されたコーパスは赤字で強調表示される。

3.4 操作フローおよび結果表示の改良

既存の ChaKi.NET では、検索クエリ入力画面と検索実行ボタンが離れた場所に配置されており、検索クエリの入力後に検索を実行する手順が直感的にわかりにくい状態であった。ChaKi.NET lite では Google 検索や COCA など、多くの想定ユーザが慣れ親しんでいると思われるインターフェイスに合わせ、検索クエリ入力場所の直下に検索実行ボタンや語彙リスト表示ボタン、検索条件リセットボタンを配置することで、直感的な操作を可能にした。他にも検索ヒット数の表示箇所をコンパクトにするなど、検索クエリ入力から例文表示、ツリー表示という一連の操作に集中できるようなデザインにした。

4. おわりに

本論文では、現在開発中の ChaKi.NET lite について、開発目的、主な改良点と想定ユーザを中心に述べた。派生元の ChaKi.NET が総合的なコーパス管理・編集・検索システムであるのに対し、本研究で開発した ChaKi.NET lite は、インターフェイスの改良とコンコーダンサとして機能を絞り込むことで、学習コストが低く容易に利用できるようにした軽量版と位置づけられる。ChaKi.NET の開発により、言語データの利用が促進されることを期待したい。なお、ChaKi.NET は開発途上であり、本論文で紹介した点以外にも今後様々な改良を施す予定である。一定の機能を追加後、ChaKi.NET lite は Github 上にて公開予定である。

謝　辞

本研究は国立国語研究所基幹型プロジェクト「実証的な理論・対照言語学の推進」サブプ

ロジェクト「アノテーションデータを用いた実証的計算心理言語学」および、JSPS 科研費 19K13180 の助成を受けたものです。

文　献

Yuji Matsumoto, Masayuki Asahara, Kiyota Hashimoto, Yukio Tono, Akira Ohtani, and Toshio Morita (2006). “An Annotated Corpus Management Tool: ChaKi.” *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

Anders Björkelund, Agnieszka Falenska, Xiang Yu, and Jonas Kuhn (2017). “IMS at the CoNLL 2017 UD Shared Task: CRFs and Perceptrons Meet Neural Networks.” *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 40–51. Vancouver, Canada: Association for Computational Linguistics.

浅原正幸・金山博・宮尾祐介・田中貴秋・大村舞・村脇有吾・松本裕治 (2019).「Universal Dependencies 日本語コーパス」　自然言語処理, 26:1, pp. 3–36.

関連 URL

コーパス検索アプリケーション『中納言』	https://chunagon.ninjal.ac.jp/
British National Corpus (BNC)	https://www.english-corpora.org/bnc/
Corpus of Contemporary American English (COCA)	https://www.english-corpora.org/coca/
ChaKi.NET	https://ja.osdn.net/projects/chaki/
ChaKi.NET lite（後日公開予定）	https://github.com/chakidev/chakinet-lite
Grew-match	http://match.grew.fr/
INESS	http://clarino.uib.no/iness
Kontext	http://lindat.mff.cuni.cz/services/kontext/corpora/corplist
PML Tree Query	http://lindat.mff.cuni.cz/services/pmltq/
SETS treebank search	http://depsearch-depsearch.rahtiapp.fi/ds_demo/
Universal Dependencies	https://universaldependencies.org/