NII RCOS
Research Center for Open Science and Data Platform

CiNii Research多言語対応のご紹介

国立情報学研究所
長瀬 友樹

図書館総合展
2024年11月7日　於 パシフィコ横浜

本日の発表

1.学術情報検索サービスCiNii とは

2.多言語対応の目的

3.試行サービスの構築

4.評価

5.まとめ

1.学術情報検索サービスCiNii とは

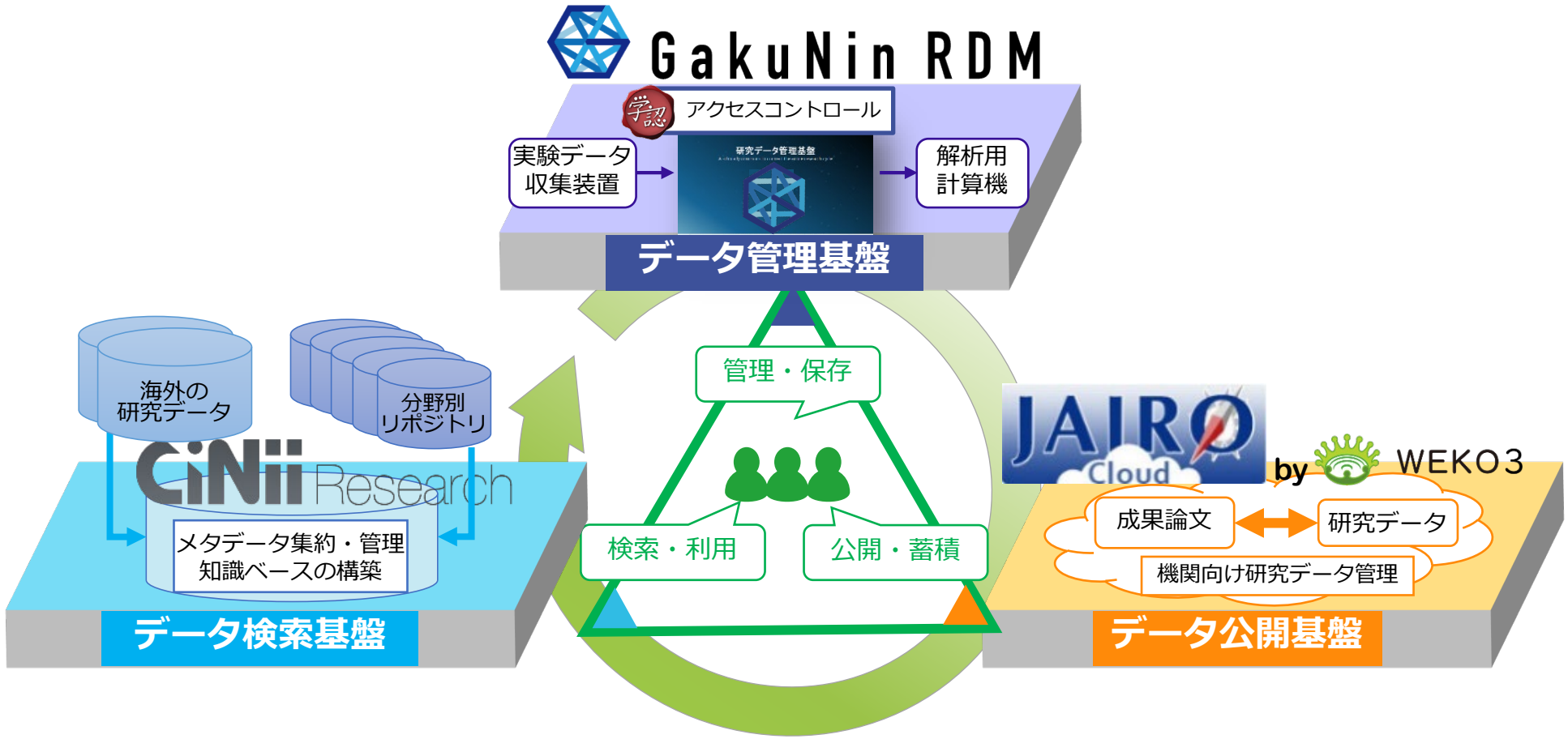
2.多言語対応の目的

3.試行サービスの構築

4.評価

5.まとめ

研究データ基盤：NII Research Data Cloud



NIIで開発・運用　2021年から正式公開

日本のオープンサイエンス推進の礎となるサービスの提供を目指す

学術検索基盤 CiNii Research

NII RCOS
Research Center for Open Science and Data Platform

一つの検索画面から多様な学術情報へ、イージーアクセス
NII RDCにおける研究データ共有の要となる機能

https://cir.nii.ac.jp/

1.学術情報検索サービスCiNii とは

2.多言語対応の目的

3.試行サービスの構築

4.評価

5.まとめ

多言語対応の目的

研究データの記録に用いられる言語の多様性が、FAIR原則のF（Findable）を損ねている可能性に着目。

自動翻訳活用による、CiNii収録データの“Findable”向上を試みる。

日本語で書かれた研究成果の、世界に向けた発信・利用促進。

FAIR原則： 研究データの管理と共有におけるベストプラクティスを示したガイドライン

- F - **発見しやすいこと**（Findable）
- A - **アクセスしやすいこと**（Accessible）
- I - **相互運用可能であること**（Interoperable）
- R - **再利用可能であること**（Reusable）

CiNii の登録データ・利用者内訳

CiNiiに登録されているデータの半数は日本語のみで書かれている

CiNii収録論文の言語別内訳

	論文件数	割合
日本語論文 (英語情報なし)	25,903,830	50.0%
日本語以外	25,996,158	50.0%

CiNii利用者の85%は日本国内からの利用である

CiNii の利用者内訳

	利用者数	割合
日本から	4,189,868	84.6%
海外から	828,490	15.4%

(2023 年 6 月 30 日のログを集計)

（参考）多言語対応へのニーズ

- **UNESCO**は、“Recommendation on Open Science”の中で、言語の平等性に言及している

https://unesdoc.unesco.org/ark:/48223/pf0000379949/PDF/379949eng.pdf.multi

> ***Equality of opportunities:***
> *all scientists and other open science actors and stakeholders, regardless of location, nationality, race, age, gender, income, socio-economic circumstances, career stage, discipline, **language,** religion, disability, ethnicity or migratory status, or any other grounds, have an equal opportunity to access, and contribute to and benefit from open science.*

- **COAR**でもタスクフォースを立ち上げて多言語に関する課題を議論


COAR Confederation of Open Access Repositories

Multilingual and Non-English Content

Multilingualism is a critical characteristic of a healthy, inclusive, and diverse research communications landscape. Publishing in a local language ensures that the public in different countries has access to the research they fund, and also levels the playing field for researchers who speak different languages. The Helsinki Initiative on Multilingualism in Scholarly Communication asserts that the disqualification of local or national languages in academic publishing is the most important – and often forgotten – factor that prevents societies from using and taking advantage of the research done where they live.

https://coar-repositories.org/what-we-do/multilingual-and-non-english-content/

1.学術情報検索サービスCiNii とは

2.多言語対応の目的

3.試行サービスの構築

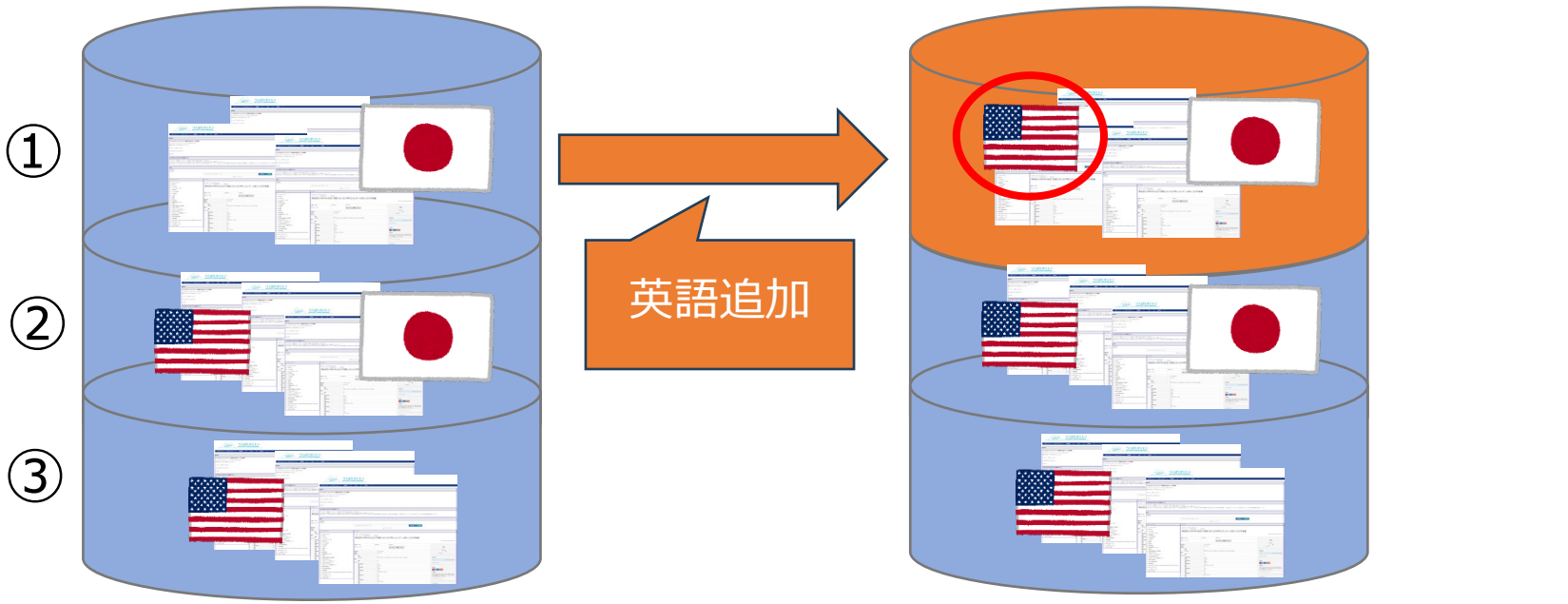
4.評価

5.まとめ

英語メタデータの追加

CiNiiのメタデータには
①**全部が日本語のみのもの** ------ ここだけを翻訳
②日本語と英語が混在するもの
③全部が英語のもの

A：従来のCiNii　　B：今回の試作システム

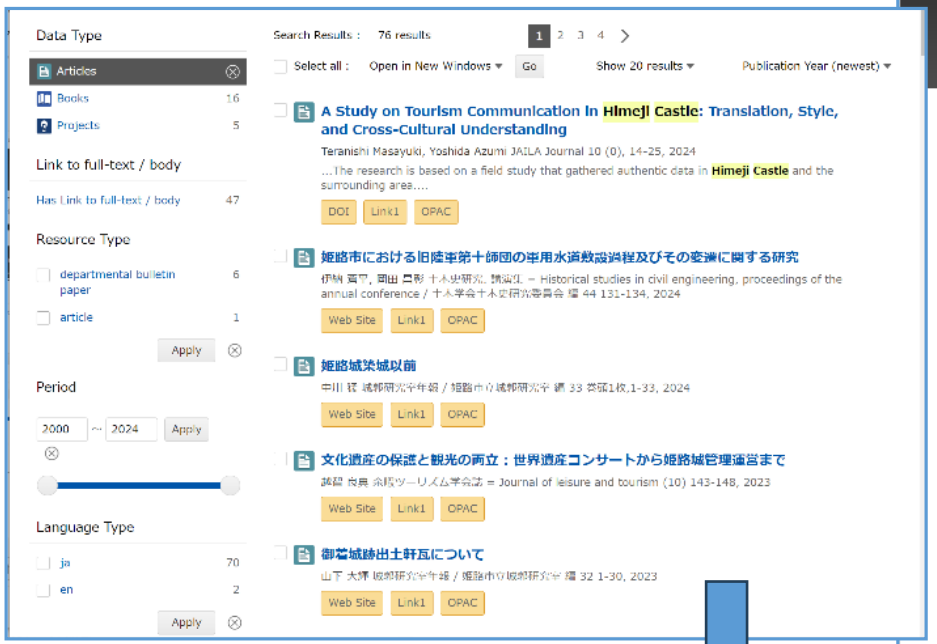


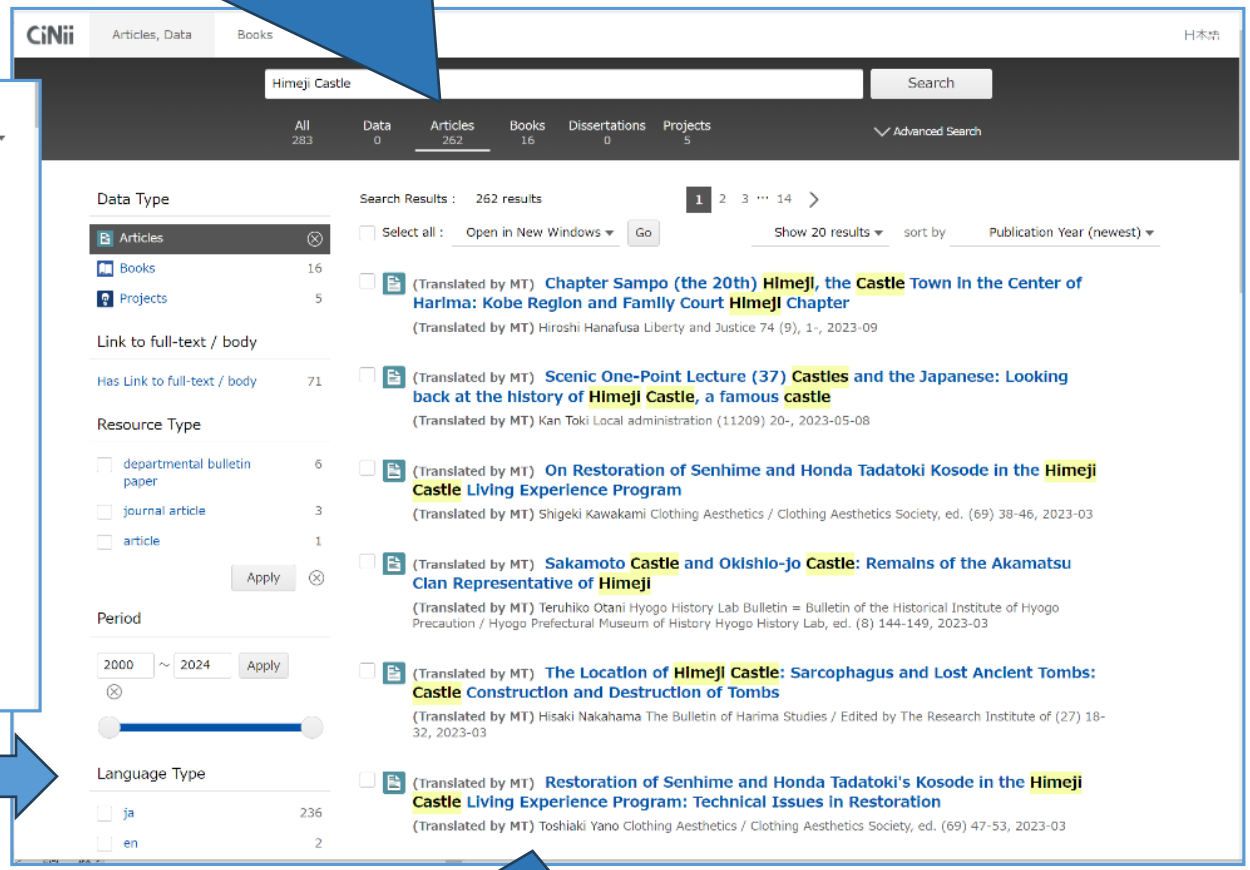
従来の（日本語の）検索結果には影響を与えない

翻訳実装済の画面イメージ（検索結果）

英語キーワードでヒットする論文が増えます

従来の検索結果表示画面





デモンストレーション
https://test.translate.cir.nii.ac.jp/?lang=en
https://cir.nii.ac.jp/

論文タイトルを英語で読めます

翻訳機能付きの検索結果表示画面

1.学術情報検索サービスC

2.多言語対応の目的

3.試行サービスの構築

4.評価

5.まとめ

評価：翻訳品質

評価方法

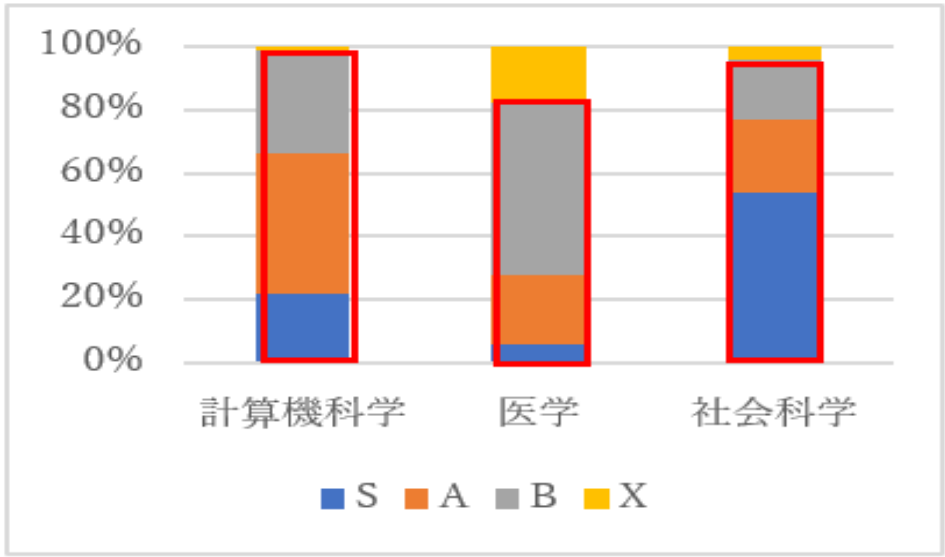
評価文　：100文×3分野　計300文
評価者　：1名（翻訳者）
評価方法：4段階主観評価

1　主観評価の評価基準

S	Aに加えて表現が自然であること
A	意味が正しく理解できること
B	一部間違いがあるが，大意は理解できること
X	S，A，B以外

評価結果

評価文の8割以上がB評価以上



スクリーニング目的であれば十分に使えるレベル

評価：検索ヒット件数

英語キーワードに対する検索ヒット件数

	キーワード	従来のCiNii A	翻訳適用後 B	ヒット増加率	日本語で検索(参考)
1	"Himeji Castle"	63	260	413%	286
2	"Plate tectonics"	1433	1630	113%	325
3	"Fossa Magna"	1013	1352	113%	796
4	"Important Intangible Cultural Property"	18	141	783%	145
5	"Zazen"	145	510	352%	688
6	"ABC conjecture"	38	59	155%	138
7	"Halley's comet"	45	189	420%	190
計		2755	4141	150%	

翻訳適用によりヒット件数が５０%増加

増加率はキーワードによって幅がある
日本固有の地名や文化に関連するもの（”Himeji Castle”, ”Zazen”, etc.）の増加率高い

ラボサイトで公開中

翻訳機能付きCiNiiは、10/21よりCiNii Labs上で公開を開始。

https://labs.ci.nii.ac.jp/

「CiNii Labs」は、CiNiiに関係する実験的なサービスやコンテンツを公開するポータルサイトです（3月スタート）。

今後の予定（課題）

- **英語メタデータの整合性維持**
  日本語のコンテンツ追加およびメタデータの修正に同期して、翻訳結果も修正（追加、更新）されるように機能強化をしたうえで、 2025年度中のCiNii Research正式サービスへの導入を予定している。

- **自動翻訳の精度向上**
  未知語（特に人名）への対応や専門用語の同義語の扱いなど、自動翻訳の精度向上について、エンジン開発元の情報通信研究機構（NICT）との共同研究の中で取組んでいく。

- **論文著者とのコラボレーション**
  著者自身による翻訳（あるいは著者による自動翻訳の修正）が容易に実現できるインターフェースを備えるなど，人と機械が協調できる枠組みについて検討していく。

- **英語以外の言語へ展開**

- **リポジトリ（NII RDCの公開基盤）等への自動翻訳適用対象の拡大**

まとめ

- 自動翻訳機能を活用してCiNii Researchのメタデータを英語化することにより、英語キーワードによる日本語論文のヒット数が大幅に増加することを確認。

- 試行サービスを構築し、10/21にCiNii Labsで公開を開始した。

- 試行サービスでは、検索結果として表示される日本語の論文リストを英語で表示する機能を併せて実装しており、日本語が理解できない利用者でも日本語論文の検索が可能になっている。

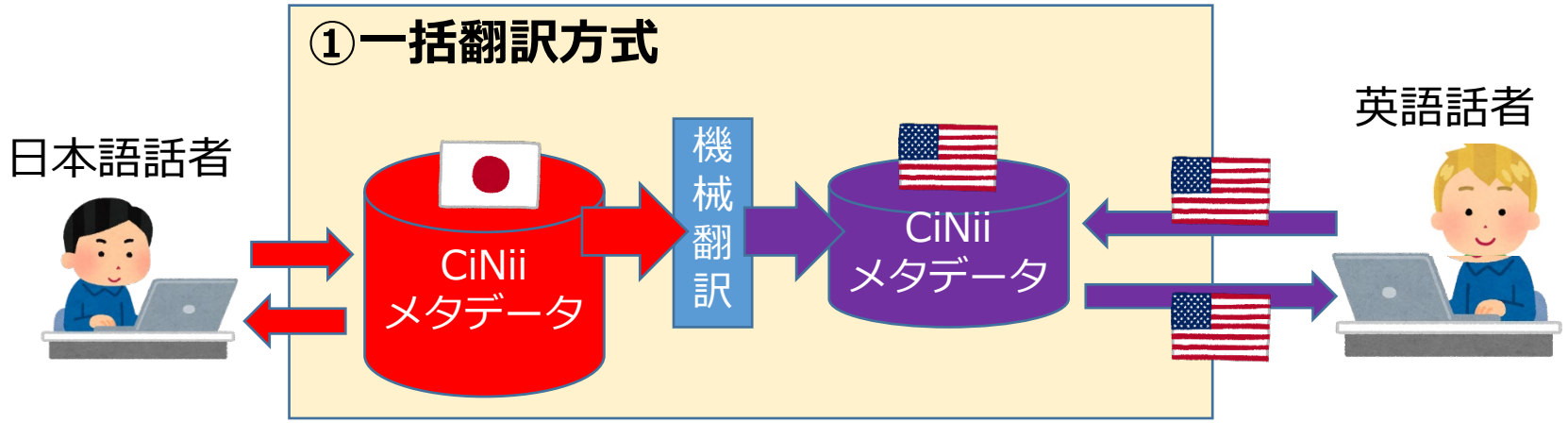
- 試行の結果を踏まえ、2025年度中に自動翻訳機能をCiNii Researchの正式サービスに組み込んで提供する計画である。

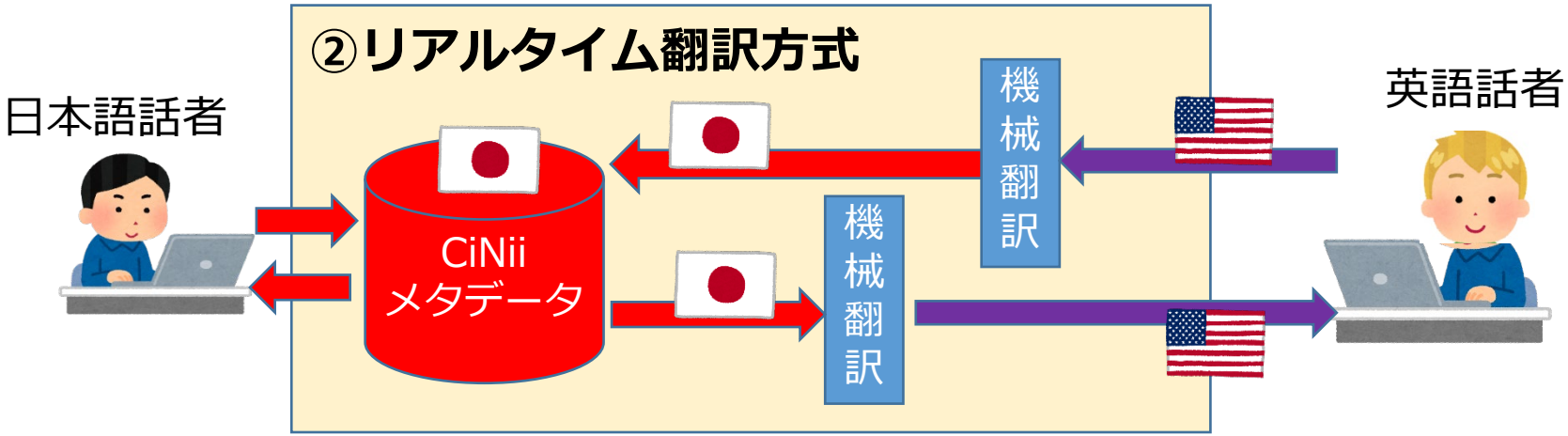
ご清聴ありがとうございました

以下、手持ち資料

（質疑応答時に適宜表示）

自動翻訳の適用方式の検討





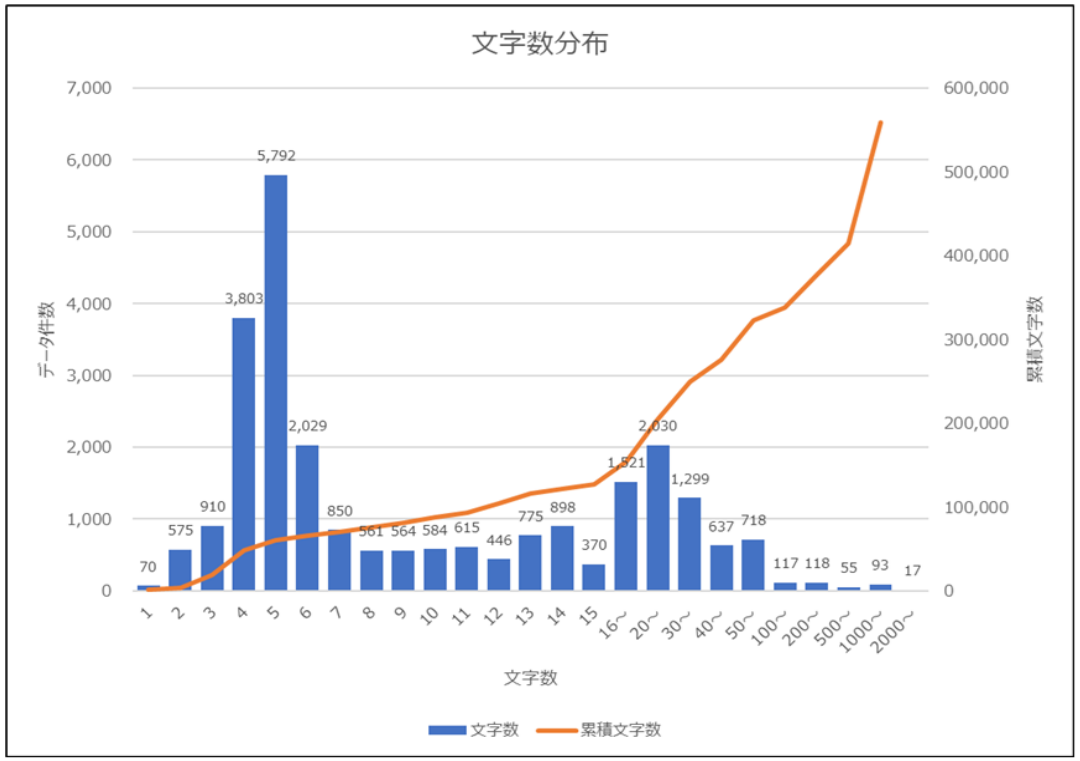
自動翻訳の適用方式の検討

	①一括翻訳方式	②リアルタイム翻訳方式
ストレージ容量	×翻訳結果の検索インデックスへの保存が必要	○検索インデッテクスへの追加不要でストレージ容量を消費しない
計算量（翻訳量）	△導入時にすべてのコンテンツを一括翻訳する	△キーワードが入力される度にキーワードを翻訳する
サービスのレスポンス	○検索時に翻訳実行が不要のため高速	×キーワード翻訳と検索結果翻訳を検索の都度実行するため遅い
論文単位の翻訳カスタマイズ	○論文単位で翻訳結果の修正が容易	×自動翻訳の論文単位の翻訳チューニングは困難
翻訳精度(検索ヒット率)	○文章単位の翻訳なので文脈を反映した多義語の翻訳精度が高い	×単語単位の翻訳となるため多義語の解釈を誤ることがある

サービスレスポンス、論文単位のカスタマイズの観点を重視し一括翻訳方式を採用

（参考）翻訳ログ分析

- 翻訳データ件数　　：2600万件
- １件あたり翻訳処理回数：10.6回
- 翻訳１回あたり文字数　：20.5文字



5文字と20文字あたりに、
２つのピーク

評価：翻訳品質（課題）

課題

- 未知の固有名詞（人名、略語、カタカナ語など）

システムにとって未知の名前が訳せない
採用したエンジンでは日本語表記のまま出力される

原文	訳文
佐藤 伊久子	Sato 伊久子
静大・理・地球科学	静大, Science and Earth Sciences
地理的条件に基づくジオ・エバキュエイタビリティ指標を用いて	Using Geo エバキュエイタビリティ indicators based on geographical conditions

評価：翻訳品質（課題）

課題

➢多訳語（同義語）

専門用語の訳語が参照訳と異なる事例が数多く見つかった。

（例）眼底鏡検査　⇔　*ophthalmoscopy*（参照訳の訳語）
fundoscopy（自動翻訳の訳）★間違いではない

意味が同じでも標記が異なるキーワードでは検索がマッチしない

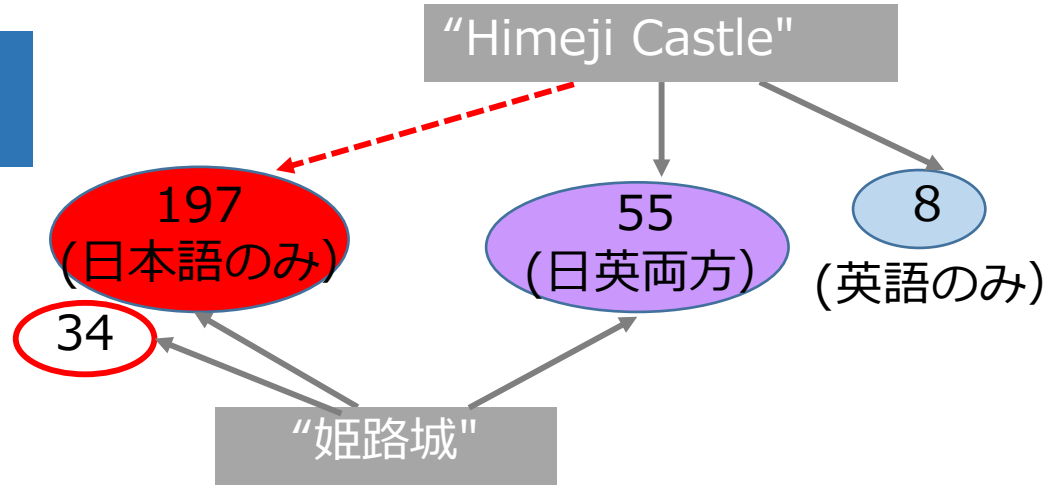
翻訳機能導入時に限定されない、検索システムで共通の既存課題

検索結果の内訳

検索結果の増加率が大きい例

英語のみの論文の比率が低い

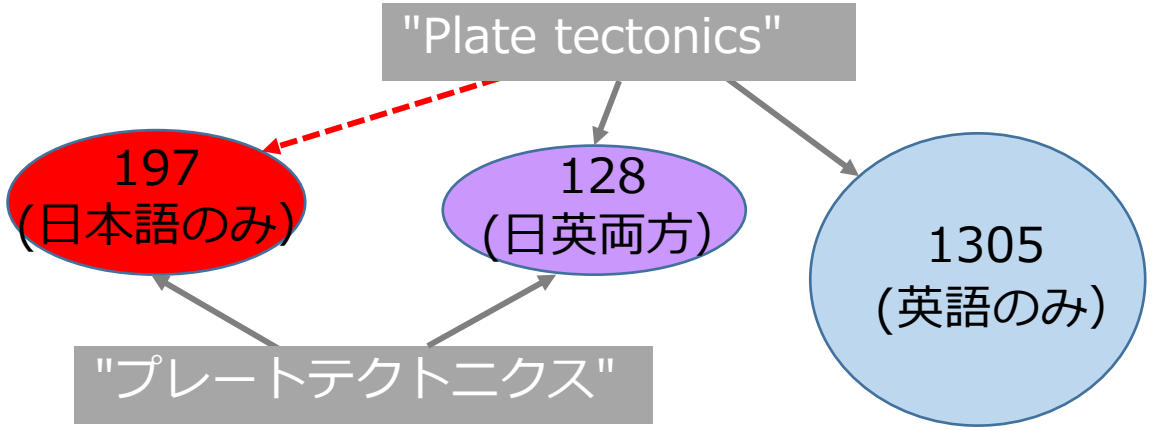
63件➡260件



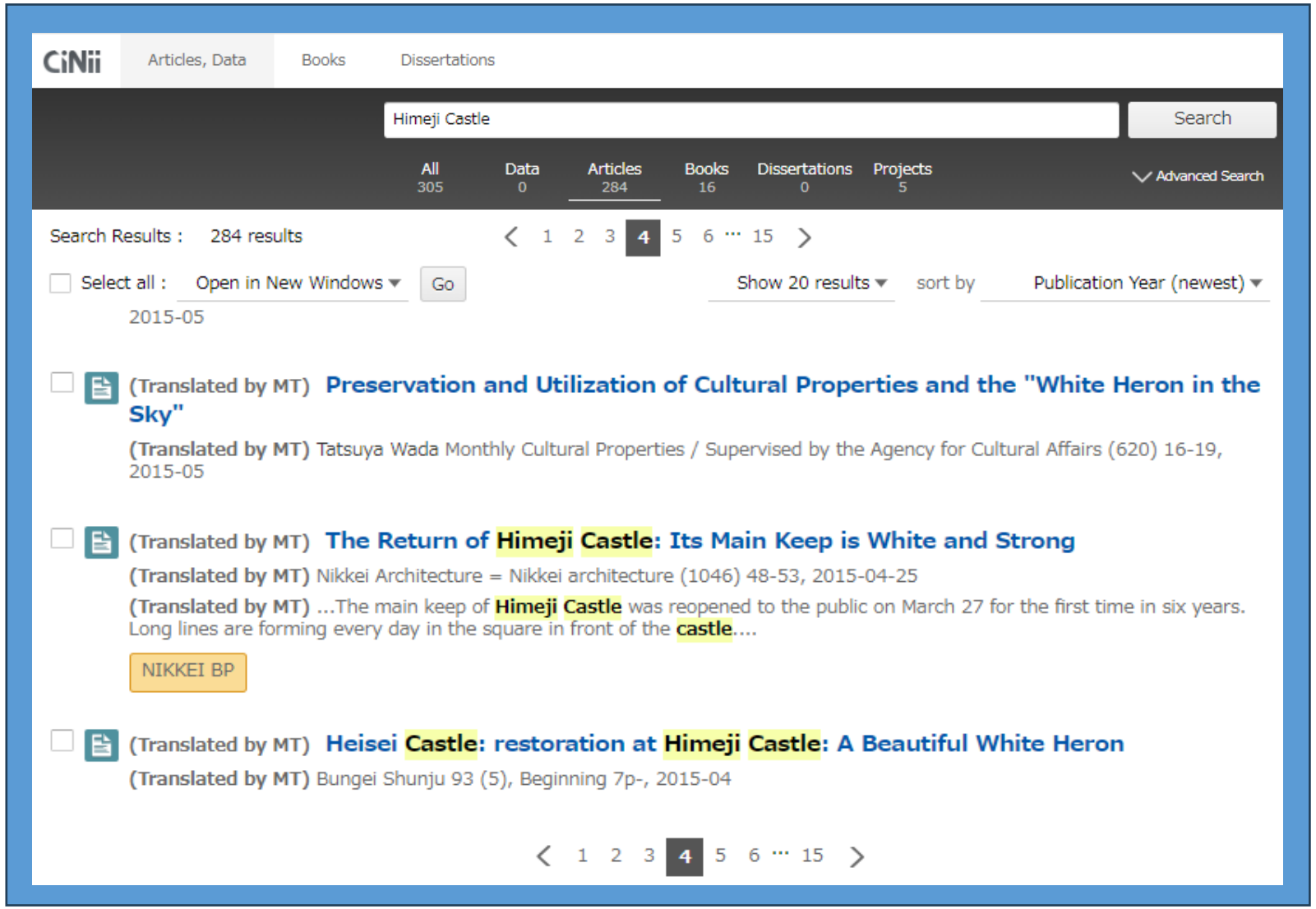
検索結果の増加率が小さい例

英語のみの論文の比率が高い

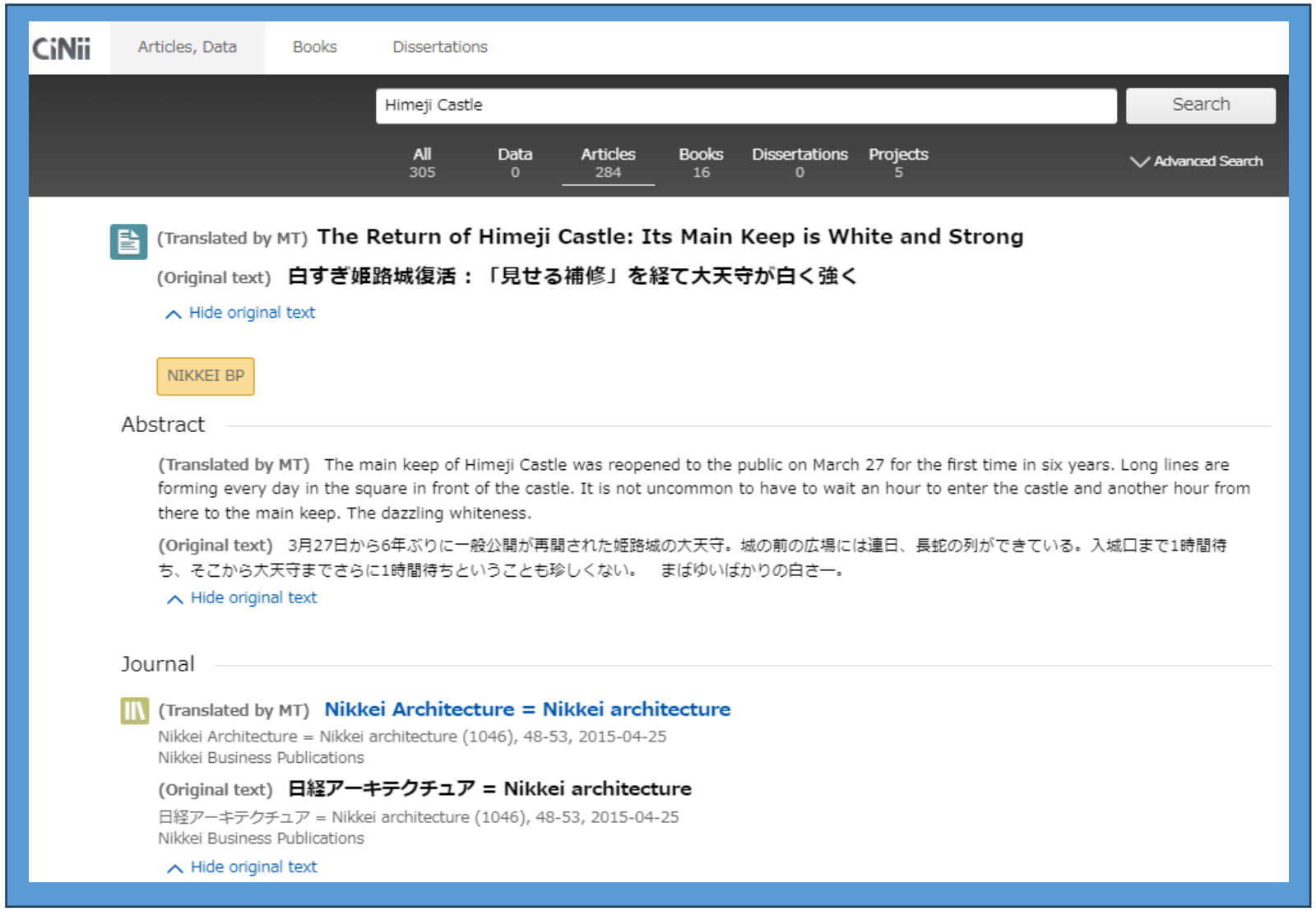
1433件➡1630件



翻訳実装済の画面イメージ（検索結果）



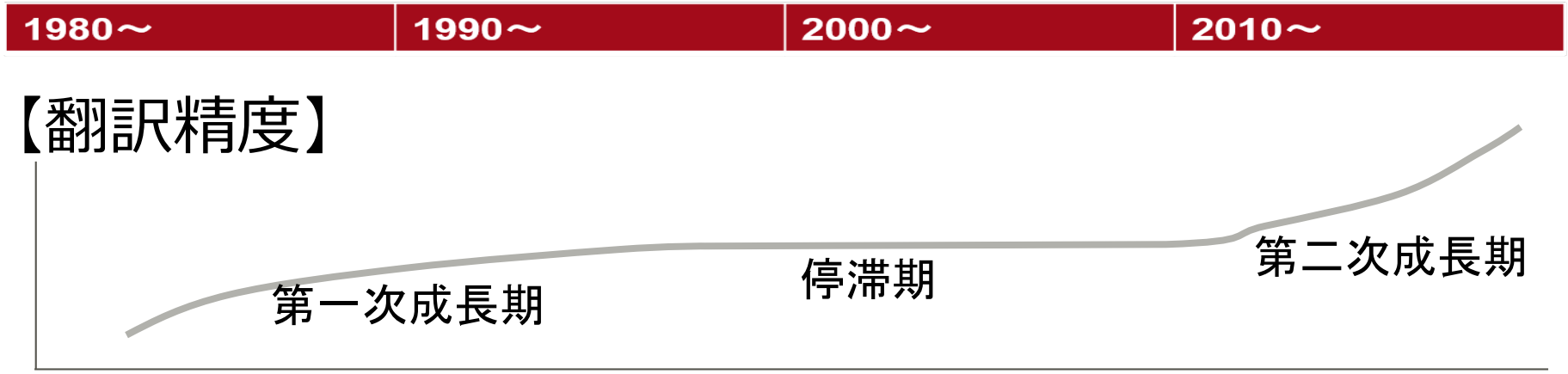
翻訳実装済の画面イメージ（詳細ページ）



翻訳済のjson

{"_id": 1010282256885782913,
"content": {"title": [{"notation": [{"language": "ja", "text": "世界遺産のより深い楽しみ方・姫路城"}], "type": "main"}],
"publication": {"publicationTitle": [{"language": "ja", "text": "日本歴史"}], "volume": "824", "startingPage": "8", "endingPage": "14", "issued": "2017",
"jointInternationalResearch": false},
"creator": [{"name": [{"language": "ja", "text": "中井 均"}, {"language": "en", "text": "NAKAI HITOSHI"}],
"affiliation": [{"language": "ja", "text": "滋賀県立大学"}],
"personIdentifier": [{"type": "ERAD", "value": "10621427"}, {"type": "CRID", "value": 1420564276159830400}, {"type": "NRID", "value": "1000010621427"},
{"type": "CINII_AUTHOR_ID", "value": "DA11023041"}, {"type": "URI", "value": "https://ci.nii.ac.jp/author/DA11023041#entity"}, {"type": "URI", "value":
"https://viaf.org/viaf/sourceID/NII%7CDA11023041"}, {"type": "NRID", "value": "9000382140483"}, {"type": "NRID", "value": "9000004300193"}, {"type": "NRID",
"value": "9000254585088"}, {"type": "NRID", "value": "9000283662643"}, {"type": "NRID", "value": "9000004608754"}]}],
"project": [{"notation": [{"language": "ja", "text": "戦国時代における石垣技術の考古学的研究"}, {"language": "en", "text": "Archaeological study of stone wall technology
in the Warring States period"}], "projectIdentifier": [{"type": "CRID", "value": 1040282256885782784}, {"type": "KAKEN", "value": "KAKENHI-PROJECT-
16K03160"}, {"type": "URI", "value": "https://kaken.nii.ac.jp/grant/KAKENHI-PROJECT-16K03160/"}]}],
"localIdentifier": 1010282256885782913,
"dataType": "Article",
"dataSourceIdentifier": [{"type": "KAKEN", "value": "PRODUCT-21209310"}], "productIdentifier": [{"type": "CRID", "value": 1010282256885782913}],
"resourceType": "journal article", "reviewed": false, "accessLevel": false},
"content_translated": {"title": [{"notation": [{"language": "en", "text": "A Deeper Way to Enjoy the World Heritage Site - Himeji Castle"}], "type": "main"}],
"publication": {"publicationTitle": [{"language": "en", "text": "Japanese History"}], "issued": "2017", "jointInternationalResearch": false},
"creator": [{"name": [{"language": "en", "text": "Hitoshi Nakai"}], "affiliation": [{"language": "en", "text": "Shiga Prefectural University"}], "personIdentifier":
[{"type": "ERAD", "value": "10621427"}, {"type": "CRID", "value": 1420564276159830400}, {"type": "NRID", "value": "1000010621427"}, {"type":
"CINII_AUTHOR_ID", "value": "DA11023041"}, {"type": "URI", "value": "https://ci.nii.ac.jp/author/DA11023041#entity"}, {"type": "URI", "value":
"https://viaf.org/viaf/sourceID/NII%7CDA11023041"}, {"type": "NRID", "value": "9000382140483"}, {"type": "NRID", "value": "9000004300193"}, {"type": "NRID",
"value": "9000254585088"}, {"type": "NRID", "value": "9000283662643"}, {"type": "NRID", "value": "9000004608754"}]}], "project": [{"notation": [{"language":
"en", "text": "Archaeological Study of Stone Wall Technology in the Warring States Period"}],
"projectIdentifier": [{"type": "CRID", "value": 1040282256885782784}, {"type": "KAKEN", "value": "KAKENHI-PROJECT-16K03160"}, {"type": "URI", "value":
"https://kaken.nii.ac.jp/grant/KAKENHI-PROJECT-16K03160/"}]}]},
"authors": [{"language": "ja", "text": "中井　均"}],
"authors_translated": [{"language": "en", "text": "Hitoshi Nakai"}],
"updated": 1676619034, "product_type": "article", "integrated_ids": [1010282256885782913]}

（参考）自動翻訳の技術/精度



ニューラルネット翻訳技術により、各段に翻訳精度が上がった。
分野にチューニングすれば、人間翻訳を超えることも珍しくない。
翻訳会社でも、自動翻訳の活用による翻訳コスト削減が目下の関心事。