嘗試將人工智慧資訊處理技術教育導入日本語言文學系之考察

落合由治

淡江大學日本語文學系　特聘教授

摘要

當前全球冠狀病毒疫情持續蔓延，對經濟社會產生了巨大影響。雖此眾人期盼的產業，人工智慧相關的資訊通信技術也帶來了便利的影響。本論文主要從以下 3 方面來探討 AI 技術在台灣日語教育中的應用現狀和挑戰。

(1) 新設課程中導入資訊處理技術。以 T 大學為例介紹各種新設課程改革之概 況。

(2) 將應用程式引導入課堂當中。將介紹可應用於課堂上的各種應用程式，以補充或改進教學方法。

(3) 程式語言之導入。從人文科學的角度出發，將程式教育作為自然語言處理技術教育的一部分，導入人文社會相關學科科學課程。並從中找出問題點以及成果。

考察結果顯示，應對人工智慧時代在台灣實施日語教育時，需要面對以下三點。第一是要為多語言環境下的程式語言教育做準備，第二是找出善用人文社會學科的特點，第三是學用無落差的實務能力培育。在活用理工科技術方面的語言處理訊息技術教育的同時，必須考量如何將此應用回歸至現今日本語文學系相關能力的培訓課程當中，以期盼依此考量在新時代中找到一個新的方向。

關鍵詞：人工智慧資訊處理技術教育、人文社會學科、課程、應用程式、電腦程式

受理日期:2022 年 02 月 22 日

通過日期:2022 年 05 月 13 日

DOI: 10.29758/TWRYJYSB.202206_(38).0001

An Attempt to Introduce Information Processing Technology Education to the Department of Japanese Language and Literature

Yuji Ochiai

Distinguished Professor, Department of Japanese, Tamkang University, Taiwan

Abstract

Currently, the global epidemic of coronaviruses continues to spread and have a great impact on the economy and society, and AI-related information and communication technology also has a great impact as an expected industry. This paper focuses on the current status and challenges of AI technology application to Japanese language education in Taiwan from the following aspects.

(1) Introduction of new courses into the curriculum: An overview of the various new curricula that have been created at the university where I work will be presented.

(2) Introduction of applications to the classroom: Various applications that can be used in the classroom to supplement or improve teaching methods will be introduced.

(3) Introduction of programming education: The course, problems, and results of introducing programming education as part of information processing technology education in the humanities and social sciences from the perspective of the humanities will be introduced.

As a result of the discussion, it can be said that three directions are necessary for Japanese language education in Taiwan to respond to the AI age. The first is to prepare for programming education in a multilingual environment, the second is to find a direction to utilize the characteristics of the humanities and social sciences, and the third is to relate it to practical business skills. While making use of information technology education in the sciences and technology, we would like to find a new direction for the new era by determining where to place the various information technologies in the existing training curriculum for Japanese language skills in Japanese-language related departments.

Keywords: information processing technology education, humanities and social sciences, curriculum, application, programming

AI 情報処理技術教育の日本語文学科への導入の試み

落合由治
淡江大学日本語文学科　卓越教授

要旨

現在、コロナウイルスの世界的流行が続き、経済や社会に大きな影響が広がっているが、期待される産業として AI 関係の情報通信技術は大きな影響を与えている。本論文では、以下の面から、台湾の日本語教育への AI 技術応用について現状と課題を取り上げた。

(1) 新設科目によるカリキュラムへの導入：勤務先の大学では各種の新しいカリキュラムの創設を試みた、その概要を紹介する。

(2) 授業へのアプリケーションの導入：授業方法の補充や改善を試みることが可能な、授業で利用できる各種のアプリケーションを紹介する。

(3) プログラミング教育の導入：人文社会系での情報処理技術教育の一部として人文系の視点でプログラミング教育を導入した課程と問題点、成果について紹介する。

考察の結果、台湾の日本語教育が AI 時代に対応するには、三つの方向性が必要と言える。第一は多言語環境でのプログラミング教育の準備、第二は人文社会系の特色を活かす方向性の模索、第三はビジネス的実用的スキルとの関係である。理系・技術系の情報技術教育を活かしながら、各種ある情報技術を日本語関係学科の日本語能力の今までの訓練カリキュラムのどこに位置づけるか、新しい時代の方向性を見いだしていきたい。

キーワード：AI 情報処理技術教育、人文社会系、カリキュラム、アプリケーション、プログラミング

AI 情報処理技術教育の日本語文学科への導入の試み

落合由治
淡江大学日本語文学科　卓越教授

1.はじめに

現在、コロナウイルスの世界的流行が続き、経済や社会に大きな影響が広がっているが、期待される産業として AI 関係の情報通信技術は、大きな影響を与えている。その一方で、技術の質的停滞も見られるようになり、今後の社会的応用に調整が必要になっている。成果が上がりやすい分野は、数値データやマルチモーダル情報データなどビッグデータの関係した処理分野であるが、人間の言語に関する自然言語処理では意味的処理の困難が新しい課題となって、試行錯誤が続いている。本論文では、以下の面から、日本語教育への AI 技術の応用について現状と課題を取り上げて見ていきたい。

（1）新設科目によるカリキュラムへの導入：勤務先の大学では、AI 関係技術を大学教育での次の中心的技術とするよう、各種の新しいカリキュラムの創設を試みている。その概要を紹介する。

（2）授業へのアプリケーションの導入：現在までのところ、直接、AI 関係技術を授業で応用して授業内容にできる状態ではないが、アプリケーションの利用を通じて、授業方法の補充や改善を試みることが可能である。授業で利用できる各種のアプリケーションについて紹介する。

（3）プログラミング教育の導入：台湾の人文社会系の学科で第三世代の AI 技術を教育や研究に結びつける方法は模索の状態である。現在の技術は、今まで利用されてきた第二世代までのコーパスや ICT 関係技術とは質的に異なり、また社会的応用もすでにはるかに広範囲に及んでいる。[1]実質的創造的な応用ができる方法として、情

[1] 第三世代の自然言語処理は、すでに今まで人文社会系の仕事だった、新聞、雑誌、Web サイト、事務、会議、文書作成、イラスト制作、電話受け付けな

報処理技術教育の一部として人文社会系の視点でプログラミング教育を導入した課程と問題点、成果について紹介する。

なお AI 技術は、狭義では第三世代の深層学習等の自己学習機能の発達に関わる技術や人間的活動ができる強い人工知能の技術開発等を指すが、ここでは広義に第一期から第三期までの自然言語処理に関わる情報処理技術の意味で用いることにする。[2]

2.新設科目による情報処理技術教育のカリキュラムへの導入

以下、試行錯誤の段階ではあるが、まず今まで勤務校で行ってきた AI 技術の人文系教育への導入と応用の実践について紹介していきたい。[3]

2.1 AI 技術の基礎について学ぶ勉強会やワークショップの開催

最初の試みは、教員の学習と現状認識を進めることから始めて、107 学年度秋から、講師を選んで定期的に、現在までの情報処理技術の発展と現在の第三次 AI ブームのトレンド、自然言語処理中でテキストマイニングを中心に学ぶ勉強会とワークショップを台湾日本語教育学会を通じて定期的に開催してきた。勉強会ではテキストマイニング技法と日本語･中国語でのそれぞれの処理を中心に学び、まず研究への応用を模索した。[4]

ど、言語を扱う分野に浸透してきている。日本での一例として TechCrunch（2021）「AI チャットボット「りんな」の rinna と UneeQ を日本展開するデジタルヒューマンが協業、顔・声・視聴覚を持つ雑談 AI 実現」https://jp.techcrunch.com/2021/05/31/rinna-uneeq-digitalhumans/。ZOOM の会議内容を自動で文字化して、商談化できる文章生成システムも実用化され、販売されている。TechCrunch（2021）「Zoom 商談を書き起こし Salesforce に自動入力するオンライン商談自動化ツール「アンプトーク」が発売開始」https://jp.techcrunch.com/2021/09/06/amptalk-fundraising/。（以上、2021 年 10 月 2 日閲覧）。

[2] 情報処理技術、人工知能技術の歴史については、総務省（2016）『平成 28 年版情報通信白書』第一部第四章を参照。http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/（2021 年 10 月 2 日閲覧）。

[3] 学科と大学での情報関係教育の導入と成果については落合由治、曾秋桂、王嘉臨、葉麥（2020）「人文系教育への情報処理・自然言語処理技術の導入と応用」『淡江日本論叢』42pp.45-63、曾秋桂（2021）「日本語教育のつながりとひろがり―AI と HI を兼ね備えた外国語（日本語）人材 2.0 の育成を目指して」『日本語教育研究』54,pp23-36 を参照。

[4] 今まで接点のなかった第三世代 AI 技術の導入は、中国語・日本語で処理ができる、ビッグデータのテキストマイニングを行っている淡江大学情報経営学

2.2 日本語教育関係学会での AI 技術関係講演者による講演と研究発表

現在まで 4 回、AI 技術と日本語教育、日本関係の人文系研究と教育との接続を試行する学術シンポジウムを開催し、日本と台湾から直接、自然言語処理技術の開発を行っている専門家を招聘して、AI 技術の発展と可能な処理、現在の開発状況の概要について状況を教示していただいた。新しい分野の概要と、第二世代までの AI 技術を応用した分野を包括する現在の AI 技術の概要について理解できた。また、AI 技術の人文社会系研究や教育に応用する研究発表を募集して、新しい時代の AI 技術の応用について討議をおこなった。

2.3 大学の授業としての外国語学部生向けの入門授業の開設

以上の学習と研究活動を活かし、台湾教育部の AI 教育推進の動きと連動して、勤務校では学科領域を超えた情報教育課程が試行されている。109 年 9 月からは、日文系主導で外国語学部と技術系学部の協力を得て、教養課程での学部生向けに AI 技術啓発の教養講座授業「AI と外国語学習」（外国語文学部教養課程 2 単位）を開設して、110 年 9 月から 2 回目の開講となった。大学外部講師（2 名、うち 1 名がデータサイエンティスト）、内部講師（12 名、そのうち、他学部の講師 6 名）を合わせ、14 名の講師が連携し、開講する形を採った。昨年は、18 週間一学期の授業の前半で、情報処理技術を専門とする工学部の教授に、AI 技術に関する機械学習、ディープラーニングなどの基本概念と応用、また技術を応用したゲーム（「Thing translator」、「AutoDraw」等）および、AI と社会変化を考えるレポート課題などを通して、学部生に新知識伝達とアプリケーション操作、AI 的思考力、AI 技術の応用力について学んでもらった。1 学期間の成果は、勤務先の情報センターが開発した iclass や PE 図（Performance Engagement）でデータとしても蓄積され、各週の受講者の授業参与と進展などの学習状況を把握することができた結果、

科魏世杰教授に指導を仰ぎ、今まで扱ったことのなかった情報学関係の新しい方法を学んだ。これによって、以降で紹介するような一連の学習活動また教育と研究への入り口が生まれた。理系と人文社会系の両方に接点のある協力者を確保することが極めて重要と言える。

受講者のモチベーションは向上しており、意欲的に新しい講座授業に参加していたことが分かった。[5]

3.授業へのアプリケーションの導入

現在まで、勤務先の葉麦、曾秋桂、王嘉臨、落合由治の 4 名の教員がテキストマイニングを質的研究として人文系研究教育に応用する試行について定期的に研究発表を行い、台湾の学術雑誌に投稿を行っている。[6]また、AI 関係テーマのシンポジウムでの研究発表も行われて、研究と授業への応用が模索されている。[7]そうした試行を通じて、授業での AI 技術の導入の試みも可能になってきている。

応用できる分野は 2 つあり、ひとつはテキストマイニングの応用で、UserLocal などの無料ツール、KH Coder などの PC アプリケーション、BI ツールの試用などの方法が可能である。[8]授業で最も簡単に使えるのは UserLocal などの無料ツールで、ニュース記事、評論記事など読解する場合、時間と手間がかかり、また内容の確認や発展的課題も容易ではない素材について容易にキーワードを重要度に応じて抽出し、そこから全体の中でまず理解すべき要点を抽出できる。2021 年 5 月からのコロナの第二次流行の時期、オンライン化された授業の中で 3、4 年生向けの授業で方法を解説し、実演して学生に操作をしてもらい、要点の抽出ができた。

---

[5] 授業の詳細は、曾秋桂（2021）「AI の学習効果についての一考察－「AI と外国語学習」講座授業を中心に」『淡江日本論叢』43pp.114-133 参照。

[6] 詳細は注 3 参照。

[7] 近年の教育応用への試行については、シンポジウムの予稿集、淡江大学日本語文学科（2021）『2021 年 AI と日本語教育国際シンポジウム―クリエイティブラーニングを目指す AI と日本語教育』淡江大学日本語文学科参照。

[8] インターネット上で使える無料テキストマイングは、UserLocal https://textmining.userlocal.jp/（2021 年 10 月 2 日閲覧）。テキストマイニングによる文学読解研究のまとめとして葉麦（2020）『AI 技術による村上春樹文学の再読―短編小説を起点として』瑞蘭國際有限公司、読解授業への応用について、王嘉臨（2021）「Amazon Comprehend 感情分析を用いた読解指導―詩教材を中心に」『淡江日本論叢』43pp.49-64、曾秋桂（2021）「グローバル時代のエコフェミニズムの視点から読む多和田葉子の『星に仄めかされて』―「神の子どもたちはみんな踊る」の意味をめぐって」『台大日本語文研究』41pp.21-40 等、一連の現代文学研究の成果がある。日本語学への応用として落合由治（2020)「日本語関係人文系研究の質的研究におけるテキストマイニング手法の応用と課題」『台大日本語文研究』39pp.101-130 参照。

もう一つは、各種のスマホアプリの利用で、信頼性があり、各種の応用が可能で、学生の自主学習や練習に応用できる。

①自動翻訳：DeepL、Papago＋Google：DeepL、Papago は日英、英日では高精度で実用レベルがあるが、日中、中日はまだ十分とは言えない場合がある。Google は精度は問題が多いが、正体字と簡体字の切り替えなど補足的な試用が便利で、各種の作業に使える。テキストの音声読み上げもできるので、発音の参考にできる。翻訳の校正をすることで、学習や読解に応用できる。[9]

②音声翻訳：VoiceTra：日本でオリンピックの機会翻訳用に開発されたアプリで、31 カ国語の音声で相互に翻訳ができる。音声の聞き取り精度が高いので、発音を入れて正しく認識されるかを試し、発音練習に使えるほか、中国語で話した内容を日本語に翻訳することで、表現辞典としても使用できる。[10]

③言語交換：Hallo Talk、HI native はネイティブと学習者を結ぶ SNS ソフトで、文字や音声を通じてやりとりして、相互の言語知識を交換したり、質問したりできる。作文の添削などもパートナーに依頼できる。コロナ流行で海外との交通は途絶しており、日本人と各国の学習者を結びつけるアプリは重要になる。[11]

④音読：OJAD はオンライン日本語アクセント辞書で、単語などのアクセントを調べ、発音することができる。韻律読み上げチュータスズキクンは文章でのプロソディーを調べることができる。Ondoku はテキストの音読ができる。[12]その他、Google 翻訳など機械翻

---

[9] 深層学習を利用した AI 翻訳は精度が上がってきている。DeepL https://www.deepl.com/ja/translator はビジネスで使用され、韓国の Naver Papago https://papago.naver.com/は韓国では広く使われている。一方、Google 翻訳 https://translate.google.com/?hl=ja は、欧米圏は別として他言語での精度は現在の他のアプリにくらべるとすでにそれほど高くないが、豊富な言語がある点でツールにできる（2021 年 10 月 2 日閲覧）。

[10] Free バージョンは VoiceTra https://voicetra.nict.go.jp/、現在、日本の多くの音声翻訳製品や音声認識製品の基礎技術になっている。ビジネス向け案内 https://gcp.nict.go.jp/news/products_and_services_GCP.pdf 参照（2021 年 10 月 2 日閲覧）。

[11] スマホアプリで手軽に使用できる Hallo Talk https://www.hellotalk.com/?lang=ja など、交流アプリの活用方法は工夫の余地が大きい（2021 年 10 月 2 日閲覧）。

[12] OJAD http://www.gavo.t.u-tokyo.ac.jp/ojad/は、初級、中級日本語教科書の単語

訳サイトも添附した文章を音読できる。

⑤グループワーク：シートを共有する形で討論や作業ができる Padlet は、コロナ流行下でオンライン授業の場合にも、共有の場を作る手助けになる。[13]

日常的なアプリの利用を習慣にすることで、発音などの練習ができ、また教室での学習にとどまらない自主学習を進める補助ができる。

4.プログラミング教育の導入

大学院での情報処理技術教育導入の基礎に関する授業の開設の試みとして、108、109 の 2 年間、勤務先の日本語文学科修士課程で、「人文社会研究におけるＡＩ日本語言語処理の応用」科目を開設し、現在までで 1 週間 2 時間、4 学期の授業を修士課程の学生に実施した。内容は、情報技術の発展、自然言語処理の基本的技術である文字コードシステム、形態素解析と KH Coder、Python を使ったデータマイニングの方法を中心にしたもので、データ収集、形態素解析の方法や設定、各種の前処理、データの読み取り等、人文系の質的研究にも活かせるように学習者と実習し、日本でのテキストマイニングコンククールである NTT「学生研究奨励賞」[14]にも出せるレベルで研究報告をまとめることができた。

現在までのところ、まだ試行錯誤の段階であり、何を授業に取り入れればよいか定見が形成できている状態ではなく、同時に、果たして今後の卒業生の進路開拓に有益な内容なのかどうかも分からない状態ではあるが、AI 技術全般のプログラミングやプログラム操作

について詳細なアクセントを調べることができる。韻律読み上げチュータスズキクン http://www.gavo.t.u-tokyo.ac.jp/ojad/phrasing は、AI 判定で、文や段落単位でプロソディー（韻律の高低）を表示し、合成音声で発音してくれる。（2021 年 10 月 2 日閲覧）。

[13] グループ活動の連絡に使える padlet https://ja.padlet.com/は 8 種類のシートから 3 つを無料で利用できる。（2021 年 10 月 2 日閲覧）。

[14] 日本では学生のテキストマイニング技術コンテストが開催されている。NTT DATA.「VMStudio & TMStudio 学生研究奨励賞. NTT DATA　TEXTMINING STUDIO」https://www.msi.co.jp/tmstudio/literature.html. （2020 年 10 月 15 日閲覧）参照。

では分からない問題点が多数あり、外国語学部でプログラミング学習を導入して、方向性を模索していくことには開拓的意味があると考えられる。以下、プログラミング学習の導入によって、明らかになったプログラミング学習とプログラム操作の課題を検討していきたい。

4.1 学習環境の問題

台湾の理科系・技術系プログラミング教育や塾でのプログラム学習では、統一したPC環境（たとえば、中文版Windows1064bit版やIOS10のMacBookなど）で同じプログラミング言語やプラットフォーム（Python3、Anaconda3など）を利用し、基本的には英語と中国語の言語資料を使用して、情報処理や自然言語処理の基礎を教えていると考えられる。しかし、人文社会系の外国語学部では、もともと大学に十分な数のないPC教室は確保できない状態であり、また、そこで使われている環境も多言語処理を前提にしていないので、日本語の自然言語処理のためには環境設定をかなり変えて準備する必要がある。結局、本学科の場合は個人のNotebook PCを持参してもらって日本語の処理を交えたプログラミング授業を実施することになった。

4.1.1 PC環境の問題

その場合、個人のPC環境の新旧（現状ではWindows7,8,10の相違があり、また32bit版と64bit版の混在もある）と利用状況の違い（中文環境と日文環境）で、pythonなどのプログラムを動かすと、それぞれ挙動が異なるなどの大きな問題が生じることが多かった。Windowsも他言語に変更できるが、中文版Windows10で日本語表示に切り替える場合は、「設定を開き日本語ランゲージパックをインストールする」→再度、サインインして「設定で日本語環境の整備」→システムロケールを変更して「OSを日本語化する」の3つの段階があり、それぞれ関係する設定を実施しなくてはならない。[15]しか

[15] 言語表示の切り替えについて、マイクロソフトのホームページ等でも公式説明はないので、個々人の試行錯誤になるが、PC関係ライターの

し、システムロケールを変更して「OS を日本語化する」と、中国語アプリケーション等は表示が文字化けしてしまうものがあり、またメールなどが読めなくなる場合もあり、試行錯誤した結果、中国語環境のままで日本語処理ができるようにするほうが影響が少ないことが分かった（4.1.2 参照）。

パソコン環境が決まったところで、次はプログラム言語として Python3 を使う設定をするが、Python と Anaconda という二つの基本的なプラットフォームがあり、どちらを選ぶか決める必要が起きた。Python3 には各種のバージョンがあり、バージョンが違うと必要なライブラリーの動作に問題が起こる。また、最初に python を入れて、後から Anaconda3 を入れるとシステム全体の挙動が異常になったりすることがある。同時に、Python のライブラリープログラムは C、C++や Java などのライブラリーが入っていないと動かない場合があるが、何が setup されているかは学生ごとに違っていて、出てくるエラーメッセージも相違していることがあり、原因を見つけるのにケースバイケースで対応しなくてはならなかった。[16]

また、Python のオリジナルの API はそのままではプログラムには向かないので、Anaconda を利用し、PyCharm、Jupyter Notebook、Spyder などの連動したエディター環境で作業したほうが学習にも研究での応用にも幅広く自然言語処理のプログラムが実行できる。[17]特に、

MURA'sHomePage（2022）「外国語版 Windows 10（と Office）を日本語化する」http://www.vwnet.jp/windows/w10/2016092501/OtherLang2jaJP.htm（2022 年 2 月 16 日閲覧）等、個人の試行を参照する必要がある。

[16] Python と Anaconda の setup は、Python Japan（2022）「Python 環境構築ガイド」https://www.python.jp/install/docs/pypi_or_anaconda.html（2022 年 2 月 16 日閲覧）。Python のプログラムは、The Python Package Index（PyPI）を利用する方法か、科学技術計算のためのプラットフォームである Ananconda を利用する方法でプログラミングができるが、管理方法が異なっているため、両方を同時に使うのは難しい。また、解説図書の説明も、それぞれの環境が出ていて、同時に使われている PC 環境も異なっているので、戸惑うことが多い。Windows、Mac、Linux で操作やプログラムの書き方はかなり異なっている。

[17] Python はインタープリタ型の高水準汎用プログラミング言語で無料で公開されている https://www.python.org。Anaconda は科学計算のための Python および R 言語の無料のオープンソースディストリビューションで、各種のエディターと Python の仮想環境を作成できる https://www.anaconda.com/products/individual。プログラムコードを書いて、順番に動かしながら、結果を見られるエディターとして PyCharm https://www.jetbrains.com/pycharm/ , Jupyter Notebook https://jupyter.org/など

Python の練習をする場合、大量のライブラリーをインポートすると動作が遅くなり、また、プログラムに失敗があって暴走したり、動作停止したりしたときに環境が破壊される場合もあるため、仮想環境を切り替えて必要なライブラリーだけを使用できる Anaconda のシステムは利便性が高い。

プログラミングを学ぶ場合は、プログラミングプラットフォームと同時に、プログラムを作成し、実行する仮想環境を作成して動かす必要があり、またプログラムを書くエディターも必要であるが、Python と Anaconda の環境では、仮想環境の作り方とエディター使用の方法も異なっていて設定に迷う。何回か試行錯誤した結果、最終的に Anaconda を利用し、Anaconda Navigator から各種エディターでの操作ができる仮想環境を造り、その環境をエディターの PyCharm で読み込んでプログラムの学習を実施することにした。[18]

こうした問題を回避するには、仮想環境をサーバーに作成して、受講者に共通した仮想環境にアクセスしてもらい、そこでプログラムを動かしてもらうという方法がある。[19]しかし、自分でプログラム環境を造る練習はできないので、プログラミングを職業として選びたい場合は、自分で環境設定をおこなう必要があり、目的に合わ

がよく使われている（2022 年 2 月 16 日閲覧）。

[18] Anaconda のインストールは Python Japan（2022）「Windows 版 Anaconda のインストール」https://www.python.jp/install/anaconda/windows/install.html（2022 年 2 月 16 日参照）。その他、必要な環境設定の方法は、金子邦彦研究室（2022）「インストール，運用」https://www.kkaneko.jp/tools/index.html にプログラミングに必要な各種 OS やアプリケーションのインストールの方法が掲載されている。Anaconda 環境の詳細な設定は、金子邦彦研究室（2022）「Anaconda 3（Python 開発環境）のインストールと，その Python 3 仮想環境に，人工知能フレームワーク類のインストール（Windows 上）」https://www.kkaneko.jp/tools/win/windows_tensorflow.html（以上、2022 年 2 月 16 日参照）。方法は随時変わっていくので、以前の内容は使えない場合が多い。常にインターネット上で最新の各種資料で確認していく必要がある。

[19] 仮想環境で共通した環境を提供するシステムは VMware と呼ばれて、広く使用されるようになっている。110 学年度 1 学期の修士課程向けプログラミング授業「外語應用 AI 思維設計（一）」では、淡江大学工学部情報工学科陳建彰教授に、python を利用した深層学習のプログラミングを教授してもらったが、環境は Microsoft Azure の VMware https://azure.microsoft.com/ja-jp/services/azure-vmware/#product-overview を使い、進階人工智慧軟體 LEADERG AI ZOO https://tw.leaderg.com/article/index?sn=10982 を共通コードにして、プログラムを動かす練習をおこなった（以上 2022 年 2 月 16 日参照）。

せた環境設定を試行錯誤する必要がある。環境によって、同じプログラム言語でも挙動がまったく変わってしまうことは試してみないと分からない泥臭い試行錯誤の世界で、そうした部分があって、AI技術が開発されていることを知る体験にもなるであろう。

4.1.2 文字コードとデータ制作

理科系・技術系の環境では、そもそもそうした元々の Windows 等の系統と違う多言語を同時並行で扱っていないので問題は発生しないが、今回実施した英語と中国語環境の Windows10 64bit 版で日本語を扱うと、MS-Word や電子メールなどのアプリケーションでは全然不便を感じなくても、プログラムの場合は文字コードの違いで、処理できない場合やエラーが解決できない場合が起こり、また表示されるテキストが完全に文字化けしたりして、非常に対応が難しいケースが多々発生する。これらは、実際に個々の環境で動かしてみるまで分からないので、原因が分からず、その時間内で処理できずに持ち越しになってしまうこともある。

プログラムでは、データとしてテキストファイルや CSV ファイルを用いる。Ms-Word など多言語対応のアプリケーションに資料を貼り付けた場合、文字コードは自動で判別されて文字化けすることなく表示できるが、テキストファイルやまた、プログラミングで使用される Windows コマンドプロンプトやターミナルは単一の文字コードしかデータ表示できない。日本語には大きく euc、s-jis、utf-8 の三種類の文字コードがあり、中国語では台湾で使う場合は Big5 である。英語と中国語環境の Windows10 64bit 版で日本語を扱う場合、文字コード環境は相互に互換性がないので、テキストデータの文字コードは、常に注意が必要である。

資料 1　漢字圏の台湾、日本、中国、韓国とユニコードの基本文字コード（Microsoft コードページ CP）と各国での呼び方[20]

[20] Microsoft コードページは Microsoft（2022）「Code Page Identifiers」https://docs.microsoft.com/en-us/windows/win32/intl/code-page-identifiers（2022 年 2 月 22 日閲覧）参照。

コードページ	名称	表示名	表示名（日本語）
932	shift_jis	Japanese （Shift-JIS）	日本語 （シフト JIS）
936	gb2312	Chinese Simplified （GB2312）	簡体字中国語 （GB2312）
949	ks_c_5601-1987	Korean	韓国語
950	big5	Chinese Traditional （Big5）	繁体字中国語 （Big5）
1200	utf-16	Unicode	Unicode
1201	unicodeFFFE	Unicode （Big endian）	Unicode （Big-Endian）
1361	Johab	Korean （Johab）	韓国語 （Johab）
10001	x-mac-japanese	Japanese （Mac）	日本語 （Mac）
10002	x-mac-chinesetrad	Chinese Traditional （Mac）	繁体字中国語 （Mac）
10003	x-mac-korean	Korean （Mac）	韓国語 （Mac）
10008	x-mac-chinesesimp	Chinese Simplified （Mac）	簡体字中国語 （Mac）
12000	utf-32	Unicode （UTF-32）	Unicode （UTF-32）
12001	utf-32BE	Unicode （UTF-32 Big endian）	Unicode （UTF-32 ビッグエンディアン）
20000	x-Chinese-CNS	Chinese Traditional （CNS）	繁体字中国語 （CNS）
20001	x-cp20001	TCA Taiwan	台湾 TCA
20002	x-Chinese-Eten	Chinese Traditional （Eten）	台湾 Eten
20003	x-cp20003	IBM5550 Taiwan	台湾 IBM5550
20004	x-cp20004	TeleText Taiwan	台湾 TeleText
20005	x-cp20005	Wang Taiwan	台湾 Wang
20290	IBM290	IBM EBCDIC （Japanese katakana）	IBM EBCDIC （日本語 カタカナ）
20833	x-EBCDIC-KoreanExtended	IBM EBCDIC （Korean Extended）	IBM EBCDIC （韓国語 拡張）
20932	EUC-JP	Japanese （JIS 0208-1990 and 0212-1990）	日本語 （JIS 0208-1990 and 0212-1990）
20936	x-cp20936	Chinese Simplified （GB2312-80）	簡体字中国語 （GB2312-80）
20949	x-cp20949	Korean Wansung	韓国語 Wansung
50220	iso-2022-jp	Japanese （JIS）	日本語 （JIS）
50221	csISO2022JP	Japanese （JIS-Allow 1 byte Kana）	日本語 （JIS-Allow 1 byte Kana）
50222	iso-2022-jp	Japanese （JIS-Allow 1 byte Kana - SO/SI）	日本語 （JIS-Allow 1 byte Kana - SO/SI）
50225	iso-2022-kr	Korean （ISO）	韓国語 （ISO）
50227	x-cp50227	Chinese Simplified （ISO-2022）	簡体字中国語 （ISO-2022）
51932	euc-jp	Japanese （EUC）	日本語 （EUC）
51936	EUC-CN	Chinese Simplified （EUC）	簡体字中国語 （EUC）
51949	euc-kr	Korean （EUC）	韓国語 （EUC）
52936	hz-gb-2312	Chinese Simplified （HZ）	簡体字中国語 （HZ）
54936	GB18030	Chinese Simplified （GB18030）	簡体字中国語 （GB18030）
65000	utf-7	Unicode （UTF-7）	Unicode （UTF-7）
65001	utf-8	Unicode （UTF-8）	Unicode （UTF-8）

Python で自然言語処理をする場合は、必ず CP65001 Unicode （UTF-8）でデータを管理する必要がある。扱いに慣れない場合、まず Ms-Word に日本語の文書を貼り付けて、テキストで保存を選び、表示される文字コード選択画面で、CP65001 Unicode （UTF-8）を選んで保存すると、python での自然言語処理で使えるテキストファイ

ルができる。[21]

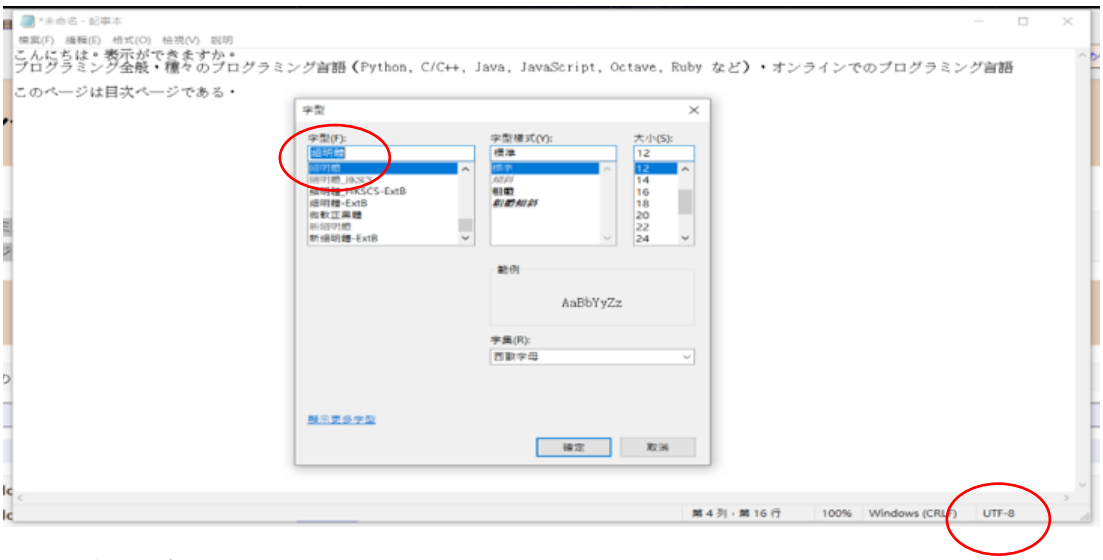
また、windows のコマンドプロンプト等で、操作する場合もあるので、windows のコマンドラインの操作方法も知っておく必要がある。コマンドラインの文字コードは chcp コマンドで表示でき、chcp=文字コードで、コードの切り替えができる。[22]

以前は文字コードの違う言語は自動では判別されず、中文版 Windows のコマンドプロンプトやメモ帳では、うまく表示できなかったが、2022 年アップデートの 21H2 の Windows10 では自動で文字コードを識別して Unicode （UTF-8）で表示ができるようになっており、コマンドプロンプトやメモ帳で多言語を貼り付けても表示が可能になった。

しかし、データの文字コードが Unicode （UTF-8）で保存されていることは Python での自然言語処理の基本なので、表示されている文字コードと、資料を保存した文字コードにはいつも注意が必要である。この更新以前のファイル内容やディレクトリ表示は、各国語の固有文字コード（日本語では S-jis、台湾では Big-5）の場合があるので、ファイル名やディレクトリ名が化けてしまう等のトラブルが起こりえる。今後、Windows11 が普及すると、文字コードに起因して新しい問題が起きる可能性がある。[23]

図 1　21H2 版中文版 Windows のメモ帳に日本語を入力し、フォントは細明體のまま自動で utf-8 に切り替わって表示されている例

[21] Microsoft（2022）「ファイルを開くときと保存するときに文字列のエンコード方法を選択する」https://bit.ly/36j28Pi（2022 年 2 月 16 日閲覧）参照。
[22] Chcp の使い方は、Windows コマンド虎の巻（2022）「chcp」https://windows.command-ref.com/cmd-chcp.html（2022 年 2 月 16 日閲覧）参照。
[23] 注意の一例として、Windows 環境ではファイル名などに使わないほうがいい文字が多数ある。



4.2 形態素解析の問題

台湾での一般的なプログラミング教育として中文版 Windows1064bit 版や IOS10 の MacBook などで同じプログラミング・アプリケーション（Python3、Anaconda3 など）を利用して、基本的には英語と中国語の言語資料を使用して自然言語処理の基礎を教えている場合、一方、日本で日本語処理を教える場合も日本語版 Windows1064bit 版や IOS10 の MacBook などで同じプログラミング・アプリケーション（Python3、Anaconda3 など）を利用して、基本的には英語と日本語の言語資料を使用して教える場合も、多言語が入らないので問題は生じにくいが、中文版 Windows1064bit 版の環境で日本語資料の処理をすると大きな壁になるのは、言語データを扱う形態素解析のプログラムの扱いである。

4.2.1 形態素解析システムの変化

表 1　日本語形態素解析システムの特性[24]

[24] 形態素解析の方法と特徴は、工藤拓（2018）『形態素解析の理論と実装』近代科学社 p.27 表 2.7 を参照。必要項目を論者が補足した。

形態素解析システム	JUMAN++[25]	ChaSen[26]（開発終了）	MeCab[27]	KyTea[28]	Sudachi[29]
動作システム／言語	Linux バイナリーから Windows にビルド	Linux/Windows	Unix/perl/ruby/Python（janome[30]）/Java（Kuromoji[31]/lucene-gosen[32]/Igo[33]）,Go（kagome）/Windows/Linux/MacOSX	Linux/Mac OSX/CygWin/Perl/Python/Ruby Node.js	Windows 10/Linux/Java/Python[34]
必要ライブラリ	gcc（4.9 以降）Boost C++ Libraries（1.57 以降）	Darts iconv	C/C++	データによる学習が必要	
開発用注釈付きコーパス	京都大学テキストコーパス	RWC コーパス	RWC コーパス BCCWJ コアデータ	BCCWJ コアデータ	RWC コーパス BCCWJ コアデータ
文字コード	Shift-JIS	Euc-JP/ Shift-Jis/ Utf-8/ ISO-8859-1	Euc-jp/ Shift-Jis/ Utf8	Utf8	Utf8
辞書	JUMAN 辞書	ipadic	Ipadic/NAIST-jdic/ UniDic/ipadic-NEologd	UniDic、	UniDic /ipadic-NEologd
品詞体系	増岡・田窪文法	学校文法 形容動詞なし	学校文法	学校文法	学校文法
単語長	長い	やや短い	やや短い、短い、とても長い	短い	短い、とても長い
語の斉一性	普通	普通	普通、高い、低い	高い	高い、低い
可能性に基づく品詞	少ない	普通	普通、多い	多い	普通、多い
台湾での応	▲Linux と	×使用例は多	◎開発終了、	◎開発が持続	◎現在も開発

[25] 京都大学黒橋・褚・村脇研究室 https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2Be（2022 年 2 月 16 日閲覧）。
[26] 奈良先端科学技術大学松本研究室 https://chasen-legacy.osdn.jp/（2022 年 2 月 16 日閲覧）。
[27] 京都大学情報学研究科−日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクト https://taku910.github.io/mecab（2022 年 2 月 16 日閲覧）。
[28] 京都テキスト解析ツールキットプロジェクト http://www.phontron.com/kytea/index-ja.html （2022 年 2 月 16 日閲覧）。
[29] ワークス徳島人工知能 NLP 研究所 https://github.com/WorksApplications/Sudachi#sudachi-%E6%97%A5%E6%9C%AC%E8%AA%9Ereadme （2022 年 2 月 16 日閲覧）。
[30] Janome https://mocobeta.github.io/janome/ （2022 年 2 月 16 日閲覧）。
[31] Atilica https://www.atilika.com/ja/kuromoji/ （2022 年 2 月 16 日閲覧）。
[32] Gosen https://github.com/lucene-gosen/lucene-gosen（2022 年 2 月 16 日閲覧）。
[33] Igo - Java 形態素解析器 http://igo.osdn.jp/ （2022 年 2 月 16 日閲覧）。
[34] Sudachi Python https://github.com/WorksApplications/SudachiPye（2022 年 2 月 16 日閲覧）。

用性	Windows の互換性のある環境	いが、開発終了のため発展性はない/辞書の切り替えはできない	中文版 Windows1064 bit 版では一番使いやすい/辞書の切り替えができる	すれば、利用する価値がある	中、ビジネス向けでは使用例が増えている

（注）　品詞体系、単語長、語の斉一性、可能性に基づく品詞は辞書の性質による違いである。工藤拓（2018）『形態素解析の理論と実装』近代科学社 p.37 を元に必要項目を補足して論者作成。

インターネットの日本語処理のプログラム例には、Python3 を直接、日本語版 Windows1064bit 版に setup して動かしている例が多数出ているが、日本語版の文字コード処理には euc、s-jis のものがあり、Python3 の文字コードは Unicode なので、Windows1064bit 版ではデータを変換しないと使えず、言語データを日本語の品詞相当の単位に区切る形態素解析の setup 時にも文字コードに注意しないと処理ができない。また、日本語の形態素解析システムには Linux 環境でしか動作しないもの、Java など Python とは別の言語でないと動かないものなども混在していて、動かすための準備に非常に手間がかかる。日本語での Python3 等で使える自然言語処理の形態素解析プログラムは以上の表１のようなものが使われているが、台湾で日本語の形態素解析を導入する場合、学習に使う中文版 Windows1064bit 版で動作確認をしなくてはならない。

また、現在、辞書を利用しない新しい方式の形態素解析システムも開発されている。[35]MeCab を使った処理が現在、日本では盛んであるが辞書の問題や品詞の問題が常に解決しきれず、処理の方法が大きく変わる可能性がある。また、参考図書やインターネットでの使用実例では作動しない場合や辞書の切り替えができないなど、使える環境を整えるのに時間が掛かる。

日本の Web サイトや図書では多くの日本語の自然言語処理のプログラムの実例があるが、書いている制作者のプログラミング環境が Linux、Mac、Windows と様々であり、それに応じて Python の式

[35] 言語データの学習による分割をおこなう方法で開発が行われている。一例として sentencepiece https://github.com/google/sentencepiece（2022 年 2 月 16 日閲覧）。

の書き方が異なり、使えるライブラリーも違っていて、形態素解析の部分だけでも環境が違うと相互に動かないため、実際に中文版 Windows1064bit 版で試行錯誤してみるほかない。現在のところは、台湾の実際の情報技術開発の場でもポピュラーな、中文版 Windows1064bit 版で Python3 を動かせる環境で、McCab を中心に各種の形態素解析のプログラムを動かす練習をしてみる必要がある。

4.2.2 形態素解析システムの辞書の選択

また、自然言語処理では表 1 のように形態素解析システムと同時に使用される辞書が様々あり、自然言語処理のプログラムの実例を出している制作者は、それぞれのシステムと辞書が優れていると報告しているが、日本語での形態素解析の課題は文章ジャンルと時代相および目的に応じてまったく違ってくるため、一般的正解は存在しない。データを扱う目的に応じて形態素解析システムを選び、また辞書を選ぶ必要がある。特に、人文社会系の言語資料を解析する場合は、現代語、近代語、古典語の相違、また、書き言葉と話し言葉の相違、文章ジャンルの相違など、理科系・技術系で普通はまったく注目していない言語の通時的共時的特徴が大きな問題になるので、こうした多様な分析例が出てくることは理科系・技術系では、ほとんどない。人文社会系で自然言語処理を行う場合、形態素解析で自分の目的に応じたシステム選択と辞書の管理や切り替えをする方法を学ぶ必要がある。

MeCab を例に、辞書の種類を説明すると、開発過程で制作されてきた大きく ipadic 系統と UniDic 系統がある。ipadic 系統には、MeCab に付属しているものと、利用者が新語を補足して制作した ipadic-NEologd がある。UniDic 系統には、MeCab サイトや Python のライブラリーMeCab Python に付属しているものと、国立国語研究所で開発した各種辞書がある。[36]MeCab は、ユーザー辞書制作が可能で、

[36] MeCab のサイト MeCab: Yet Another Part-of-Speech and Morphological Analyzer https://taku910.github.io/mecab/では、ipdadic、jumandic が公開されている。Unidic は、https://ja.osdn.net/projects/unidic/downloads/58338/unidic-mecab-2.1.2_src.zip/にある。国立国語研究所 UniDichttps://ccd.ninjal.ac.jp/unidic/では、

現在、多様な新語に対応するため各使用者が自分のデータ用にユーザー辞書を作成している。[37]実施した授業では、まだ試行していないので、こうした試行も授業に加える必要がある。

また、出力された形態素解析結果は、基本となる文法による品詞分類はあっても、日本語学や日本語教育で想定している単語の区切れや該当品詞とは大きく違っている場合が多々ある。自然言語処理での形態素は、形態素への分割、品詞の推定、語形変化の処理を経て分析された結果であり、たとえば「外国人参政権」であるが、「外国人参政権」と一形態素とするか、「外国」「人参」「政権」のように二語を基本とした区切れで分けるか、「外国人」「参政権」のように三語基本とした区切れにするか、あるいは、「外」「国」「人」「参政」「権」のようにできるだけ細かく分けるかなど、形態素解析システムと辞書の組み合わせで期待される区切れは大きく変わってしまう。

4.3 Anaconda を使った Python 環境での MeCab 運用

中文版 Windows1064bit 版で Anaconda を使った Python 環境を作り、日本語自然言語処理に関わるプログラミングを練習する場合、MeCab が元々32bit であるため扱いが難しい場合が多い。日本語のサイトや本での説明のままの例示プログラムでは MeCab が認識できない、辞書の切り替えができないなどのトラブルが起こる。また Windows に MeCab を直接 setup しても、Python 自体から認識できない場合が起こる。また、PyPI を使って、Python のライブラリーからプログラム内に MeCab をインポートすると、PyPI の中に MeCab を利用したライブラリーが複数あって、どのバージョンがインポートされたのかはっきりせず、辞書が認識できない、切り替えができない等のトラブルが発生しやすい。

繰り返し試行錯誤した結果、以下の手順で比較的安定した実行環

現代書き言葉、現代話し言葉、古文用（旧仮名口語、近代文語、近世口語（洒落本）、中世口語（狂言）、中世文語（説話・随筆）、中古和文、上代（万葉集））の各 UniDic が公開されている（以上、2022 年 2 月 16 日参照）。

[37] MeCab のユーザー辞書作成の方法は、http://taku910.github.io/mecab/dic.html 参照。しかし、文字コードの問題もあって各環境で試行するしかない。

境を作成できた。

（1）MeCab の setup：公式ではないが MeCab64bit 版[38]をまず中文版 Windows に setup する。その場合、setup 時の文字表示は文字化けするが、文字コード選択で utf-8 を選択して setup すれば問題はない。

（2）Anaconda 環境との連携：作成しておいた Anaconda の Python3 のディレクトリを確認し、必要な DLL を Anaconda の Python3 の仮想環境から認識できるディレクトリにコピーする。[39]Anaconda Navigator を使って作った仮想環境が複数あると、それぞれの名前のディレクトリができるので、そこにも同じ操作が必要である。

（3）Anaconda の Python 環境への MeCab ライブラリーのインストール：2022 年 2 月までの Python 内では、mecab 0.996.3 のライブラリーを指定してインポートするのが一番、安定した環境になるようである。[40]

（4）使用辞書の切り換え：仮想環境内で MeCab をインポートしてプログラムが作動するのを確認したら、使用辞書の切り替えを試して見る。辞書の切り替えは、一般的な python の解説や日本語のサイトや本での説明には必要な操作が出ていないので、工夫が必要である。辞書の切り替えは、以下の操作が必要である。

（a）MeCab に標準で付属している ipadic 以外の使いたい辞書（ipadic-Neologd や UniDic または国立国語研究所 UniDic（現代書き

---

[38] MeCab64 ビット版はいくつか種類があるが、MeCab 0.996 64bit version（2019）https://github.com/ikegami-yukino/mecab/releases（2022 年 2 月 16 日閲覧）ファイル名は、「mecab-64-0.996.2.exe」を使用する。

[39] Python がインストールされているフォルダに MeCab_64bit_Installer.exe でインストールした libmecab.lib と libmecab.dll をコピー＆ペーストする。libmecab.lib が入っているのは「MeCab をインストールした場所¥MeCab¥sdk」で、libmecab.dll があるのは、「MeCab をインストールした場所¥MeCab¥bin」である。この二つをペーストするべき Python がインストールされているフォルダは「Python をインストールした場所¥Lib¥site-packages」となっている。Anaconda 仮想環境なので、通常は C:¥Users¥ユーザー名¥anaconda3¥Lib¥site-packages となっているはずである。ここに、libmecab.lib と libmecab.dll をコピー＆ペーストする。

[40] Anaconda Prompt で使用している仮想環境を起動し、2021 年のバージョンでコマンド[pip install mecab] https://pypi.org/project/mecab/によって、mecab 0.996.3 のライブラリーをインストールできる。他にも、Pypi のライブラリーに MeCab のライブラリーが多数あるが、Anaconda の仮想環境から認識できない、pip での install に失敗する、辞書の切り替えができないなどの問題が起きやすいので、どれが作動するか試行錯誤で確認していく必要がある。

言葉、現代話し言葉、古文用（旧仮名口語、近代文語、近世口語（洒落本)、中世口語（狂言)、中世文語（説話・随筆)、中古和文、上代（万葉集))をサイトからダウンロードして、setup または解凍する。

（b)「MeCab をインストールした場所¥MeCab¥etc」にある「mecabrc」をエディターで編集して、使いたい辞書を MeCab の標準（デフォルト）辞書とする。mecabrc に、「dicdir＝$（rcpath）¥..¥dic¥ipadic」という箇所があるので、「=」の後を、使いたい辞書がある場所へと変更して、上書き保存する。

（c) MeCab は出力のフォーマットを自由に変更できるので、形態素解析結果の品詞分類等を変えたい場合は、各辞書のフォルダーにある dicrc の記述を変更する。たとえば、UniDic を使って ChaSen と同じ形式で出力する設定では、UniDic をインストールしたフォルダの「dicrc」をテキストエディタで開いて、末尾に以下の 3 行を追加する。[41]

```
node-format-chasen = %m¥t%f[6]¥t%f[7]¥t%F-[0,1,2,3]¥t%f[4]¥t%f[5]¥n
unk-format-chasen = %m¥t%m¥t%m¥tUNKNOWN¥t¥t¥n
eos-format-chasen = EOS¥n
```

Python 内のライブラリーでは辞書の切り替えは、うまくいかないため、必要な外部環境を、あらかじめ PC 側の環境で作っておく必要がある。

4.4 プログラミングの実行

自然言語処理のプログラミング練習は、基本的に扱うデータの種類が異なるため、いわゆる Python の練習やそれを応用した機械学習、深層学習とは異なる部分がある。理科系・技術系で練習しているデータは、各種の数値データや統計データ、数値化できる画像データや音声データで、いわゆるビッグ・データと呼ばれている分野での処理で使われるようなデータである。一方、人文社会系での主なデータは、書き言葉や話し言葉をテキスト化した言語データで、

[41] MeCab の出力形式変更のコマンドは、MeCab「出力フォーマット」https://taku910.github.io/mecab/format.html（2022 年 2 月 16 日閲覧）参照。

形態素解析が成功して初めて、数値的処理ができる（解析された品詞を分類する等）ようになるデータである。言語データをテキストから読み込んで、形態素解析する部分が動かないと、プログラムを動かすことができない。今まで、述べてきた内容は、プログラムを動かすための前提を作る準備作業で、実際に手本となるプログラムを教室で学生達が動かすまでに、こうした準備をしておく必要がある。多くの本やネットの資料で為されているような Python の練習やそれを応用した機械学習、深層学習等とはすべき準備がかなり異なっており、また、そこに多言語環境という課題があることを知っておく必要がある。

プログラム学習と言っても、どのようなデータを扱って、どのような結果を得たいのかという目的設定がないと、人文社会系でのプログラミング学習は、理系・技術系に追随するだけになり、独自性を失ってしまう危険がある。しかし、自然言語処理をする準備が整えば、後は本やネットに出ている例題を動かして、自然言語処理の基本操作を学ぶことは容易である。自然言語処理の成果として、現在、すでに知られているように翻訳、要点抽出、文書作成、対話、音声の文字化、情報検索など様々な分野で実用的プログラムが作動しており、ビジネス化が進んでいる。そうした実用化の方向と接続できる学習デザインを人文社会系では目指していくことが求められると言えよう。[42]

実践した授業では、以下のような、自然言語処理に関する資料を手本にして、プログラムの操作を学んだ。

山内長承（2017）『Python によるテキストマイニング入門』オーム社：自然言語処理の基礎となる資料の読み込み、形態素解析後の統計的処理を具体的に操作して、学ぶことができる。

[42] ビジネスの自然言語処理動向の紹介として、AISmaily「自然言語処理とは！？できることをまとめた NLP 入門書」https://aismiley.co.jp/ai_news/what-is-natural-language-processing/等を参照（以上、2022 年 2 月 16 日参照）。

中山光樹（2020)『機械学習・深層学習による自然言語処理入門
ーscikit-learn と TensorFlow を使った実践プログラミング』
マイナビ：2010 年以降の第三世代 AI の機械学習と深層学習のライブラリーを Python で動かして、体験できる。

これらの図書は自然言語処理の基本的な手続きを紹介しており、実際に動かせるプログラムもサイトからダウンロードできるので、手本のプログラムを動かして、作動を確認し、また各部分で必要な処理を確認できる。理系・技術系のプログラミング教育では、新しい処理を開発するために元になっている式の数学的理解が必要だが、人文社会系の場合は、まずプログラミングに慣れることが重要と言え、必要な処理について実例を見ながら処理を組み立てたり、必要部分を書き換えていくなどで、操作を学ぶことが最初の出発点として重要と言えよう。

プログラミング言語そのものの作動を知るには、Python など言語そのものを学ぶ課程を別に受けておく必要があるが、これは、今回述べたような自然言語処理環境なしでも学べるので、既存のコースを履修することで、基礎を学ぶことができる。多言語環境での自然言語処理の課程は、既存のコースとは別に人文社会系の学科の特色が活かせるように内容をデザインしていく必要がある。

4.5 テキストマイニングの導入

現在、ビジネス用の BI ツールが販売され、データマイニング、テキストマイニング、感情評価分析などが一体でできる環境が生まれているが、有料なのでどこまで授業や研究で使うか判断は難しい。無料で使えるツールでも目的に応じて学習や研究で有益に使えるので、ツールの導入は使用目的の見極めが重要であろう。無料で使える KH Coder は、中文版 Windows1064bit 版に setup し、中国語、英語をはじめ多言語での分析が可能であり、人文社会系の学科では導入して、テキストマイニングの実習をおこなうのに適している。

特に、日本語の資料の場合、MeCab の辞書を切り替えて多様なジャンルの資料でも簡単に使用できる、一単位で取る語や排除する語

を指定でき形態素の区切りを自由に入れられる、外部変数として資料にコーディング項目をつけて解析できるなど、人文社会系のテキストマイニングでは極めて有益で、使いやすいツールである。[43]

108、109 学年度、1 学期を使って、量的分析と質的分析の違いを理解しながら、KH Coder を使って、テキストファイルなどの資料制作と文字コードの処理、コーディングできる Excel を使った資料作成、スクレイピング等を使用したインターネットでの資料蒐集、各種辞書の切り換えをおこなっての MeCab による形態素解析、分析結果の各種出力（対応分析、多次元尺度構成法、階層的クラスター分析、共起ネットワーク、自己組織化マップ）と解読、ネット資料を使った分析と結果解読実習を実施し、第 2 節でも触れたように、日本でのテキストマイニングコンククールである NTT「学生研究奨励賞」[44]にも出せるレベルで研究報告をまとめることができた。

テキストマイニングは、現在、実際にビジネス現場で使用されている応用スキルとして、今後さらに一般化していくと考えられるので、プログラミングとは別に練習の機会を設け、また、活用方法について認識する必要がある。特に、第二世代 AI 時代の言語資料を一般化する量的統計的処理に止まらず、ジャンルや分野別に非常に異なる言語資料の特徴に即した要点抽出、特徴抽出には内容の質的理解が必要で、質的分析を踏まえた量的分析の活用ができるように、内容をデザインしていくことが求められる。今までなされてきた第二世代 AI 時代のテキストマイニングから一歩進んだ次元に入りつつあることを踏まえて、何が可能で、何が不可能か技術の特性を理解しながら活用する方向性を目指していきたい。

4.6 今後の方向性の探究

以上、勤務先で 2 年間、人文社会系分野へのプログラミング教育の導入について試行錯誤してきた結果、明らかになった多言語環境

[43] 樋口耕一（2020)『社会調査のための計量テキスト分析―内容分析の継承と発展を目指して第 2 版』ナカニシヤ出版参照。
[44] 同注 14 参照。

での自然言語処理のトラブルと対応を中心に述べてきた。大事な点は、以下のように整理できよう。

（1）多言語環境でのプログラミング教育の準備：従来からの理系・技術系での単言語環境でのプログラミング教育はアルゴリズムやプログラミング言語の理解の基礎を作ることはできる。しかし、多言語環境でのプログラミング教育では文字コードの混在や形態素解析の環境設定など、特別な問題が起こるため、環境に応じた特別な準備が必要になる。

（2）人文社会系の特色を活かす方向性：従来からの理系・技術系での単言語環境でのプログラミング教育では、量的処理だけで扱えるデータを中心に扱っているが､それは人文社会系では応用できない。多言語という人文社会系の特色を活かせる自然言語処理など､理系･技術系とは異なった新しい方向性を見つけて、理系・技術系での単言語環境でのプログラミング教育とは異なった教育目標と内容を形成していく必要がある。

（3）ビジネス的実用的スキルとの関係：自然言語処理から生まれた第三世代 AI の技術により、翻訳、対話、資料制作、情報収集などで人間の言語を処理する新しい AI プログラムが現在、ビジネスや生活の各分野に応用されるようになっており、その傾向は今後も拡大すると考えられる。そのトレンドの中、人文社会系の特色である多言語能力を自然言語処理に結びつけることで、新しいビジネス的実用的スキルを開拓していくことは、今後の人文社会系研究と教育の存続のために不可欠になっている。人文社会系でのプログラミング教育を応用可能性の探究の中で、学習の系統性を持って形成していく必要がある。

5．おわりに

現在、使われている AI 技術を概観し、カリキュラム開設、アプリケーション導入、プログラミング教育創設の各レベルで台湾の日本語教育に情報処理技術教育を活かす方法を考えてきた。台湾の高等

教育における日本語教育が AI 時代に対応するには、基本的に三つの方向性が必要と言える。第一は多言語環境でのプログラミング教育の準備、第二は人文社会系の特色を活かす方向性の模索、第三はビジネス的実用的スキルとの関係である。理系・技術系の情報処理技術教育を活かしながら、各種ある情報技術を日本語関係学科の日本語能力の今までの訓練カリキュラムのどこに位置づけるかで、日本語教育また人文社会系学科でのカリキュラム開設、アプリケーション導入、プログラミング教育創設の位置づけと内容は大きく変わってくる。未だ、試みは各現場で始まったばかりであり、試行錯誤を繰り返しながら、新しい時代の方向性を見いだしていきたい。

参考文献・資料

韻律読み上げチュータスズキクン http://www.gavo.t.u-tokyo.ac.jp/ojad/phrasing（2021 年 10 月 2 日閲覧）

王嘉臨（2021）「Amazon Comprehend 感情分析を用いた読解指導―詩教材を中心に」『淡江日本論叢』43pp.49-64

落合由治（2020）「日本語関係人文系研究の質的研究におけるテキストマイニング手法の応用と課題」『台大日本語文研究』39pp.101-130

落合由治、曾秋桂、王嘉臨、葉麥（2020）「人文系教育への情報処理・自然言語処理技術の導入と応用」『淡江日本論叢』42pp.45-63

金子邦彦研究室（2022）「Anaconda 3（Python 開発環境）のインストールと，その Python 3 仮想環境に，人工知能フレームワーク類のインストール（Windows 上）」https://www.kkaneko.jp/tools/win/windows_tensorflow.html（2022 年 2 月 16 日参照）

金子邦彦研究室（2022）「インストール，運用」https://www.kkaneko.jp/tools/index.html（2022 年 2 月 16 日参照）

京都大学黒橋・褚・村脇研究室 https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2Be（2022 年 2 月 16 日閲覧）

京都大学情報学研究科–日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクト https://taku910.github.io/mecab （2022 年 2 月 16 日閲覧）

京都テキスト解析ツールキットプロジェクト
http://www.phontron.com/kytea/index-ja.html （2022 年 2 月 16 日閲覧）
工藤拓（2018）『形態素解析の理論と実装』近代科学社 p.27
国立国語研究所 UniDic https://ccd.ninjal.ac.jp/unidic/（2022 年 2 月 16 日参照）
曾秋桂（2021）「AI の学習効果についての一考察－「AI と外国語学習」講座授業を中心に」『淡江日本論叢』43pp.114-133
曾秋桂（2021）「グローバル時代のエコフェミニズムの視点から読む多和田葉子の『星に仄めかされて』―「神の子どもたちはみんな踊る」の意味をめぐって」『台大日本語文研究』41pp.21-40
曾秋桂（2021）「日本語教育のつながりとひろがり―AI と HI を兼ね備えた外国語（日本語）人材 2.0 の育成を目指して」『日本語教育研究』54,pp23-36
総務省（2016）『平成 28 年版情報通信白書』
http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/（2021 年 10 月 2 日閲覧）。
淡江大学日本語文学科（2021）『2021 年 AI と日本語教育国際シンポジウム―クリエイティブラーニングを目指す AI と日本語教育』淡江大学日本語文学科
奈良先端科学技術大学松本研究室 https://chasen-legacy.osdn.jp/（2022 年 2 月 16 日閲覧）
樋口耕一（2020）『社会調査のための計量テキスト分析―内容分析の継承と発展を目指して第 2 版』ナカニシヤ出版
ビジネス向け案内
https://gcp.nict.go.jp/news/products_and_services_GCP.pdf （ 2021 年 10 月 2 日閲覧）
葉麥（2020）『AI 技術による村上春樹文学の再読―短編小説を起点として』瑞蘭國際有限公司
ワークス徳島人工知能 NLP 研究所
https://github.com/WorksApplications/Sudachi#sudachi-%E6%97%A5%E6%9C%AC%E8%AA%9Ereadme （2022 年 2 月 16 日閲覧）
AISmaily「自然言語処理とは！？できることをまとめた NLP 入門書」
https://aismiley.co.jp/ai_news/what-is-natural-language-processing/（2022 年 2 月 16 日閲覧）
Anaconda https://www.anaconda.com/products/individual（2022 年 2 月 16 日閲覧）

Atilica https://www.atilika.com/ja/kuromoji/（2022 年 2 月 16 日閲覧）
DeepL https://www.deepl.com/ja/translator（2021 年 10 月 2 日閲覧）
Google 翻訳 https://translate.google.com/?hl=ja（2021 年 10 月 2 日閲覧）
Gosen https://github.com/lucene-gosen/lucene-gosen （2022 年 2 月 16 日閲覧）
Hallo Talk https://www.hellotalk.com/?lang=ja（2021 年 10 月 2 日閲覧）
Igo - Java 形態素解析器 http://igo.osdn.jp/ （2022 年 2 月 16 日閲覧）
Janome https://mocobeta.github.io/janome/ （2022 年 2 月 16 日閲覧）
Jupyter Notebook https://jupyter.org/（2022 年 2 月 16 日閲覧）
LEADERG AI ZOO https://tw.leaderg.com/article/index?sn=10982（2022 年 2 月 16 日閲覧）
MeCab: Yet Another Part-of-Speech and Morphological Analyzer https://taku910.github.io/mecab/（2022 年 2 月 16 日参照）
MeCab「出力フォーマット」 https://taku910.github.io/mecab/format.html（2022 年 2 月 16 日閲覧）
MeCab のユーザー辞書作成の方法 http://taku910.github.io/mecab/dic.html（2022 年 2 月 16 日閲覧）
MeCab 0.996 64bit version （ 2019 ） https://github.com/ikegami-yukino/mecab/releases（2022 年 2 月 16 日閲覧）
Microsoft（2022）「Code Page Identifiers」https://docs.microsoft.com/en-us/windows/win32/intl/code-page-identifiers（2022 年 2 月 22 日閲覧）
Microsoft（2022）「ファイルを開くときと保存するときに文字列のエンコード方法を選択する」https://bit.ly/36j28Pi（2022 年 2 月 16 日閲覧）
Microsoft Azure の VMware https://azure.microsoft.com/ja-jp/services/azure-vmware/#product-overview（2022 年 2 月 16 日閲覧）
MURA’s HomePage（2022）「外国語版 Windows 10（と Office）を日本語化する」 http://www.vwnet.jp/windows/w10/2016092501/OtherLang2jaJP.htm（2022 年 2 月 16 日閲覧）
Naver Papago https://papago.naver.com/（2021 年 10 月 2 日閲覧）
NTT DATA.「VMStudio & TMStudio 学生研究奨励賞. NTT DATA TEXTMINING STUDIO」

https://www.msi.co.jp/tmstudio/literature.html. （2020 年 10 月 15 日閲覧）
OJAD http://www.gavo.t.u-tokyo.ac.jp/ojad/（2021 年 10 月 2 日閲覧）
padlet https://ja.padlet.com/（2021 年 10 月 2 日閲覧）
PyCharm https://www.jetbrains.com/pycharm/（2022 年 2 月 16 日閲覧）
Python https://www.python.org（2022 年 2 月 16 日閲覧）
Python Japan（2022）「Python 環境構築ガイド」https://www.python.jp/install/docs/pypi_or_anaconda.html（2022 年 2 月 16 日閲覧）
Python Japan（2022）「Windows 版 Anaconda のインストール」https://www.python.jp/install/anaconda/windows/install.html（2022 年 2 月 16 日参照）
sentencepiece https://github.com/google/sentencepiece（2022 年 2 月 16 日閲覧）
Sudachi Python https://github.com/WorksApplications/SudachiPye（2022 年 2 月 16 日閲覧）
TechCrunch（2021）「AI チャットボット「りんな」の rinna と UneeQ を日本展開するデジタルヒューマンが協業、顔・声・視聴覚を持つ雑談 AI 実現」https://jp.techcrunch.com/2021/05/31/rinna-uneeq-digitalhumans/（2021 年 10 月 2 日閲覧）
TechCrunch（2021）「Zoom 商談を書き起こし Salesforce に自動入力するオンライン商談自動化ツール「アンプトーク」が発売開始」https://jp.techcrunch.com/2021/09/06/amptalk-fundraising/（2021 年 10 月 2 日閲覧）
Unidic https://ja.osdn.net/projects/unidic/downloads/58338/unidic-mecab-2.1.2_src.zip/（2022 年 2 月 16 日参照）
UserLocal https://textmining.userlocal.jp/（2021 年 10 月 2 日閲覧）
VoiceTra https://voicetra.nict.go.jp/（2021 年 10 月 2 日閲覧）
Windows コマンド虎の巻（2022）「chcp」https://windows.command-ref.com/cmd-chcp.html（2022 年 2 月 16 日閲覧）

註記

本論文は、2021 年 11 月「2021 年度台湾日本語教育研究国際学術シンポジウム」での研究発表を修正、加筆したもので、科技部専題研究計画 109-2410-H-032 -061 -MY3 の研究成果の一部である。査読のご意見、研究へのご支援に感謝申しあげる。