「バイリンガルコーパス・ナビゲーター」オンライン日伊並列コンコーダンサの構築と活用

Zotti Patrizia（奈良先端科学技術大学院大学、国立国語研究所）
Apolloni Riccardo，松本裕治（奈良先端科学技術大学院大学）

Compilation and Use of the Bilingual Corpus Navigator (BCN): A Japanese-Italian Online Concordancer

Patrizia Zotti (NAIST, NINJAL)
Riccardo Apolloni, Yuji Matsumoto (Nara Institute of Science and Technology)

要旨

日本語を学ぶイタリア学生の数は２０年の間にほとんど４倍に増加した（１９９３年の１９７８人から、２０１２年の７４２０人)。２０１２年には、日本語が全国２１の大学で教えられていた。それにも関わらず、オンラインリソースはほとんど存在しない。
近年では、言語処理で基本的なリソースと見なされてきた並列コーパスは、言語教育、辞書編集、翻訳の研究など多様な分野で重要であることが認められ始めている。しかし、より広く、より効果的な利用を確保するためには、簡単で直感的な方法で並列コーパスに含まれるすべての情報へのアクセスを可能にするプラットフォームを開発する必要がある。JAICOという１０，０００ペアー日伊並列コーパスと「バイリンガル・コーパス・ナビゲーター」という日伊並列オンライン・コンコーダンサを開発した。コーパスデータの検索結果をKWIC形式で、対応する対訳文と共に表示するツールである。

１．はじめに

近年では、言語処理で基本的なリソースと見なされてきた並列コーパスは、言語教育、辞書編集、翻訳の研究など多様な分野で重要であることが認められ始めている。データ駆動型学習（Data Driven Learning – DDL）では、外国語教育への実践利用には，コーパスの言語分析結果を教材やシラバスに応用する間接的利用と，コーパス検索から得られた用例を見て学習者自身が言語の規則性を発見する直接的利用が考えられている。しかし、利用の容易さを確保するためには、簡単で直感的な方法で並列コーパスを利用することができる検索ツールが必要である。

本稿では, JAICOという１０，０００ペアー日伊並列コーパスと「バイリンガル・コーパス・ナビゲーター」という日伊並列オンライン・コンコーダンサを紹介する。２節では、関連研究を紹介し、３節ではBCN (BilingualCorpusNavigator) 「バイリンガル・コーパス・ナビゲーター」の機能と使い方を紹介し、４節では，JAICO日伊コーパスを紹介し、５節でまとめを行う。

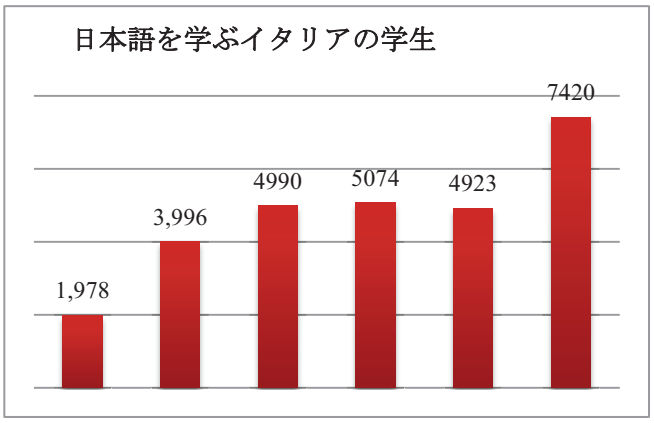

図１－日本語を学ぶイタリアの学生

出典: Japan Foundation “Survey on Japanese-Language Education Abroad” (1993-2012).

２．関連研究

伊日オンラインリソースを調べてみると、次の３つのリソースしか見つけることができない。それぞれを以下に簡単に説明する。

１）ITADICT（http://virgo.unive.it/itadict/）は、オンライン日伊辞書の計画である。現在は４万の見出し語を含んでいる (Mariotti, Mantelli 2012)。

２）OPUS Corpus (http://opus.lingfil.uu.se/)は、ウェブから取られた翻訳した文書のオープンソースデータベースである。多言語の中で、日伊パラレルデータにアクセスできるが、ジャンルは字幕だけであること、ソフトウェアのマニュアルが不十分であること、また、出力結果の日本語文が同じ語の繰り返しが含まれるなど完全な文として出力されないなどの問題がある。

３）Lang-8 （http://lang-8.com/）は言語交換ソーシャルネットワーキングサイトで、外国語で書いた文章をネイティブの話者が添消をしてくれる語学学習プラットフォームである。

３．ＢＣＮ「バイリンガル・コーパス・ナビゲーター」とは

BCN「バイリンガル・コーパス・ナビゲーター」は、イタリア語と日本語の二言語ＪＡＩＣＯコーパスの一部を搭載している（４節参照)。

BCNはhttp://cl.naist.jp/~zottip/tools.htmlにアクセスするか, あるいは「BilingualCorpusNavigator」のインターネット検索で最初に得られる検索結果をダブルクリックすると，図１のBCN の初期画面が現れ，検索作業が可能となる。

イタリア語、または日本語の検索したい語句を,「string」という文字列のボックスに入力し，ジャンルを「genre」のボックスに５つから（すべて、ニュース、小説、議事録、白書、雑録）選択し、検索方法を選択して、最後に「submit」ボタンを押すと，図２に示したようなKWIC（Key Words in Context) 検索結果の画面が得られる。

「Finding」と「Matching」の２つの検索方法がある。「Matching」の場合には、単一のトークンか、トークンの組み合わせ、または完全なテキストのセグメントを入力して、結果は優先度順に並べられる。図３示すように、すべての検索したトークンを含む文が最初に表

示され、その後検索したそれぞれトークンの１つを含む文が表示される。
図３は,「vita libertà sicurezza」の３つの伊トークンの「Matching」検索結果である。「vita libertà sicurezza」を含む伊文の検索結果が、画面の右側に１文表示され、「vita sicurezza」を含む伊文の検索結果が１文表示され、「vita libertà」を含む伊文の検索結果が１文表示され、「vita」、「sicurezza」、「libertà」のうち１つのトークンを含む伊文の検索結果が表示されている。または、それらの伊文に対応する日文が、画面の左に表示される。
「Finding」の検索方法の場合には、一つのトークンか、日本語の場合には、スペースがない文字列の検索しかできない。
ＢＣＮでは，図３と図４に示すように，検索語を含む日本語文とそれに対応するイタリア語文が一画面に表示されるので，日本語とイタリア語の文例を対照させながら学習することができる。
研究の現段階では、二つのインターフェスがある。一番目では、文のレベルで対応付けた７０００文の検索が可能で、イタリア語の検索の場合には「Matching」と「Finding」の検索が行われる。日本語の場合には「Finding」検索しか行われない。二番目では、単語レベルアラインされた白書の１００文の検索ができ、両言語でもMatchingとfindingの検索が行われる。すべてのデータに対して、単語をもちいた検索を実現することを目標にしているが、単語のアラインメントを全データに対して行うにはまだ時間を要する。現在の開発の実験段階では、データに関する検索は両言語とも文字レベルが対象であり、見出し語や品詞を用いた検索は、まだ行うことができない。

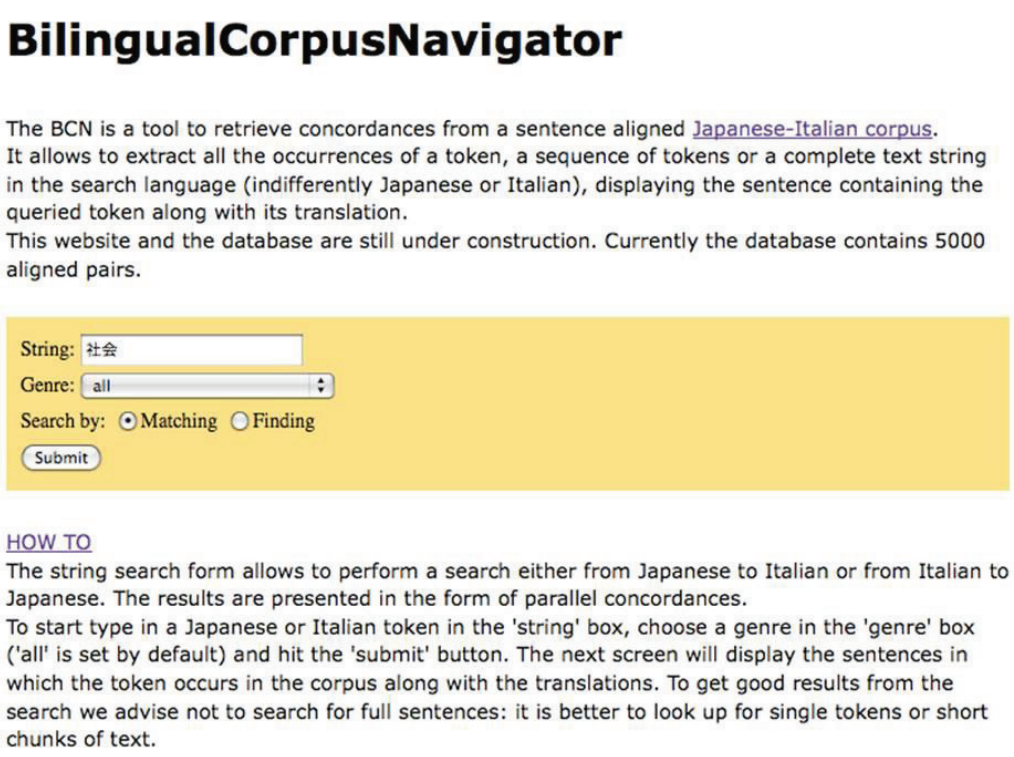

図　２　初期画面

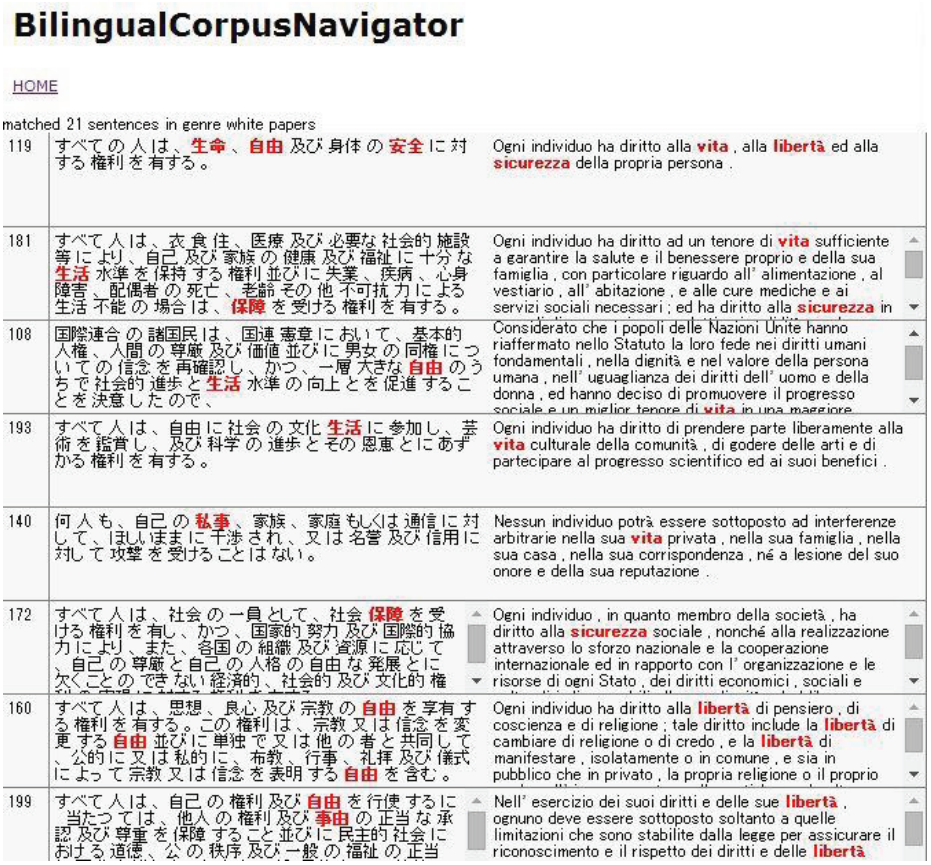

図　3　「vita libertà sicurezza」の「Matching」検索を示すス画面

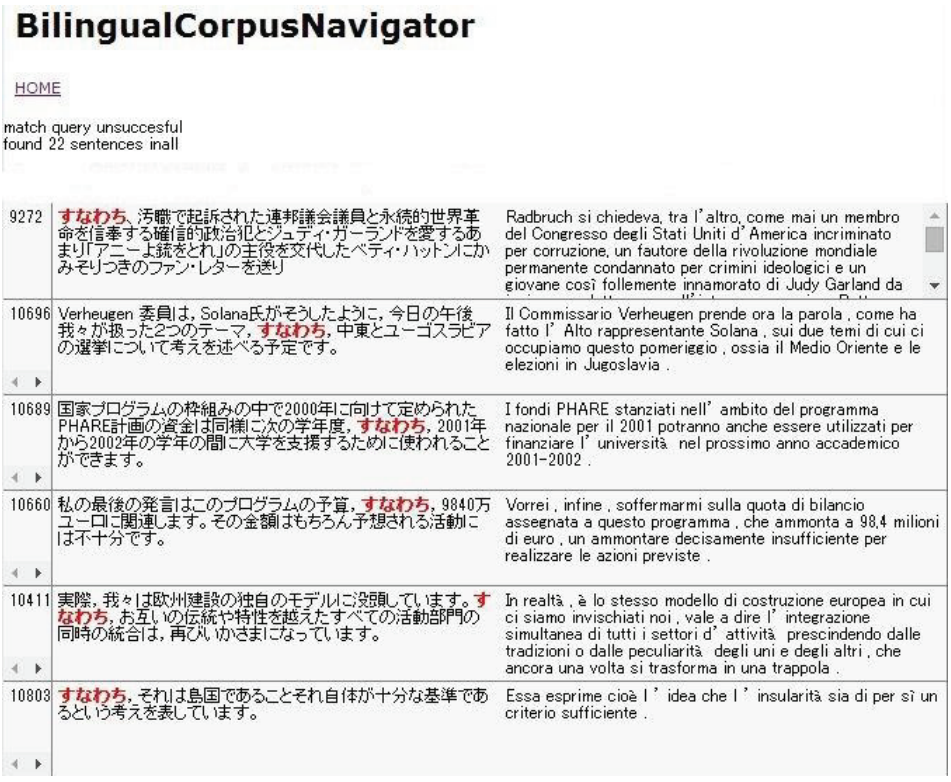

図　4　「すなわち」の「Finding」検索を示すス画面

４．JAICO 日伊パラレルコーパスとは

JAICO 「JApaneseItalianCorpus」という日伊並列コーパスのデータは，イタリア語・日本語各１万文の文対応付けコーパスである（Zotti 2013）。最初の６０００文は半自動的に対応付けされ、次の４０００文はChurch and Galeアルゴリズム（Gale, Church 1993）に基づく実装で自動的に対応付けされた（Zotti, Apolloni, Matsumoto 2010；JISA – JapaneseItalianSentence Aligner, http://cl.naist.jp/~zottip/tools.html）。

ＢＣＮで使用しているJAICOのデータは、７１００伊日並列文である（詳細については、図５を参照のこと）。最近、単語と単語の対応の作業を始めたが、まだオンラインBCNのデータベースに１００しかアップロードしていない。

			Japanese		Italian	
Domain	Year/s	Sentence Pairs	Tokens	Types	Tokens	Types
N	1989-2001	2000	54,849	5,873	41,238	6,452
PP	2000	2000	74,740	5,539	54,267	7,357
LW	1965-2004	3000	62,849	8,890	48,763	8,893
WP		100				
		7,100	**192,438**	**20,302**	**144,178**	**22,702**

N= News (Yomiuri Shinbun translated into Italian)- Utiyama, Isahara 2003; Nichols et al. 2010
PP= Parliamentary Proceedings (Europarl Shared Task ACL 2007 dataset translated into Japanese)
LW= Literary Works (Excerpts from Japanese novels and their translation)
WP= White Papers (Universal Declaration of Human Rights)

図 5　BCNで使用しているJAICOのデータ

５．まとめ

日伊並列コーパスのデータを検索するフリーウェアオンライン・コンコーダンサを開発した。現在の開発の実験段階では、7000対応付け文と100単語と単語対応付け文の検索が可能である。

これから、日本語とイタリア語について無料のツールを提供するために、コーパスやオンラインデータベースを増やす予定である。

参考文献

Barlow, M. (2008) Parallel Texts and Corpus-Based Contrastive Analysis, in Gómez González, M., Mackenzie, L. and González Alvarez, E. (eds.), *Current Trends in Contrastive Linguistics: Functional and Cognitive Perspectives*, Benjamins, 101-121.

Gale, W.A., Church, K.W. (1993) A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19/1, 75-102.

Koehn P. (2005) Europarl: a Parallel Corpus for Statistical Machine Translation, in *Conference Proceedings: the Tenth Machine Translation Summit, 79-86. Phuket, Thailand.*

Japan Foundation (1993, 1998, 2003, 2006, 2009, 2012) *Survey on Japanese-Language Education*

Abroad. Planning and Coordination Section, Japanese Language Dept., Japanese-Language Group. Tokyo.

Johansson, S. (1998) On the Role of Corpora in Cross-linguistic Research, in Johansson, S., S. Oksefjell (eds.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*. Amsterdam and Atlanta, GA, Rodopi, 3-24.

Mariotti, M.M., Mantelli A. (2012) ITADICT Project and Japanese Language Learning. *Acta Linguistica Asiatica* 2/2, 65-82.

Tiedemann, J. (2012) Parallel Data, Tools and Interfaces in OPUS, in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*

Utiyama M., Isahara H. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences, in *Proceedings of the 41st Annual Meeting of the ACL - Association for Computational Linguistics*, 72-79.

Zotti, P. (2013) Costruire un corpus parallelo Giapponese-Italiano. Metodologie di compilazione e Applicazioni, in M. Casari, P. Scrolavezza (eds), *Giappone, storie plurali*, 351-363. I libri di Emil-Odoya Edizioni. Bologna.

Zotti, P., Apolloni, R., Matsumoto, Y. (2014) Sentence Alignment of a Japanese-Italian Parallel Corpus. Towards a web-based interface. 言語処理学会第 20 回年次大会発表論文集, 23-26, 18 March 2014.

関連 URL

http://cl.naist.jp/ zottip ~ /