98-01039

多言語対応理工系留学生のための日本語読解学習支援システムの開発研究

代表研究者　仁科　喜久子　東京工業大学留学生センター
共同研究者　奥村　学　東京工業大学精密工学研究所
共同研究者　杉本　茂樹　東京工業大学大学院情報理工学研究科
共同研究者　八木　豊　東京工業大学大学院情報理工学研究科
共同研究者　阿辺川　武　東京工業大学大学院情報理工学研究科
共同研究者　戸次　徳久　東京工業大学大学院社会理工学研究科
共同研究者　傳　亮　有限会社フウズラボ

1　研究の背景と目的

大学院レベルの理工系留学生にとっては論文読解も論文作成も英語である場合が多い。それにも拘わらず、研究室でのコミュニケーションは日本語が基本である。日本での留学期間が長い学生の中には、その隙間を埋めるために日本語学習を続けるという例がしばしば見られる。しかし、多くの場合は研究に忙しく、独習に頼ることが多い。また、日本に留学する学生の８割は非英語圏国からであるが、日本語学習の媒介言語はほとんど英語である。留学生は高い英語の能力をもっていることが前提とされているが、非英語圏の留学生が母語⇔英語⇔日本語というプロセスで目的の科学技術を学ぶ際、技術用語を誤読することが多く、母国ではない言語を通して学習目標言語を正しく理解するのは障害が多い。一方、留学生向けの日本語学習の現場では、それぞれ違った分野の技術系日本語を学びたい留学生が、それぞれ異なった日本語能力を持っているため、全ての留学生に対応するのは不可能であると言える。

本研究の目的は日本の科学技術を学ぶことを目的とする理工系留学生のための多言語対応日本語学習支援システムを構築することである。留学生の多くが既に母語での基本的な専門分野の概念知識をもっていることを利用して、母語から日本語が学べる読解学習支援システムとする。システムは、学習者が入力した日本語文章に対し、文章中の単語の訳と文章構造を出力することを主な機能とし、日本語の構造や初歩的な単語を知りたい日本語初級者から、単に技術用語の意味を知りたい上級者までが利用できるものであるとする。現在の日本語学習システムおよび辞書の殆どは英語訳を導くものであるが、多言語で出力できることを目標とする。また、留学生が自分のレベルにあわせて自由に自分の興味対象を読解できるように、インターネット上で利用できるものとする。

本システムの実現により、科学技術修得を目的とする理工系留学生が効率的に目的を果たすことが期待できる。また、科学技術に関するコミュニケーションがオンライン上でそれぞれの母語によって出来るようになり、日本語学習が現行の方法より魅力的で効率的になれば、日本留学も魅力的となると考えられる。自然言語情報処理・日本語学・教育工学の成果を広く利用した上で、すでに存在する多国語の電子化辞書を利用しやすく整理することや、構文辞書を利用した多義語絞込み機能の実装、文法解析における教育的な提示法の論理的な考究を行うことにより、学際的視点から各分野に新しい知見を加えることができると考えられる。

2　開発手順と成果

(1)　システムデザインと構築

学習者はインターネットを利用して、学習したい教材を教材データベースから選ぶか、学習者自身が調べたい論文などを持ち込み、文章をシステムに入力する。システムに読み込まれた文章は、形態素解析にかけられる。形態素解析プログラムは京都大学で開発されたJUMANを利用する。プログラム処理により解析された形態素が、学習者が選択した言語の辞書データベース中の一つの対訳辞書に行き、単語項目を問い合わせる。参照したい語の単語項目があれば、単語の意味が学習者の母国語で表示される。同時に、入力した文章中の任意の一文について、構文解析行う。構文解析プログラムはJUMANと同じく京都大学で開発されたKNPプログラムを利用する(参考文献５)。

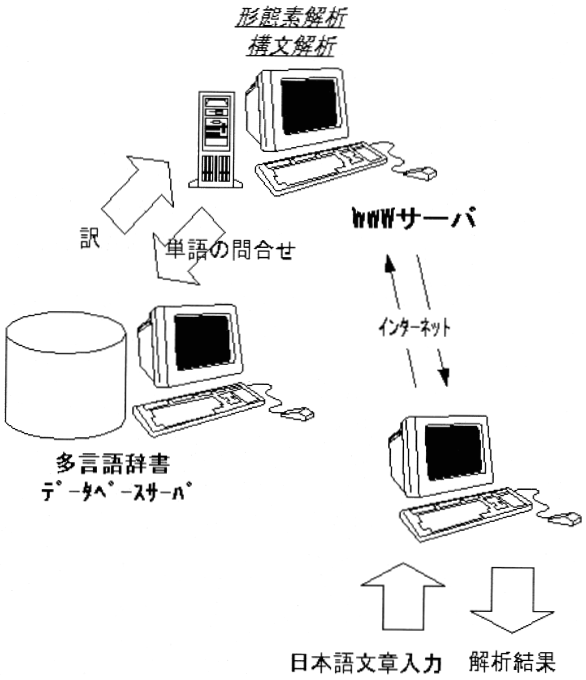

図1　システム構成

上述のシステムを構成するためには、システムサーバは形態素／構文解析プログラムを実行したのち、要素の対訳を辞書データベースから検索して応答する必要がある。解析プログラムの実行と、データベースの検索はどちらも計算機の負荷が大きく、アクセスが同時に発生するような環境下において１台のサーバで行うことには無理がある。そこで、解析プログラムを実行するサーバとデータベース問い合わせを行うデータベースサーバを分け、サーバがデータベースサーバに対訳問い合わせを行うようにシステムを構築する(図１)。すなわち、窓口のWWWサーバは解析プログラムを実行し、訳語問い合わせを他のサーバに依存して応答する。このようなシステムを構成する利点は、将来アクセスが膨大になったときにも窓口WWWサーバを増やすことで対応できると同時に、辞書データベースの内容更新が容易に行えることにある。

WWWサーバ、データベースサーバとも、一般のDOS/V計算機を用い、コストパフォーマンスの良いLinux ベースのOSを採用した。辞書データベースは、日本語・英語・中国語・タイ語・インドネシア語・マレー語の各対訳辞書と、概念辞書、日本語構文辞書などの膨大なデータをリレーション付で格納するため、商用リレーショナルデータベースにおいてコスト、高速性、信頼性で定評のあるOracle8i for Linux(Oracle社)を採用した。

(2)　インターフェースの作成

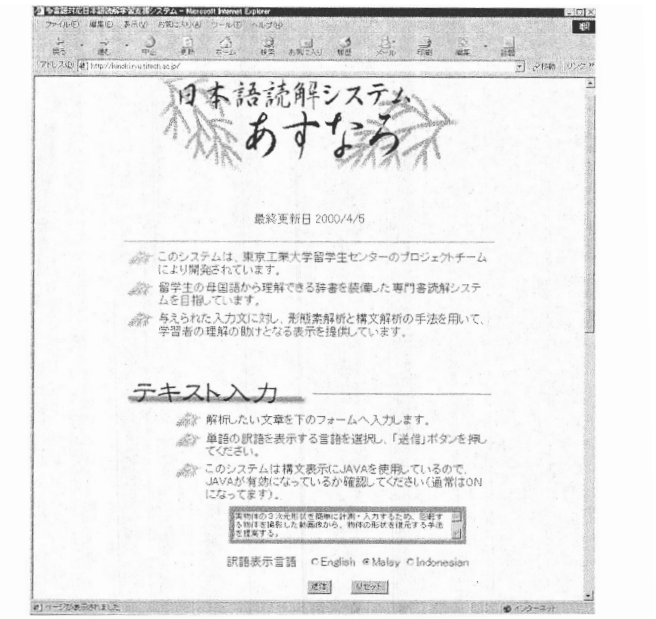

図2　初期画面

インターフェースは、学習者が入力したテキスト中の任意の１文について構文解析を行い、学習者がこのツールを利用することで、単語情報や文法構造が学べるように構成される。KNPプログラムが出力する文法構造に関する情報を、学習者にとって分かりやすく親しみの持てる形式で表示する方法として、木構造での提示を行う。文法構造を木構造で表示するプログラムはJAVAで書かれていて、インタラクティブに木構造を開閉できるようになっており、マーカーをクリックしていくと、次々に木構造の詳細が表示されていく。ユーザインターフェースを設計および作成し、システムのプロタイプを実験的にhttp://hinoki.ryu.titech.ac.jp に実装中である。

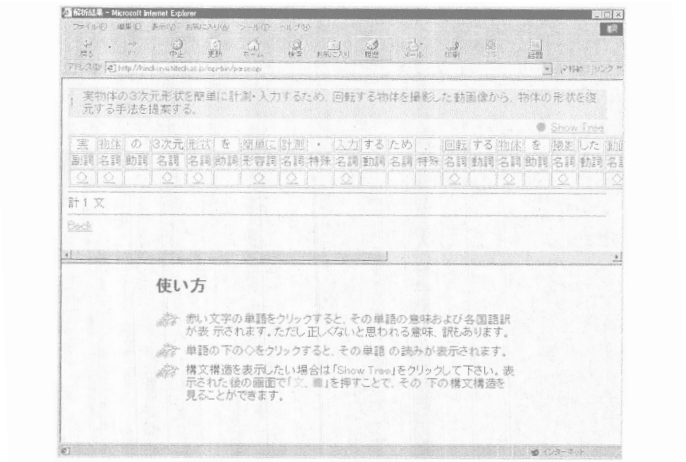

図3　解析結果画面

初期画面において学習者は、テキスト入力部に日本語のテキストをペーストまたは入力する(図２)。表示用の言語は学習者が選べるようになっていて、現在、英語・マレーシア語・インドネシア語が利用可能である。中国語・タイ語に関しては表示機能がまだ不十分なため実装されていない。

入力として、「実物体の３次元形状を簡単に計測・入力するため、回転する物体を撮影した動画像から、物体の形状を復元する手法を提案する。」という文章を入力して送信すると、JUMANプログラム(形態素解析)をおこなった結果と、解析結果の利用方法がフレームで表示される(図３)。

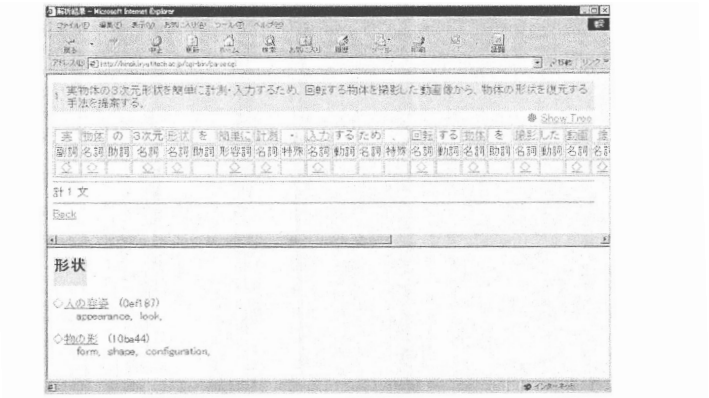

図4　訳語表示の様子

上部フレームにおいて、文章中の単語を選択クリックすると、下部フレームにその単語の訳と意味が表示される。英語の場合を図４に示した。訳は初期に選択済みの言語で表示されるが、意味は単語辞書が日本語のものしか実装されていないために日本語である。また、上部フレームの◇をクリックすると、下部フレームに日本語の読みが表示されるようになっている。

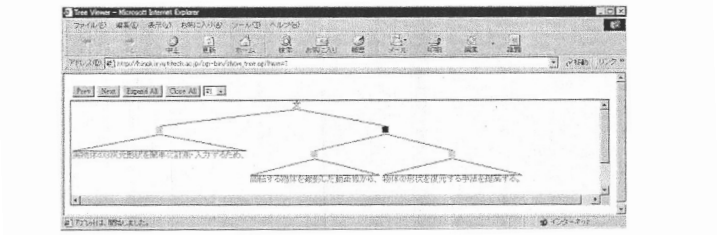

図5　木構造の中途解析結果の表示例

上部フレームに存在する「Show Tree」をクリックすると、新しいウィンドウが開き、文章構造が木構造で表示される。□をクリックすることにより木構造が開閉するようになっており、全体の文章構造を大雑把に表示したり全部開いて詳細を見たりすることができる(図５,６)。

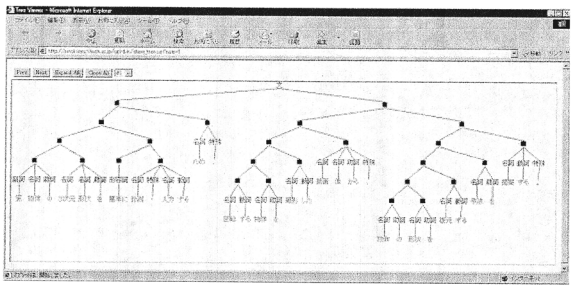

図6　木構造末端までの解析結果表示の例

⑶　既存辞書の検討と語彙の選定

システムに必要な辞書データは、日本語を多言語で表示するための多言語対訳辞書情報と、多義語の絞り込みを行うための日本語の語彙に関する充実した文法情報であり、どちらもすでに電子化されている必要がある。

多言語の語彙情報をまとめたものに、(財)国際情報化協力センター(CICC)が発行したものがあり、アジア諸国の言語に関してまとめて電子化された辞書としては、管見ではこれが唯一である（参考文献３）。この辞書は、中国語・マレー語・インドネシア語タイ語の基本語辞書と、英語専門用語に対する中国語・マレー語・インドネシア語・タイ語・日本語の対訳専門語辞書で構成される。2万5千語余りの専門語に対して対訳データが得られることは、本システムにとって大きな利用価値がある。しかし、比較的充実した専門語辞書に対し、基本語辞書には対日本語訳および概念辞書が存在しないことなどから、基本語情報をデータベースとして利用するには余りに不十分であることがわかった。

上述の通り、多言語辞書にCICC基本語辞書がそのまま利用できないため、基本語に関しては独自で辞書を作成する必要がある。辞書が電子化されていることを利用して、英語を介して対訳を結びつける方法が検討されたが、英語に訳された時点で多義語が生じ、さらに母国語に訳された時点で多義語が生じるため、ありえないものを含めた複数の対訳が出力されてしまう。

そこでCICCの基本語辞書を利用することは諦め、学習辞書としての基礎語の選択と情報系専門用語を選択することを考えた。日本語が利用できるCICC専門語辞書とIPALと日本語教育における高頻度語彙との共通語彙を調べたが、百語程度しか一致しないことがわかったため、日本語能力検定試験の１級と２級で頻度の高い動詞と名詞を選びだし、中国語に関してのみ、約1万語に関する対訳辞書を作成するに至った。
他の言語に関する対訳辞書は、現在作成中であるが、コストと時間がかかる作業となっている。

一方、日本語の語彙、文法に関するデータベースとして代表的なものに、（株）日本電子化辞書研究所(EDR)が発行したEDR電子化辞書がある（参考文献４）。日本語単語辞書だけでも26万語あり、40万概念に関する概念辞書、90万語の共起辞書などが整備されていて、概念辞書では概念上の上下関係(概念シソーラス)、共起辞書では文章中で要素がどの要素と同時に使われるかなどの文法的な情報を含んでいる。

EDR辞書が語彙を品詞に分けずに単語辞書として扱っているのに対し、IPALは動詞・名詞・形容詞を別の辞書としてそれぞれに詳細に単語情報を収録している。多義語の絞り込みを考慮した場合、IPALでは充実した意味素性と構文情報を利用した選択制限が可能であり、EDRでは意味素性と構文情報の質は劣るが、共起情報を利用した選択が可能といった特色があることがわかった。

(4) データベース設計と構築

データベーステーブルは、日本語単語辞書・概念辞書・日本語共起辞書と、英語・中国語・インドネシア語・マレー語・タイ語に対する日本語対訳辞書を作成した。日本語辞書には、本システムで学習のための難易度を表記するために、日本語能力試験で定められている級のフィールドを追加した。対訳に関しては、概念番号でリレーションをつないである。

(5) 多義語絞り込みに関する調査および研究

形態素として分解された要素に対し、辞書を引いた結果として複数の意味をもつことがありうる。それらの意味を学習者に羅列しただけでは、学習者がその要素がどのような意味で使われているのか確定できずに混乱を招く。ある要素が複数の意味をもつことを知識としてもつことは重要であるが、その文章内でどのような意味で利用されているのかを明確にし、どのような状況でその意味をもつのかの情報を提示する必要がある。本システムではまだ絞り込み機能は実装されていないが、現在、EDRの共起辞書を用いたアルゴリズムを開発中であり、問題点などの検討を行った結果を述べる。

多義語絞り込みの手法は、文中の動詞に注目し、文内の格情報をシソーラスに基づいて解析し、動詞の構文情報(使われ方の情報)に基づいて意味を決定する方法が提案されている（参考文献10、11)。具体的には、文中に現れる「〜が」「〜で」などの格に対し目的格、場所格など判断し、対象となる動詞の構文情報からその格で表現されるものを検索して意味を決定する。シソーラスを本研究調査内で整備するのは時間もコストもかかりすぎる。

豊富なシソーラス情報を電子化して公開されたものとしてEDRの概念辞書があり、この辞書では概念の上位下位を明記して概念シソーラスを構成している。また、構文情報に関して共起辞書が存在し、格情報として利用される概念がデータベース化されている。本研究において、EDRの概念辞書と共起辞書を用いた多義語絞り込みアルゴリズムを検討した。アルゴリズムと例文を用いた結果に関して以下に述べる。

共起辞書を用いた多義語絞り込みアルゴリズム

○ 使用辞書（以下は全てEDR製作）

・日本語共起辞書

日本語共起辞書（基本語)、日本語動詞共起パターン副辞書があり、ここでは動詞に着目し、後者の辞書を使用する。約5千の主要動詞に約1万4千の共起表現が掲載されている。

・日英対訳辞書

日本語単語に対して、その概念、概念説明、英訳などが収納されている。ここでは動詞の概念情報をこの辞書から得て

いる。

・日本語単語辞書

日本語単語に対して、その概念、概念説明、連接属性などが収納されている。基本的に日英対訳辞書と重複している情報が多いと思われる。ここでは名詞の概念情報を得ている。

・概念辞書

概念体系辞書、概念見出し辞書、概念記述辞書の三つが存在する。ここでは上位下位の関係が記述されている概念体系辞書を用いる。

・ アルゴリズム

1. 与えられた文に対し、JUMANプログラムを用いて形態素解析する。
2. KNPプログラムを用いて構文解析をし、本動詞に対する格パターンを抽出する。
3. 共起辞書から本動詞をキーにして共起表現を検索する。
4. 複数ある候補から格パターンに適合する単語との距離を計算する。
5. スコア順に動詞概念を表示。

多義語の絞り込みの一例として、「かむ」という動詞に関して調べたものを簡単に述べる。「かむ」は「上下の歯で食物などをくだく(ガムをかむ)」「鼻汁を出してきれいにする(鼻をかむ)」「ある事柄に関わる(プロジェクトにかむ)」「動物が歯で傷つける(手をかむ)」「歯車がぴったり組合う(動作がかむ)」の5つの意味を持つ。EDR共起辞書には次のようなデータとして表現されている。

「かむ」の共起辞書パターン

概念ID　：格パターンとその取りうる概念(=で結ぶ)

・0ea5f0(鼻汁を出してきれいにする)
　：source(を)=30f6d8： agent(が)=30f6b0
・0ea5f6(動物が歯で傷つける)
　：object(を)=30f6b0／30f6bf： agent(が)=30f6b0／30f6bf-30f746
・1faecb(ある事柄に関わる)
　：goal(に)=30f7e4： agent(が)=30f6b0／30f746
・3bcc43(上下の歯で食物などをくだく)
　：implement(で)30f6d8
　：object(を)=3f9639： agent(が)=30f6b0／30f6bf
・3bf986(歯車がぴったり組合う)
　：object(が)3aa92f

EDR共起辞書を用いた場合、共起表現の格パターンの取りうる概念IDは個々の物体ではなく、抽象化された概念で表記されているため、文から求めた名詞のIDがそのまま一致することはほとんどない。そこで名詞のIDから上へたどり一致した概念までの距離を共起格との距離として導入する。例えば、文章中に「人が」を見つけた場合には、まず「人」の概念IDを調べ、このとき共起表現が格と一致していれば距離を0とする。次に概念体系の上に向かって調べていく。共起表現における一番目の行の 「agent(が)＝30f6b0」 との距離は3である。一番上の概念まで検索してもマッチしないときは距離を10 とした。また、格が存在しない場合は距離を求めることができない。以上をふまえ、ここでは、距離が近いほど高い評価値を出すように次のような計算をおこなった。

格パターンの評価値 ＝ 10 - 格パターンの距離

したがって概念がマッチしなかった場合や格が存在しない場合の評価値は0となる。最終的にはそれぞれの格パターンの評価値を足し合わせる。

○ 例文を用いた実験例

例文をいくつか入力して「かむ」の多義語絞込みをおこなった。入力サンプルの質と量に関する検討が必要であるが、簡単な文章を入力した結果の正解率は70%程度である。

正しい結果の例

「人が鼻をかむ」

評価値	概念ID	評価値詳細	概念説明
17	0ea5f0	鼻=8 人=9	鼻汁を出してきれいにする
9	3bcc43	鼻=0 人=9	上下の歯で食物などをくだく
9	1faecb	人=9	ある事柄に関わる
9	0ea5f6	鼻=0 人=9	動物が歯で傷つける
0	3bf986	人=0	歯車がぴったり組合う

ちなみに「かむ」で格パターンに他の名詞を採用すると誤った結果が見られる。

誤った結果の例

「犬が手をかむ」

評価値	概念ID	評価値詳細	概念説明
14	0ea5f0	手=7 犬=7	鼻汁を出してきれいにする
14	0ea5f6	手=7 犬=7	動物が歯で傷つける
7	3bcc43	手=0 犬=7	上下の歯で食物などをくだく
7	1faecb	犬=7	ある事柄に関わる
0	3bf986	犬=0	歯車がぴったり組合う

「犯人が事件にかむ」

評価値	概念ID	評価値詳細	概念説明
5	0ea5f0	犯人=5	鼻汁を出してきれいにする
5	3bcc43	犯人=5	上下の歯で食物などをくだく
5	1faecb	事件=0 犯人=5	ある事柄に関わる
5	0ea5f6	犯人=5	動物が歯で傷つける
0	3bf986	犯人=0	歯車がぴったり組合う

○ 考察

「犬が手をかむ」の例に関しては、「動物が歯で傷つける」と「鼻汁を出してきれいにする」が同じ評価値を得る。常識的に考えて、不自然である。「動物が歯で傷つける」と「鼻汁を出してきれいにする」にとって、「犬」という主格と「手」という目的格がそれぞれ同じ評価値をもつ事は問題である。これは、共起辞書における共起概念が抽象的であるために、主格としての「犬」／「人」も“生き物”の概念であり、目的格である「手」／「鼻」も“体の一部”の概念となっていて、「動物が歯で傷つける」と「鼻汁を出してきれいにする」に記述されている抽象的な共起概念からみれば、どちらも同じ距離になってしまうからである。「鼻汁をだしてきれいにする」の目的格の共起概念として、「鼻」を明示的に記述しておく必要があり、むしろその上位の概念を記述する必要はないといえる。

「犯人が事件にかむ」に関しては、「事件」が「ある事柄に関わる」における目的格の共起概念には含まれていない。すると「〜が」格だけが選択の情報源となるが、「犯人」は特別な共起表現ではないため、それだけでは判別できない。「事件」を「ある事柄にかかわる」の目的格の共起概念に明示しておけば、「かむ」との共起における「事件」の評価値は大きいものとして得られるであろう。

したがって、正解率が期待通りに高くない原因は、共起辞書における概念が概念体系の上位に属するの抽象概念に限られることにあり、使われるべき対象を具体的に表現していないからと言える。もっとも、EDRの共起辞書は文章における可能性を全て網羅するために記述された包括的な辞書であるため、意味の絞り込みを行うために適していないと考えられる。

○ 課題

共起辞書を用いて絞込みを正しく行うには、共起辞書における共起概念をなるべく下位のものとなるように構成しなおす必要がある。この場合、それぞれの格で利用される可能性のある共起概念を細かく明示する必要があり、相当な作業量となることが予想される。これとは別に、「貨幣・紙幣」の概念に「お札」が含まれないなどといった概念の問題や、英語を基本にした概念体系であるため、日本語にしか存在しない概念が概念辞書に存在しないなど、概念体系内の問題も明らかになってきており、要素に対して新たに概念を加えるなどの概念体系の不備を補う作業が必要になると考えている。EDR

の共起辞書から多義語絞り込みを行うには精度が良くないことから、コーパスを利用した手法も検討中である。工学的な処理技術向上の努力はさることながら、教育現場での学習者への対処は時を待てない。意味の曖昧性の解消法として,学習者には出来るだけ多くの例文を示すことも考えている。そのためには学習者のレベルと日本語学習の対象領域（専門分野）とが合致するようなコーパスを整える必要がある。

(7)　学習シラバス作成

(a)テキスト

学習者がシステムに入力する為の素材となるテキスト文章を作成した（参考文献８）。最終的にはWWW上の科学的な文章が読めることを目標とするが、その事前段階として、日本語学習者が学びたい文章を選択するための基礎的な題材を提供することと、学習者のシステムに対する感触を得るためである。国際教育協会主催の日本語能力試験出題基準に現れる語彙と文型を学習シラバスの標準とし,特に理工系留学生に必要な項目を選定し、調整した。テキストは第１課：インターネットホームページ、第２課：電子メール、第３課：実験をする、第４課：太陽と地球と月、第５課：日本のエネルギー問題、といった科学的なトピックを扱っている。これらの内容は、名詞・動詞・形容詞などの品詞に関して初級で学んだことを復習しながら、科学技術用語と科学技術文によく見られる文型(助動詞相当句)や文章の型(定義・比較・分類・列挙・原因・理由・条件)などを学習できるように構成した。

(b)練習問題

学習者が自習するための練習問題を作成した。ディスコースの理解ができたか否かの問題が必要であるが、現在可能な学習機能からは効果が図りにくい。現状では、(ア)漢字の読み、(イ)単語の認定、(ウ)単語の意味取得、(エ)構文理解（どの語句からどの語句までがひとかたまりか、それらの修飾関係はどうか等)のレベルは現段階では学習可能で、評価も可能と考えられる。上記のテキストには練習問題を入れた。この練習問題はオンライン上でそのまま利用することを想定している。練習問題作成という予定の目標は中級用については達成した。

(b) 学習履歴

学習履歴は、学習者が学習成果を知るための履歴と、対訳システムのログとして履歴が必要である。現在、システムログとしてどの単語が多く引かれているかの統計情報を記録するようになっているが、利用者単位でその情報を利用できるようになっていない。また、練習問題がオンラインに実装されていないために、問題に関する履歴もとられていない。これらは今後実装する予定である。

3　今後の課題

対訳機能に関しては、対訳辞書の作成をCICCに記述された2万5千語に対してすべて行い、専門語だけではなく、基本語に関しても対訳を表示できるようにすること、および、多義語絞り込み機能を改良することが当面の課題である。学習機能に関しては、すでに作成されたテキスト・練習問題をオンライン上に実装し、学習履歴をとるプログラムを実装する必要がある。

また、教育工学的なシステム評価実験を行い、本システムを用いることによってどの程度学習成果があがるか正しく評価することが必要であると考えている。本システムを使わない場合と使った場合に、学習効率がどのように変化するのかに関し、12年7月末に日本語学習者（初中級レベル）留学生を対象に学習評価実験を行う予定である。

4　特記事項

昨年11月13日には留学生センター5周年記念シンポジウム「留学生教育とテクノロジー支援」において「留学生教育と情報技術―日本語日本語学習システムの紹介―」と題して分担者奥村学が講演および本システムのデモンストレーションを行ったことに、多くの反響を得たことは特記すべきことである。

参考文献

(1)　情報処理振興事業協会技術センター（1990）計算機用日本語基本形容詞辞書IPAL（Basic Adjectives）

(2)　国際交流基金・財団法人国際教育協会（1993）『日本語能力試験出題基準外部公開用』国際交流基金

(3)　Machine Translation System Laboratory Center of the International Cooperation for Computerization (1995) Basic Dictionary

(4) （株）日本電子化辞書研究所（1994）EDR電子化辞書利用マニュアル 第2.1版
(5) 仁科喜久子（1999）多言語対応専門日本語読解学習支援システムの構想について 専門日本語教育研究創刊号 No.1 pp.40-43
(6) 仁科喜久子・奥村学（1999）科学技術日本語学習支援多言語対応辞書の項目設定に関する研究 日本語教育方法研究会誌 Vol.6 No.2 pp.32-33
(7) 仁科喜久子・奥村学（2000）「やさしい科学技術日本語読解入門」－多言語対応オンライン科学技術日本語学習支援読解教材としての利用法－ 日本語教育方法研究会誌 Vol.7 No.1 pp.16-17
(8) 仁科喜久子編（1999）やさしい科学技術日本語読解入門（試作版）東京工業大学留学生センター
(9) NTTコミュニケーション科学研究所(1997)日本語語彙大系
(10) 白井清昭、乾健太郎、徳永健伸、田中穂積(1998)統計的構文解析における構文的統計情報と語彙的統計情報の統合について。自然言語処理、Vol. 5、 No. 3、pp. 85-106
(11) 松本祐治・徳永健伸・奥村学・杉浦芳樹・大林正晴（1998）言語活動を知的支援するための辞書の作成・利用技術の開発　創造的育成事業最終成果発表会論文集知識情報処理・エージェント技術　情報処理振興事業会