読解支援システムのための言語非依存フレームワーク構築

阿辺川 武 † 八木 豊 † 戸次 徳久 † 傅 亮 ‡
Slaven Bilac† 奥村 学 † 仁科 喜久子 †

† 東京工業大学 ‡ フウズラボ

1 はじめに

著者らは，日本語読解学習支援システム “あすなろ”[1] の開発を続け，インターネット上で公開してきた[1]．本研究の最終目標は，以下に挙げる 3 項目である．(1)Web 上で学習可能な第二言語読解学習支援を多言語対応システムとして開発する．学習者の母語による支援により文章理解できることを目指す．(2) 学習者の能力別の学習を可能にする．一斉授業で個々の学習者が満足できる専門読解を目指すことはむずかしいが，Web 上では，個別に学習者に最適な内容を選択でき，学習レベルに合わせた時間配分も可能となる．(3) 自然言語処理，日本語学，第二言語習得理論 (外国語学習理論)，教育工学などの学際的視点から新たに各分野に新しい知見を加えることである．

現在のシステムの主な機能は，学習者が入力した日本語の文章に対し，文章中の単語の意味や対訳，そして文法項目の意味を表示することである．その際，Web 画面表示や辞書データベースを Unicode で構成することにより，日本語，英語，マレー語，インドネシア語の他，中国語，タイ語等の特殊な文字を含めた多言語表示ができる．

Web で利用できる同様な読解支援システムとして “リーディング チュウ太[2]” や “rikai.com[3]” などが存在する．本システムの特徴は，入力された日本語文に対し，文法項目や慣用句などの複合要素を提示できること，そして文節ごとの係り受け関係を表示できるといったことが挙げられる．

“リーディング チュウ太” では日本語から英語だけでなく日独への対応が現在行われており，また “rikai.com” では，英日，英西，中英といった対訳にも対応している．このように読解支援システムは基本的な枠組みがあれば，どの言語でも同様の対応が可能であると思われる．したがって辞書および形態素解析ツールを追加することで容易に多言語へと拡張できる．ただし，本システムのように，構文構造や複合要素を扱う場合，単純に辞書を用意するだけでは多言語に対応できない．構文構造は節構造や句構造などのように言語によって異なり，複合要素は構文構造に依存した形で記述されるからである．したがってこれらの要素を言語に依存しない形で扱える仕組みが必要となる．本稿では，読解対象言語を日本語だけでなく任意の言語へと対応させる際に構築するフレームワークについて紹介する．

[1] http://hinoki.ryu.titech.ac.jp/
[2] http://language.tiu.ac.jp/
[3] http://rikai.com/

2 言語非依存フレームワーク

現状のシステムは，日本語学習者を想定し，日本語の文章の読解を支援する目的で開発されている．入力された日本語文を解析し，分かち書きされた単語や抽出された文法項目に対して日本語の意味を表示している．また学習者の母語に応じた単語の対訳を表示できる．現在，対訳が表示可能な言語は，英語，中国語，タイ語，インドネシア語，マレー語の 5 言語であるが，新たに言語を増やす際には，日本語との対訳辞書さえ用意すれば対応できる．

本システムの拡張を考える時，日本語から多言語への読解支援が可能なら，逆に多言語から日本語への読解支援システムも可能な枠組みであることが望ましい．さらに展開すると多言語から多言語の読解支援システムへの拡張が可能になると思われる．ただし，構文解析などの言語処理ツールを言語に応じて用意する必要がある．一方で，データ保存形式や文法項目の検索手法などが言語依存の構造であると，容易に言語の追加ができなくなる．そこで，本稿では言語に依存しない部分を統一した形式で扱えるフレームワークを提案する．

2.1 構成

言語非依存のフレームワークを構成していく上で，言語解析ツールや辞書などの言語依存部と，それ以外の非依存部とに明確に分離する必要がある．図 1 は，本システムの構成図である．以下，入力部，言語処理部，出力部の順にそれぞれ説明する．

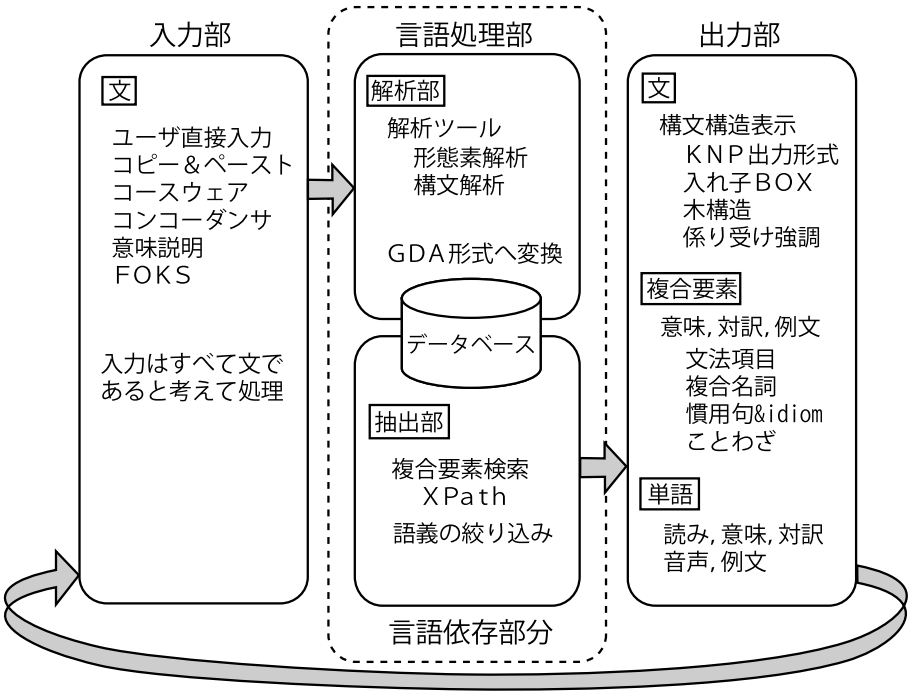

図 1: 構成図

2.1.1 入力部

入力部はユーザインタフェースの一部であり，説明文が学習者の母語で書かれる他は，言語非依存である．本システムで想定している入力は，ユーザが読解の際に入力する文やコースウェアだけでなく，単語そのものであったり，文法項目や慣用句などの文の一部も入力として受けつける．また，直接入力だけでなく FOKS[2] のような入力支援ツールから入力することも可能である．さらに，システム側が出力した語句の意味を説明した文や例文などを再びシステムの入力として扱うことができる．入力された文もしくは語句は，内部的にはすべて文として扱い，処理を簡単化している．

2.1.2 言語処理部

言語処理部はシステム内部で言語が処理されることから，言語に依存したリソースが必要となる．最初に文から辞書に登録された形で単語を切り出すために形態素解析ツールが必要となる．次に本システムの特徴の 1 つである構文表示 [1] と，文法項目や慣用句などの複合要素を抽出するためには，構文解析ツールが必要である．現在，日本語の解析には JUMAN,KNP[3] を使用している．そして今後拡張を考えている英語と中国語では Stanford Lexicalized Parser[4][5] を使用する予定である．

構文解析後，入力された文をデータベースに格納し，ユーザの入力履歴をとる．入力言語として日本語のみを対象としていた以前のシステムでは，構文解析ツールが出力した文節の係り受け構造をそのまま格納していた．しかし，この方式では日本語のような文節を中心として扱う言語しか扱えない．そこで英語のような句構造を持った言語にも対応したデータの保存形式が必要となる．ここでは橋田らが提唱する GDA(Global Document Annotation: 大域文書修飾) [6] を利用する．GDA については次の 2.2 節で詳しく紹介する．

単語の分かち書きは形態素解析ツールにより行われているが，文法項目や慣用句などの抽出は既存のツールではできない．そこで，データベースに保存された文からそれらを検出する機能が必要となる．これは 3 節で詳しく説明する．

2.1.3 出力部

出力部は，言語処理部の出力をユーザに提示する部分である．表 1 に本システムでユーザに提示できる要素を挙げる．

表 1: 提示できる要素

文	構文構造，音声 (コースウェアのみ)
単語	読み，意味，発音，例文
複合要素	意味，例文

単語や複合要素の意味は，EDR 電子化辞書 [7] を使用し，日本語と英語で表示している．EDR に含まれていない中国語，タイ語，インドネシア語，マレー語に

ついては，その一部の単語に対して，独自に EDR 概念識別子との対応付けをおこなっている．その結果として，概念識別子を中間言語とすることで任意の言語間の対訳を得ることができる．

また，本システムでは留学生の日本語授業で実際に使用されているテキストをコースウェアとして掲載しており，コースウェア中の文と単語に対しては，日本語教師による朗読音声を聞くことができる．

単語や複合要素の一部では，実際の文中での使用例を見ることができる [8]．例文には中学生新聞などの比較的わかりやすい文章を採用している．

2.2 GDA について

最終的に必要となるデータの形式は，言語に依存せず，なおかつ構文構造をも保持した形式であることが望ましい．当初は独自にデータ構造を定義することを考えたが，本システムでは GDA(Global Document Annotation; 大域文書修飾)[6] を利用することにした．GDA は，文書の意味的，語用論的構造を計算機が自動的に認識することを可能にする XML のタグ集合を規定している．翻訳，照応解析，情報抽出，情報提示などさまざまな分野で，GDA を前提とした処理が行なわれており，今後これらの技術が実用化されたとき，容易に本システムと統合することができる．

GDA では意味的関係の記述に対して，多くのタグと属性が定義されているが，本システムで必要な形態素に関する属性は少ない．例えば，複合要素の検索で使用する単語の原形を記述する属性が存在しない．そこで本システムでは，タグ内の属性として原形を表わす “orig” を導入する．「昨日私は言語を絶する体験をした」を KNP で解析し，GDA の形式へと変換した例を図 2 に載せる．

```
<su>
  <np prn="きのう" orig="昨日">昨日</np>
  <adp>
    <n prn="わたくし" orig="私">私</n>
    <ad prn="は" orig="は">は</ad>
  </adp>
  <adp>
    <adp>
      <np prn="げんご" orig="言語">言語</np>
      <ad prn="を" orig="を">を</ad>
    </adp>
    <v prn="ぜっする" orig="絶する">絶する</v>
    <ad>
      <vp prn="たいけん" orig="体験">体験</vp>
      <ad prn="を" orig="を">を</ad>
    </ad>
  </adp>
  <v prn="した" orig="する">した</v>
</su>
```

図 2: GDA の例

3 複合要素の検索

3.1 XPath を用いた検索方式

複合要素とは複数の形態素から構成されている要素のことである．複合名詞のように隣接して形態素が並んでいるものもあれば,「決して〜ない」のように離れた位置で呼応しているものもある．表 2 に本システムで表示可能な複合要素の例を挙げる．

表 2: 複合要素の例

文法項目	はおろか，なければならない
慣用句，イディオム	顔が広い，point of view
ことわざ	猿も木から落ちる
複合名詞	民主主義，自然言語

慣用句の中には名詞が動詞に係っていれば，間に他の格要素や副詞が挿入されていてもよいものがあり，このような係り受け関係を考慮した検索は，grep などの正規表現を利用したマッチングでは実現が難しい．

GDA は XML 形式で構成されており，データの検索には既存の XML を扱う種々の手法が使用できる．そこで複合要素の検索には伊藤らの手法 [9] と同様に XML のデータ構造を検索する形式の 1 つである XPath[10] を用いる．XPath は，XML データを表す木構造をたどり，ある条件を満たす要素や属性を検索する記述方法で，W3C により規定されている仕様である．複合要素の検索を XPath の検索式に置き換えれば，実際の検索には XML データベースに実装されている検索エンジンが利用できる．これにより独自に検索部分を実装する必要はなくなるとともに，言語に依存しない検索が実現できる．

従来のシステムでは，用意している例文に対してどの複合要素を含んでいるかというインデックスを予め作成し，検索に利用していたのだが，これでは例文や複合要素を追加する時に，再びインデックスを作成し直さなければならない．しかし XPath を用いた検索方式を採用することにより，リアルタイムで複合要素の検索ができるようになり，管理のコストが軽減される．

3.2 実装例

言語に依存しないといっても，複合要素を検索する XPath 式は，その言語の形態素の単位や構文構造に応じて記述しなければならない．次頁表 3 に複合要素に対する XPath 式の例を示す．GDA のタグには，統語構造の種類を示す syn 属性や交差する依存関係を示す dep 属性などがあるが，本システムへの導入は現時点では考慮していない．

表 3: XPath 式の例

自然言語	//np[.=”自然”][following::*[1]=”言語”]
顔が広い	//np[.=”顔”][parent::*[following-sibling::*//@orig=”広い”]]
point of view	//n[.=”point”][parent::*[following-sibling::adp[ad=”of ”][np=”view”]]]

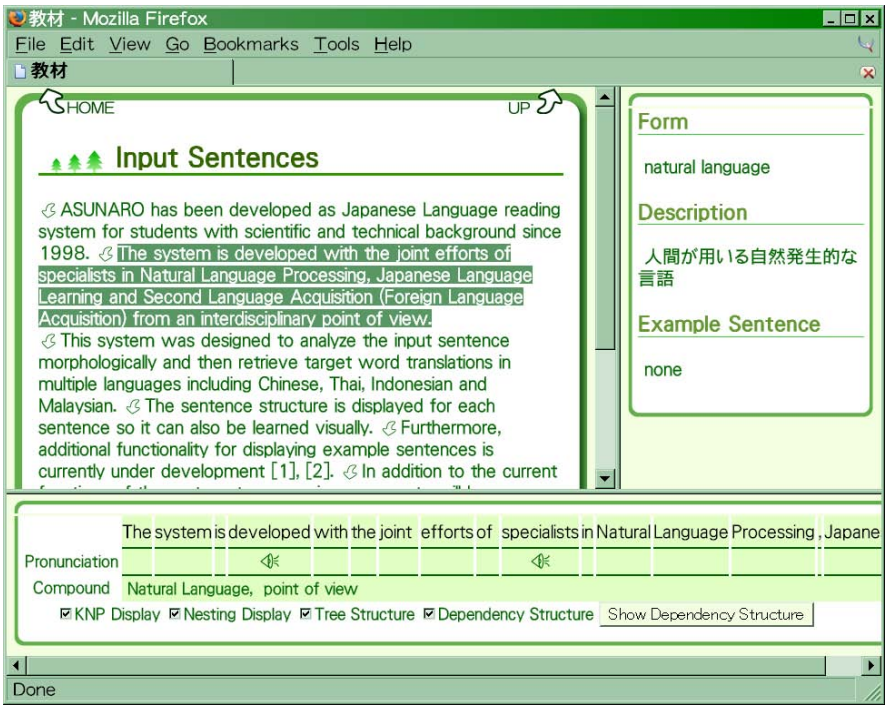

図 3: 画面の例

最後に実装例として英語の読解支援を想定したスクリーンショットを図 3 に載せる．複合要素である “natural language” を含んだ文が選択され，画面右でその意味を表示している．

4 おわりに

本稿では，日本語学習支援システム “あすなろ” について，学習対象言語を日本語だけでなく任意の言語へと対応可能なフレームワークを紹介した．システムを言語に依存する部分と依存しない部分に分割し，非依存部は共通の枠組みで扱えるような実装方法を説明した．今後は，扱える言語を増やし，システムを実際に使用してもらいながら発生する課題について検討する．そしてシステムのさらなる改良に取り組んでいきたい．

参考文献

[1] 阿辺川武, 八木豊, 戸次徳久, 澤谷孝志, 奥村学, 仁科喜久子, 杉本茂樹, 傅亮. 日本語学習システム「あすなろ」開発の新しい展開–構文学習とその評価–, 情報処理学会第 65 回全国大会 特別トラック (6), 2003.

[2] Bilac,S. Baldwin,T and Tanaka,H. Biringing the Dictionary to the User: The FOKS System, COLING2002.

[3] 黒橋禎夫, けっこうやるな KNP, 情報処理学会誌, Vol.41, No.11, 2000.

[4] Dan Klein and Christopher D. Manning. Accurate Unlexicalized Parsing, Proceedings of the Association for Computational Linguistics, 2003

[5] Dan Klein and Christopher D. Manning. Fast Exact Inference with a Factored Model for Natural Language Parsing. Advances in Neural Information Processing Systems 15, 2002.

[6] 橋田浩一, GDA 日本語アノテーションマニュアル, http://i-content.org/gda/tagman.html.

[7] 日本電子化辞書研究所,EDR 電子化辞書仕様説明書第２版, Technical Report TR–045, 1995.

[8] 澤谷孝志, 仁科喜久子, 赤堀侃司, 日本語学習者のための Web-Concordancer の開発, 日本教育工学会第 17 回大会講演論文集, pp.469-470, 2001.

[9] 伊藤一茂, 斎藤博昭, マルチモーダル対話コーパス検索/再生ツールの実装, 自然言語処理 142-5, 2001.

[10] W3C, XML Path Language (XPath) Version 1.0, http://www.w3.org/TR/xpath.