

## 「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 コンコーダンスの構築と活用

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・コンコーダンスを開発した。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・コンコーダンスを紹介する。2節では、関連研究を紹介し、3節ではBCN (Bilingual Corpus Navigator) 「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。

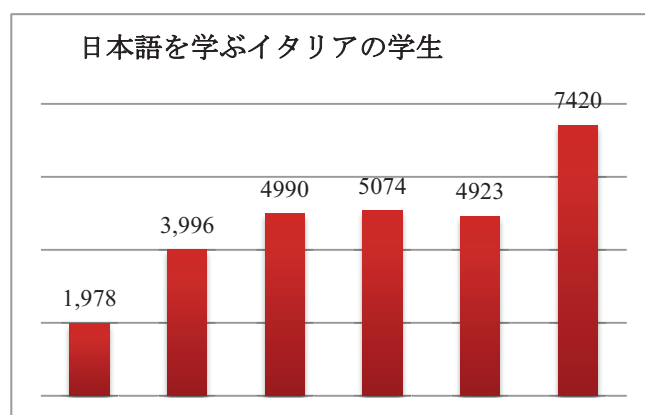


図1 - 日本語を学ぶイタリアの学生

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削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 J A I C O 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ースがある。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。

## BilingualCorpusNavigator

The BCN is a tool to retrieve concordances from a sentence aligned [Japanese-Italian corpus](#). It allows to extract all the occurrences of a token, a sequence of tokens or a complete text string in the search language (indifferently Japanese or Italian), displaying the sentence containing the queried token along with its translation. This website and the database are still under construction. Currently the database contains 5000 aligned pairs.

### HOW TO

The string search form allows to perform a search either from Japanese to Italian or from Italian to Japanese. The results are presented in the form of parallel concordances.

To start type in a Japanese or Italian token in the 'string' box, choose a genre in the 'genre' box ('all' is set by default) and hit the 'submit' button. The next screen will display the sentences in which the token occurs in the corpus along with the translations. To get good results from the search we advise not to search for full sentences: it is better to look up for single tokens or short chunks of text.

図 2 初期画面



## BilingualCorpusNavigator

HOME

matched 21 sentences in genre white papers

119	すべての人は、 <b>生命</b> 、 <b>自由</b> 及び身体の <b>安全</b> に対する権利を有する。	Ogni individuo ha diritto alla <b>vita</b> , alla <b>libertà</b> ed alla <b>sicurezza</b> della propria persona .
181	すべて人は、衣食住、医療及び必要な社会的施設等により、自己及び家族の健康及び福祉に十分な <b>生活</b> 水準を保持する権利並びに失業、疾病、心身障害、配偶者の死亡、老齢その他不可抗力による生活不能の場合は、 <b>保障</b> を受ける権利を有する。	Ogni individuo ha diritto ad un tenore di <b>vita</b> sufficiente a garantire la salute e il benessere proprio e della sua famiglia , con particolare riguardo all' alimentazione , al vestiario , all' abitazione , e alle cure mediche e ai servizi sociali necessari ; ed ha diritto alla <b>sicurezza</b> in
108	国際連合の諸国民は、国連憲章において、基本的人権、人間の尊厳及び価値並びに男女の <b>同権</b> についての信念を再確認し、かつ、一層大きな <b>自由</b> のうちに社会的進歩と <b>生活</b> 水準の向上とを促進することを決意したので、	Considerato che i popoli delle Nazioni Unite hanno riaffermato nello Statuto la loro fede nei diritti umani fondamentali , nella dignità e nel valore della persona umana , nell' uguaglianza dei diritti dell' uomo e della donna , ed hanno deciso di promuovere il progresso sociale e un miglior tenore di <b>vita</b> in una maggiore
193	すべて人は、自由にして社会の文化 <b>生活</b> に参加し、芸術を鑑賞し、及び科学の進歩とその恩恵とにあずかる権利を有する。	Ogni individuo ha diritto di prendere parte liberamente alla <b>vita</b> culturale della comunità , di godere delle arti e di partecipare al progresso scientifico ed ai suoi benefici .
140	何人も、自己の <b>私事</b> 、家族、家庭もしくは通信に対して、いかに干渉され、又は名譽及び信用に対して攻撃を受けることはない。	Nessun individuo potrà essere sottoposto ad interferenze arbitrarie nella sua <b>vita</b> privata , nella sua famiglia , nella sua casa , nella sua corrispondenza , né a lesione del suo onore e della sua reputazione .
172	すべて人は、社会の一員として、社会 <b>保障</b> を受け権利を有し、かつ、国家的努力及び国際的協力により、また、各国の組織及び資源に応じて、自己の尊厳と自己の人格の自由な発展とに欠くことのできない経済的、社会的及び文化的権	▲ Ogni individuo , in quanto membro della società , ha diritto alla <b>sicurezza</b> sociale , nonché alla realizzazione attraverso lo sforzo nazionale e la cooperazione internazionale ed in rapporto con l' organizzazione e le risorse di ogni Stato , dei diritti economici , sociali e
160	すべて人は、思想、良心及び宗教の <b>自由</b> を享有する権利を有する。この権利は、宗教又は信念を変更する <b>自由</b> 並びに単独で又は他の者と共同して、公的に又は私的に、布教、行事、礼拝及び儀式によって宗教又は信念を表明する <b>自由</b> を含む。	▲ Ogni individuo ha diritto alla <b>libertà</b> di pensiero , di coscienza e di religione ; tale diritto include la <b>libertà</b> di cambiare di religione o di credo , e la <b>libertà</b> di manifestare , isolatamente o in comune , e sia in pubblico che in privato , la propria religione o il proprio
199	すべて人は、自己の権利及び <b>自由</b> を行使するに当たっては、他人の権利及び <b>事由</b> の正当な承認及び尊重を保障すること並びに民主的 <b>社会</b> における道徳、公の秩序及び一般の福祉の正当	▲ Nell' esercizio dei suoi diritti e delle sue <b>libertà</b> , ognuno deve essere sottoposto soltanto a quelle limitazioni che sono stabilite dalla legge per assicurare il riconoscimento e il rispetto dei diritti e delle <b>libertà</b> .

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

## BilingualCorpusNavigator

HOME

match query unsuccessful  
found 22 sentences in all

9272	<b>すなわち</b> 、汚職で起訴された連判議会議員と永続的世界革命を信奉する確信的政治犯とジュディ・ガーランドを愛するあまり「アニー」銃をどれかの主役を交代したベティ・ハットンにかみそりつきのファン・レターを送り	Radbruch si chiedeva, tra l' altro, come mai un membro del Congresso degli Stati Uniti d' America incriminato per corruzione, un fautore della rivoluzione mondiale permanente condannato per crimini ideologici e un giovane così follemente innamorato di Judy Garland da
10696	Verheugen 委員は、Solana氏がそうしたように、今日の午後我々が扱った2つのテーマ、 <b>すなわち</b> 、中東とユーゴスラビアの選挙について考えを述べる予定です。	Il Commissario Verheugen prende ora la parola , come ha fatto l' Alto rappresentante Solana , sui due temi di cui ci occupiamo questo pomeriggio , ossia il Medio Oriente e le elezioni in Jugoslavia .
10689	国家プログラムの枠組みの中で2000年に向けて定められたPHARE計画の資金は同様に次の学年度、 <b>すなわち</b> 、2001年から2002年の学年の間に大学を支援するために使われることができます。	I fondi PHARE stanziati nell' ambito del programma nazionale per il 2001 potranno anche essere utilizzati per finanziare l' università nel prossimo anno accademico 2001-2002 .
10660	私の最後の発言はこのプログラムの予算、 <b>すなわち</b> 、9840万ユーロに関連します。その金額はもちろん予想される活動には不十分です。	Vorrei , infine , soffermarmi sulla quota di bilancio assegnata a questo programma , che ammonta a 98,4 milioni di euro , un ammontare decisamente insufficiente per realizzare le azioni previste .
10411	実際、我々は欧州建設の独自のモデルに没頭しています。 <b>すなわち</b> 、お互いの伝統や特性を越えたとすべての活動部門の同時の統合は、再びいかにまじっています。	In realtà , è lo stesso modello di costruzione europea in cui ci siamo invischiati noi , vale a dire l' integrazione simultanea di tutti i settori d' attività prescindendo dalle tradizioni o dalle peculiarità degli uni e degli altri , che ancora una volta si trasforma in una trappola .
10803	<b>すなわち</b> 、それは島国であることそれ自体が十分な基準であるという考えを表しています。	Essa esprime cioè l' idea che l' insularità sia di per sé un criterio sufficiente .

図 4 「すなわち」の「Finding」検索を示す画面

#### 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。

Domain	Year/s	Sentence Pairs	Japanese		Italian	
			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				
		<b>7,100</b>	<b>192,438</b>	<b>20,302</b>	<b>144,178</b>	<b>22,702</b>

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・コンコーダンスを開発した。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~/>