研究科委員長印
㊞

２０１４年度
大学院奨励研究員研究報告書

年　　月　　日

関西学院大学教務機構長　殿

奨励研究員

氏　　名	田中良　　印

指導教員

所属・職名	
氏　　名	印

以下のとおり、報告いたします。

研究課題	多言語対応コンコーダンサー『HASHI』 ―日本語と日本語教育と社会言語学の研究を中心に―
採用期間	２０１４年　４月　１日　～　２０１５年　３月　３１日

研究科受付印	教務機構受付印

研究発表状況（奨励研究員採用期間内に発表したものおよび、近く発表予定のもの）

（１）学会誌等への発表（著者、発表論文名、学会誌名、巻号、発表年月、掲載頁等）

雑誌論文	著者名	田中良	論文題目	日本語研究の一般的手法を機械上で実現するための言語分析ソフト『HASHI』

		巻号	発行年月	掲載頁
雑誌名	日語研究	第10号	2015年予定	

雑誌論文	著者名		論文題目	

		巻号	発行年月	掲載頁
雑誌名				

図書	著者名	于康・田中良	論文題目	

		発行年月	頁
書名	中国語作文添削と指導―タグ付けプログラムTNR	2014年10月	総頁：201
			担当箇所：

※論文題目：共著の場合の担当部分のタイトル

（２）学会発表（口頭・ポスター：学会名、開催地、発表論文名、発表年月日等）

学会名	第６回日中対照言語学研究会	開催地	北京、人民大学
題目	日本語学習者の作文への添削結果のコーパス化と利用	発表年月日	2014年８月20日

学会名		開催地	
題目		発表年月日	

学会名		開催地	
題目		発表年月日	

研究経過状況（３０００字程度）

奨励研究員採用期間内における研究の進展と結果に関し以下のことが挙げられる。

【研究の背景】

まず、本研究の土台となる要素として、「コーパス」が存在する。コーパスとは、実際に使用された言語を大量に収集しデータベース化したものであり、これを調査することで、これまでのような少量の用例の収集では行えなかった、大量の用例からのみ発見できる言語の実態を、質的にも量的にも解明することができるものである。このコーパスを分析するためのソフトウェアを「コンコーダンサー」と呼ぶ。

【問題点】

本研究は、コンコーダンサーにおいて、現在多く求められているにも関わらず実現されていなかった問題を解決することを目的とする。

研究計画書にて提示した問題点は以下の通りである。

（１）収集した用例に対し、各研究者が独自のタグを付与することで、あらゆる研究の視点からの分析が行えるようにする。その際に、同時に各分野内で必要とされるタグは共通したものが多いと考えられるため、雛形となるタグリストを開発する必要がある。

（２）タグの付与の際にデータの形式が２種類考えられる。語の区切りの無い文字列のものと、語の区切りがある形態素解析のものである。これの両方の利点を併せ持つことで、言語にまつわるあらゆる情報を扱えるデータ形式が必要である。

（３）本研究が主に対象とする言語である日本語以外にも、代表的言語である英語、中国語、韓国語などさまざまな言語に対応させることで、言語を問わず扱える一般性を獲得する必要がある。

（４）本研究で提示するコンコーダンサーは大幅に多機能なものであるため、そのすべてを一気に提示することは使用者にとって分かりにくさにつながるため、初歩的なものから順に機能を提示し、それぞれのソフトにまとめる必要がある。

研究の進展に伴い、これらの問題点に対し、新たに起こってきた問題点や課題を合わせてまとめ直し、本研究の目的を以下の５点の解決とした。

①分析対象となる表現が使用される用例の、自在で簡易的な方式での収集

単独で使用される語や、複数の語が完全に決まった並びで使用される表現の収集は既存の方法で十分可能である。それに対し、「～が ～に 動詞」などの不特定の語が間に入る文型の収集は難しい。複雑な正規表現や位置ごとに語を個別に指定するなど、使用が複雑なうえ、検出結果も不完全なものである。

②収集した用例に対してのあらゆる研究視点からのタグの、効率的な付与

現状の研究では用例の分類はほぼ手作業で行われている。研究のさまざまな視点からの分類にはタグが効果的であり、まさに効率、精度、先入観の３点の問題の解決に直結する。しかしこれを効果的に行えるものが無い。

③計量的分析処理による、言語の傾向性の自動的な抽出

コーパスを主体とした研究では統計をはじめ計量的な処理が非常に重視されている。しかし単独の語の傾向性や、対象の語と周囲の語との関係性を測ることはできるが、対象の語が使用される際にどのような文型の中で使用されるかを効果的に、主観や経験に頼らずに抽出する方式が存在しない。

④研究の立場や興味に即したデータの作成

より深く各人の研究の立場や視点に根ざして行うためには、根本となるデータであるコーパス自体を対応させる必要がある。まず日本語では語を区切る単位を規定する必要があるが、研究の立場や研究したい内容によって有効な単位は変わる。これへ深く対応できるものが無い。また無数にある各人の持つ研究対象への興味にすべて答えたコーパスが存在しない以上、それらは研究者各人が付与する必要がある。そのためにコーパス全体へのタグ付与が必要である。しかしこれを効果的に行うソフトは現在無い。

⑤一般的な言語研究手順のすべての工程の、統合した仕組み上での実現

言語研究の手順の各工程を、それぞれ別の独立した方式で順に行うと、効率が悪く、整合性について難点が出やすく、データの連動の点で不安定になりやすい。

【解決内容】

以上の問題点に関し、以下のように解決を行った。

1）「～が ～に 動詞」のような一般的に用いられる文型の表し方と非常に近い書き方で検索を行えるようにしたことで、使用者にとって理解しやすくなった。また、目的の表現を高精度に収集することができるようになった。

2）タグ付与ソフトが使用されない原因として、使用の複雑さや使用目的を完全に叶えたものでないことがある。具体的には次の点がある。基本的にキーボードからのタグ入力になりミスが生じやすい点、特定の語には決まったタグが付与される場合の対応がない点、各研究に即したタグが自由に設定できない点、付与したタグを利用する方法が弱い点である。これらを解決した結果、以下のような仕組みとなった。あらかじめ作成したタグリストからの選択式での簡易的なタグ付与、特定の語への機械による自動判別でのタグ付与、自由なタグ設定、付与したタグでの検索や集計である。これにより、研究者の独自の研究視点からのタグを自在に付与できるようになり、用例の分類による研究の精度の向上を果たせることになった。

3）ある語が実際に使用された文の集まりである KWIC と、ある語と周囲の語の位置関係の傾向を数値で知ることができる Picture を合わせ、その両方の特性を持った処理として「POPAK」を開発した。これにより自動的に文型が浮かび上がる。これで、例えば動詞「なる」を調べた場合、「の ○ が ○ に なった 。」という文型が多く使用されるなどということが分かるようになった。

4）コーパスを研究者の立場や興味に対応したものにするためには、「語」の単位の規定と、研究視点の付与が重要である。これに対応するため、日本語の語の単位として仁田（1997）、庵（2012）、短単位などさまざまなものを各研究者の立場に合わせて選択し、また研究に最も効果を発揮するように微調整できるようになった。さらに、そのように区切られた語に対し、研究の視点や興味からの情報をタグとして付与できるようにし、完全に各研究者の研究の立場や目的に深く合致するコーパスが作成できるようになった。

5）既存の課題、①～⑤をすべて解決し、その結果が図 5-1 のようになった。課題①の解決結果が『ToriBASHI』、課題②の解決結果が『SaiBASHI』、課題③④の解決結果を『HASHI』というコンコーダンサーになった。

以上の研究結果を博士論文にまとめ、2015 年 3 月に提出をした。

以　上