様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ２９ 日現在

機関番号：１１１０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ００５４１

研究課題名（和文）Developing a program for language teaching with parsed corpora

研究課題名（英文）Developing a program for language teaching with parsed corpora

研究代表者
バトラー　アラステア（Butler, Alastair）

弘前大学・人文社会科学部・准教授

研究者番号：９０５８８８７３

交付決定額（研究期間全体）：（直接経費）　　1,300,000円

研究成果の概要（和文）：本研究では、外国語教育に役立てるために、ほぼ並行して開発を行った2つの統語解析情報付きコーパス、NPCMJ および TSPC の応用について研究を行った。扱う言語は異なるが、これら2つのコーパスは、パラレル・データを含み、言語学的解析の原理や同一のウェブ・インタフェースの利用という点で共通している。
これらのコーパスを利用した英語教育プログラムの開発を行った。具体的には、1）日本人大学生向けの英語学習用教科書の開発、2）動詞の項-述語パターン表示コードの付加等、文法分析情報の補強、および 3）学生が自分で文法分析を行えるようにするためのツールキットの開発、を行った。

研究成果の学術的意義や社会的意義
本研究の独創的な点は、外国語教育に直接文統語解析を利用することにある。このことは、類似のアノテーションを施したコーパスが日本語、英語双方に存在することにより可能となる。学習者は、実際に学習項目の文法事象の統語解析を、解析結果の視覚化システムの補助を受けて進めることによって理解するようになる。
本研究により、実際のテクストデータに対する言語分析ツールキットの適用例を示すことができた。テクストの解析結果を見ることは、文法理解の過程をリアルに知ることに他ならない。これにより、学生は、豊かな文法解析情報を持つコーパスの参照を通じて、自分自身の言語使用を客観視しつつ向上させることが可能になる。

研究成果の概要（英文）：This project aimed to utilise for the purposes of language teaching two parsed corpora developed largely in tandem: the NINJAL Parsed Corpus of Modern Japanese (NPCMJ) and the Treebank Semantics Parsed Corpus (TSPC). While these corpora are for very different languages, Japanese and English respectively, there is considerable overlap, both in terms of parallel data, as well as principles of linguistic analysis, and the accessibility of the analysis through a shared web-interface with search functionality.
The research plan involved developing a program for language teaching using these parsed corpora. Components developed were: 1) a grammar textbook focused on English language learning for Japanese students at university level, 2) further enhancements to the grammatical analysis of the corpora, notably adding codes to mark the complementation patterns of verbs, and 3) the development of a "toolkit" for analysis creation, for students to start analysing their own written language.

研究分野：linguistics

キーワード：grammatical analysis　parsed corpora　semantic dependencies　language teaching　English　Japanese

１．研究開始当初の背景
Background

There is now widespread use of corpora in language education for stages of textbook creation, where they provide the grounds for establishing what is real language use. Corpora of second language learners have been created as a basis for identifying common mistakes. For education in the classroom, there is a tradition of using concordance software for students to gain access to the language data of corpora (see e.g., Reppen 2010).

Parsed corpora add linguistic mark-up and constitute extremely careful considerations of the data. Parsed corpora have been created for many languages, and some include wide historical timespans. When data is given very detailed analysis, there can be automated extractions of insight from the pre-analysed data, so students can gain experience from consequences of analysis that will be exposing explanations for productive language use.

This project aimed to utilise for the purposes of language teaching two parsed corpora developed largely in tandem: the NINJAL Parsed Corpus of Modern Japanese (NPCMJ; Yoshimoto et al. 2022) and the Treebank Semantics Parsed Corpus (TSPC; Butler 2023). While these corpora are for very different languages, Japanese and English respectively, there is considerable overlap, both in terms of parallel data, principles of linguistic analysis, and the accessibility of the analysis through a shared web-interface with search functionality.

２．研究の目的
Purpose

The NPCMJ and TSPC parsed corpora, as well as the parsing systems that underlie their creation, have great potential for use in scenarios of language teaching. In particular, the parsing systems amount to the realisation of a “toolkit” for analysing sentences of Japanese and English.

The innovation of the research is the ambition to use parsed analysis directly for language teaching. This is now possible because there are Japanese and English corpora to search with related analysis. There are also automatic tools to complete much of the tedious aspects of analysis creation.

The way offered for language learners to be able to make the most sense of the parsed analysis is to first become involved with the actual creation of parsed analysis, followed by supporting visualisations to assist with making the analysis approachable and relatable to points of grammar students are learning.

The research has contributed demonstrations of how to apply the analysis toolkit to real textual materials, including students own work. Seeing text analysed is to see the embodiment of grammatical understanding. This promises to equip students with skills to criticise and improve their own language use through experiencing implications drawn from rich grammatical analysis.

３．研究の方法
Method

The research plan involved developing a program for language teaching using the NPCMJ and the TSPC parsed corpora. Components developed were:

1) a grammar textbook focused on English language learning for Japanese students at university level,
2) further enhancements to the grammatical analysis of the corpora, notably adding codes to mark the complementation patterns of verbs, and
3) the development of a "toolkit" for analysis creation, for students to start analysing their

own written language.

Textbook development was branched into: (i) an introductory guide for the English parsed corpus, and (ii) a supplement to a published textbook linking to corpus queries. Many “to do” tasks were created to encourage active learning. Results were released on the web.

While the size of the TSPC for English parsed data didn't change (43,850 trees; 468,868 words), many improvements were made to the analysis. Most notably, there was adding of verb codes to mark verb complementation information together with markings to signal when adverbial items (adverbial particles and preposition phrases) are verb selected. Also, more marking was included to resolve anaphoric connections, so as to capture records of discourse coherence.

The largest amount of work went into improving the "toolkit" for producing syntactic parse and semantic dependency analysis for English sentence input. The parsing system is accessible on the web.

The developed parsing system provides very wide input coverage while also enforcing grammatical rules. This has involved developing techniques for wide coverage parsing analysis of English to build up parse information from rich word information. The rich word information is made with markers of word class and grammatical codes. Grammatical codes given to verbs double as partial indicators of word sense, offering key insights into word sense that are independently of value for English language learners to know.

４．研究成果
Results

A notable achievement is that parsed results present full syntactic parse content, in the sense of offering constituent structure with: (i) rich tag labels that give form and function information, (ii) zero elements with indexing to mark the place of interpretation for unbounded dependencies, and (iii) traces for relative clauses. This rich parse information supports resolving semantic dependencies beyond internal clause relations, such as control relations. Consequences of analysis are made clear from visualisations of the derived semantic dependencies.

The rich word information required to support the parsing can be supplied by students who are learning about grammatical analysis. This takes away the mundane crunching tasks of reaching parse analysis while leaving the task of providing the essential information (word class and grammatical codes) to determine the directions a parse takes. These are really hard decisions of parsing that computers are still not very good at making, but this is exactly the information that humans excel at giving and are representative of in-depth insight into language competency, so skills language learners need to master. Applying the system in the classroom has demonstrated the great personal satisfaction learners can gain from being responsible for successful language analysis.

References

Butler, Alastair. 2023. The Treebank Semantics Parsed corpus (TSPC) Web Site. Hirosaki University. Available at: entrees.github.io.
Reppen, Randi. 2010. Using Corpora in the Language Classroom. Cambridge University Press, Cambridge.
Yoshimoto, Kei, Prashant Pardeshi, Iku Nagasaki and Alastair Butler. 2022. NINJAL Parsed Corpus of Modern Japanese **の構築と公開** (Development and publication of the NINJAL Parsed Corpus of Modern Japanese). **自然言語処理** 29(3), pp. 1015-1022.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　4件／うち国際共著　4件／うちオープンアクセス　4件）

１．著者名 Alastair Butler	４．巻 12758
２．論文標題 Knowledge Acquisition from Natural Language with Treebank Semantics and FLORA-2	５．発行年 2021年
３．雑誌名 Lecture Notes in Computer Science, New Frontiers in Artificial Intelligence. JSAI-isAI 2020	６．最初と最後の頁 37-49
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-79942-7_3	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Alastair Butler	４．巻 12331
２．論文標題 From Discourse to Logic with Stanford CoreNLP and Treebank Semantics	５．発行年 2020年
３．雑誌名 New Frontiers in Artificial Intelligence. JSAI-isAI 2019	６．最初と最後の頁 182-196
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-58790-1_12	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Koichi Takeuchi, Alastair Butler, Iku Nagasaki, Takuya Okamura and Prashant Pardeshi	４．巻 -
２．論文標題 Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus	５．発行年 2020年
３．雑誌名 Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)	６．最初と最後の頁 3153-3161
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 竹内孔一，バトラー アラステア，長崎郁，ホーンスティーブンライト	４．巻 -
２．論文標題 PropBank形式を考慮したNPCMJに対する意味役割付与~態の違いと経験者の付与~	５．発行年 2020年
３．雑誌名 言語処理学会 第26回年次大会	６．最初と最後の頁 633-636
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 竹内 孔一, Alastair Butler, 長崎 郁, Prashant Pardeshi	4．巻 2019-NL-241(4)
2．論文標題 NPCMJに対する述語項構造シソーラスの意味役割と概念フレームの付与	5．発行年 2019年
3．雑誌名 SIG Technical Reports	6．最初と最後の頁 2188--8779
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Alastair Butler	4．巻 -
2．論文標題 From discourse to logic with Stanford CoreNLP and Treebank Semantics	5．発行年 2019年
3．雑誌名 Proceedings of the Sixteenth International Workshop of Logic and Engineering of Natural Language Semantics (LENLS 16)	6．最初と最後の頁 1-14
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　3件）

1．発表者名 Alastair Butler
2．発表標題 Parsed corpus development with a quick access interface
3．学会等名 Logic and Engineering of Natural Language Semantics 18 (LENLS18)（国際学会）
4．発表年 2021年

1．発表者名 Alastair Butler
2．発表標題 Knowledge acquisition from natural language with Treebank Semantics and Flora-2
3．学会等名 Proceedings of the Seventeenth International Workshop of Logic and Engineering of Natural Language Semantics (LENLS 17)（国際学会）
4．発表年 2020年

１．発表者名
Alastair Butler

２．発表標題
From discourse to logic with Stanford CoreNLP and Treebank Semantics

３．学会等名
Logic and Engineering of Natural Language Semantics (LENLS 16)（国際学会）

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
Treebank Semantics Parsed Corpus (TSPC) website
http://www.compling.jp/ajb129/tspc.html
Treebank Semantics website
http://www.compling.jp/ajb129/ts.html

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関