

翻訳ゆらぎ

翻訳ゆらぎや関連用語をグラフで視覚化し
翻訳範囲を共有する

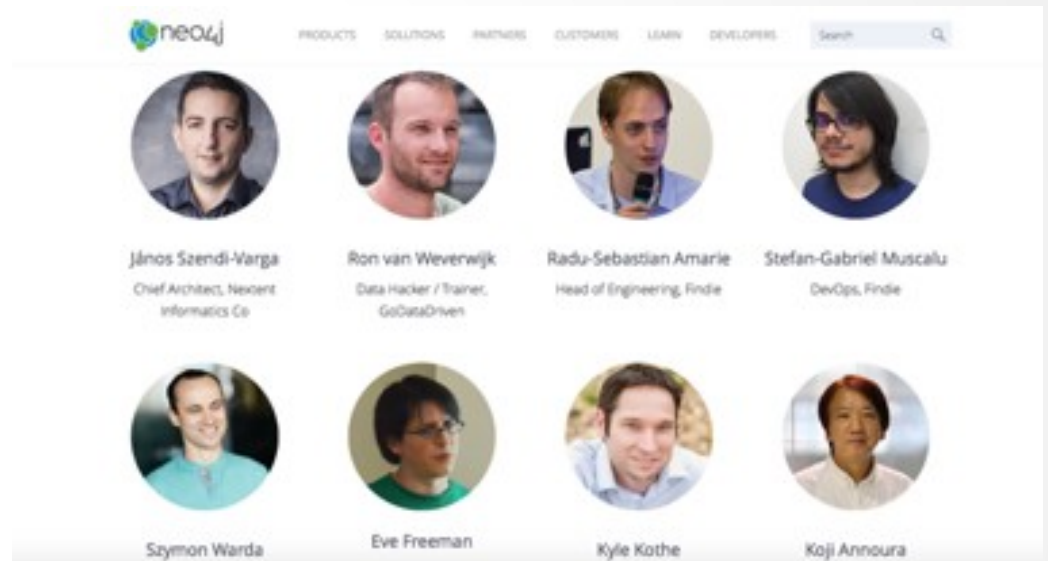
2019-05-25, 13:30 – 13:55

Koji Annoura

koji@annoura.com
@kojiannoura

about Me

- Neo4j
 - founder Neo4j UsersGroup Tokyo
 - Neo4j Ambassador
 - Certificate Neo4j Professional
- Agile
 - Certified Scrum Master
 - Certified Scrum Product Owner
 - LeSS Practitioner
- Time Management
 - The Pomodoro Technique® Trainers
- Member
 - LibreOffice Japanese Team
 - Japan PostgreSQL Users Group (JPUG)
- Coffee
 - UCC Certified Coffee Professional
 - SCAJ Coffee Meister
 - JCQA Certified Associate Instructor in Coffee



翻訳のゆらぎ

- Computerの日本語訳は？
 - 電子計算機
 - 電算機
 - コンピュータ
 - コンピューター
- どれが正しいわけではない
 - 時代とともに変わる
 - でも、1つの中ではなるべく統一してほしい

コンピュータ

- JIS Z 8301:1951

- 末尾に長音が付く言葉は、3文字までならそのまま記述する。3文字を超える場合は、末尾の長音を省く
- 3文字まで
 - キー
 - フロー
- カナ3文字を超える例
 - コンピューターは、コンピュータ
 - プリンターは、プリンタ
 - データーは、データ

コンピューター

- JIS Z 8301:2011

- 長音符号を付けるか，付けないかについて厳格に一定にすることは困難であると認め，各用語集の表記をそれぞれの専門分野の標準とするが，**長音符号は，用いても略しても誤りでない**ことにしている。

G.6.2.2 英語の語尾に対応する長音符号の扱い

英語の語尾に対応する長音符号の扱いは，通常，次による。

なお，英語の語末の -er, -or, -ar などは，ア列の長音とし，長音符号を用いて表すものに当たるとみなす。

a) 専門分野の用語の表記による。

する場合に，長音符号を付けるか，付けないかについて厳格に一定にすることは困難であると認め，各用語集の表記をそれぞれの専門分野の標準とするが，長音符号は，用いても略しても誤りでないことにしている。

b) 規格の用語及び学術用語にない用語の語尾に付ける長音符号は，表 G.3 による。

表 G.3-外来語の表記に語尾の長音符号を省く場合の原則

原則	例
a) その言葉が3音以上の場合には，語尾に長音符号を付けない。	エレベータ (elevator)
b) その言葉が2音以下の場合には，語尾に長音符号を付ける。	カー (car), カバー (cover)
c) 複合の語は，それぞれの成分語について，上記 a)又は b)を適用する。	モーターカー (motor car)
d) 上記 a)~c)による場合で，長音符号を書き表す音 (例 1)，はねる音 (例 2) 及びつまる音 (例 3) は，それぞれ1音と認め，よう (拗)音 (例 4) は1音と認めない。	1 テーパー (taper) 2 ダンパ (damper) 3 ニッパ (nipper) 4 シャワー (shower)

Pootle (Online translation tool)

- <https://translations.documentfoundation.org/ja/>

検索

致命的なエラーを修正 **5,153** 提案の査読 **410** 翻訳続行 **471,136** すべて表示 **2,276,769**

名前	進捗状況	最終更新日時	重大	提案	不完全	合計
LibreOffice Online	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Pootle 1 日	0	3	340	2,126
LibreOffice master – Help	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Pootle 1 日	1,052	187	138,020	473,056
LibreOffice master – UI	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Pootle 1 日	2	43	8,087	104,808
LibreOffice 6.2 – Help	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Pootle 3 週間	1,073	0	127,362	467,087
LibreOffice 6.2 – UI	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Pootle 3 週間	67	1	2,473	99,500
Website	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Pootle 1 か月	0	9	649	2,427
LibreOffice 6.0 – UI	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Jun NOGATA 4 か月	67	6	1,614	98,200
LibreOffice 6.1 – UI	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Jun NOGATA 4 か月	58	23	1,649	98,761
LibreOffice 6.1 – Help	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Pootle 6 か月	1,175	12	103,569	466,208
Impress Remote (Android)	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Ikuya Awashiro 3 年	0	56	0	259
Impress Remote (iOS)	<div style="width: 100%;"><div style="width: 100%;"></div></div>	matuaki 4 年	0	62	274	346
Terminology	<div style="width: 100%;"><div style="width: 100%;"></div></div>	Ikuya Awashiro 6 年	0	8	0	1,128
LibreOffice 6.0 – Help	<div style="width: 100%;"><div style="width: 100%;"></div></div>		1,659	0	87,099	462,863

LibreOffice master – UI

- Browse
 - ブラウズ (Browse)
 - 検索 (Search)

The screenshot shows the LibreOffice master UI for the string 'Browse'. The breadcrumb path is 'Japanese → LibreOffice master – UI → accessibility/messages.po → ユニット #137790249'. The string ID is 'RID_STR_ACC_NAME_BROWSEBUTTON' and the location is 'strings.hrc:25'. The current language is 'English'. The string 'Browse' is shown in the input field, with the Japanese translation 'ブラウズ' entered. A '提案' (Propose) button is visible. The '用語' (Glossary) section shows 'Browse 閲覧'. The '類似翻訳' (Similar Translations) section lists two entries: 'Browse ブラウズ' and 'Browse 検索', both attributed to '誰かによる "LibreOffice 5.3 – Help (EOL)" プロジェクトの翻訳'.

Japanese → LibreOffice master – UI → accessibility/messages.po → ユニット #137790249

優先度: 1.0 English この文字列の問題を報告する

文脈: RID_STR_ACC_NAME_BROWSEBUTTON

場所: strings.hrc:25

用語: Browse 閲覧

コメント: be4e7

Browse

ブラウズ

2年

類似翻訳

Browse	ブラウズ
誰かによる "LibreOffice 5.3 – Help (EOL)" プロジェクトの翻訳	
Browse	検索
誰かによる "LibreOffice 5.3 – Help (EOL)" プロジェクトの翻訳	

提案

LibreOffice master – UI

- Check
 - チェック
 - チェックする

Japanese → LibreOffice master – UI → accessibility/messages.po → ユニット #137790253

優先度: 1.0

文脈:
RID_STR_ACC_ACTION_CHECK

場所: strings.hrc:29

用語:
Check bit チェックビット

コメント:
zZTzc

English

この文字列の問題を報告する

Check

チェックする

2年

類似翻訳

check	チェック
Pootle による "LibreOffice 5.4 – UI" プロジェクトの翻訳 ()	
Check	チェックする
誰か による "LibreOffice 5.3 – UI (EOL)" プロジェクトの翻訳	

提案

手順

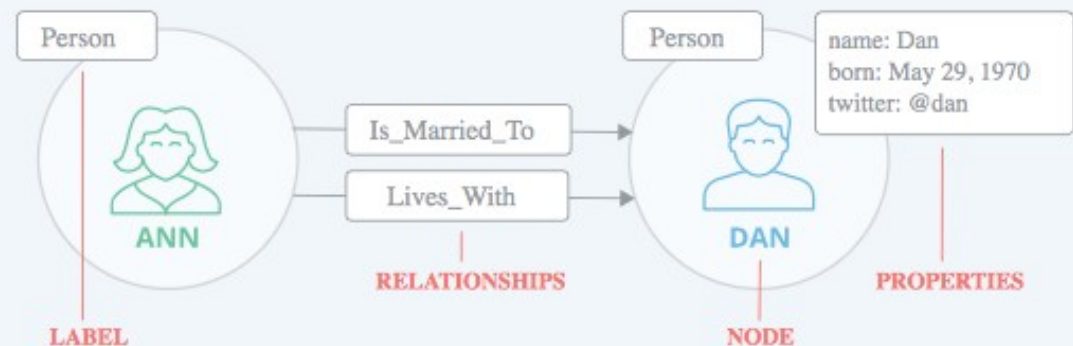
- Pootle
 - <https://translations.documentfoundation.org/ja/>
- Calc
 - PoからCalcにデータを作る (.po data to Calc)
- CSV
 - Kanji code to UTF-8
- Neo4j
 - グラフを作成して翻訳を共有する (CSV to Graph)

Neo4jとは

- ネイティブなグラフデータベース
Native Graph Database
 - ノード (Nodes)
 - リレーションシップ (Relationships)
 - プロパティ (Properties)
 - ラベル (Labels)

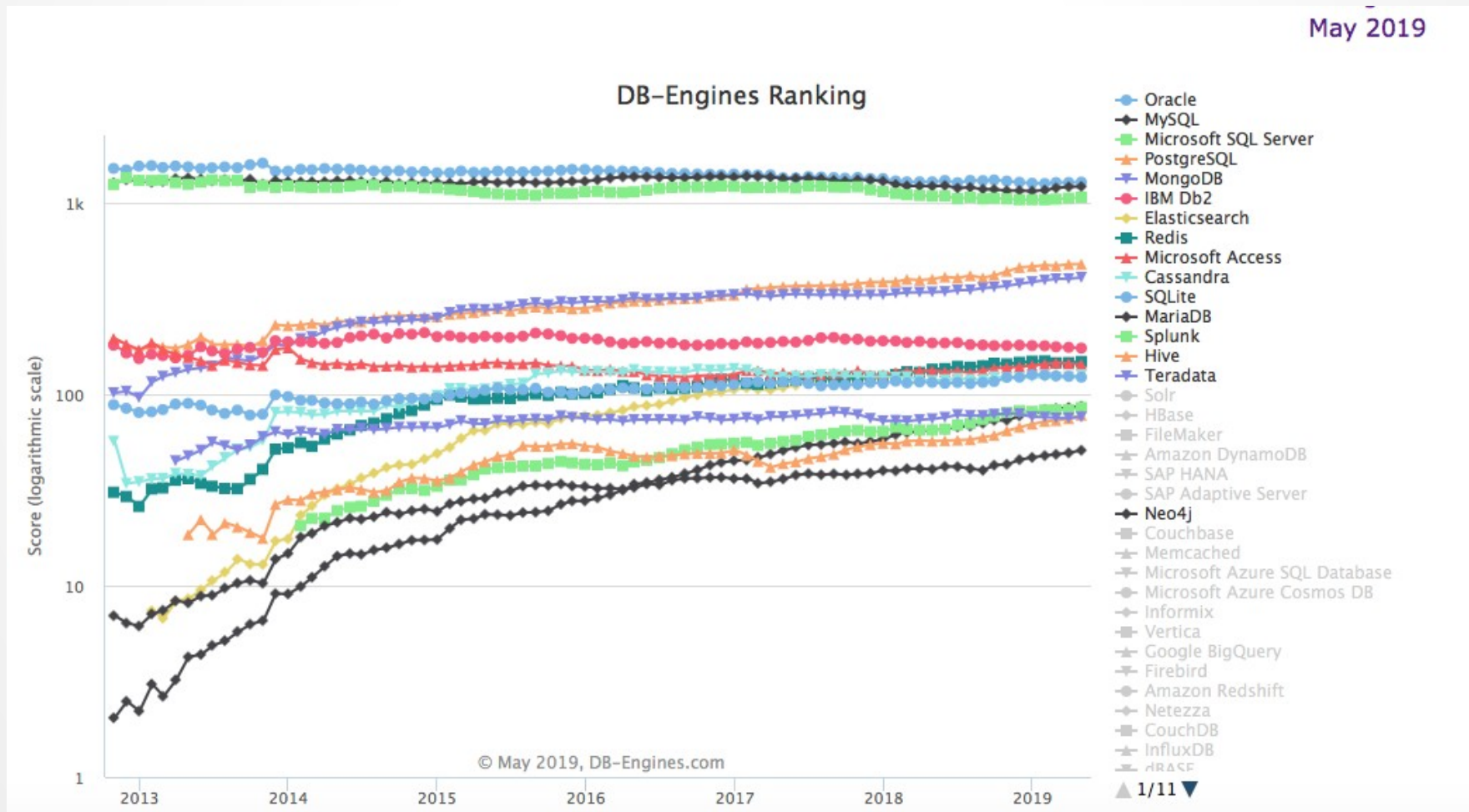
Neo4j Basics

The Labeled Property Graph Model



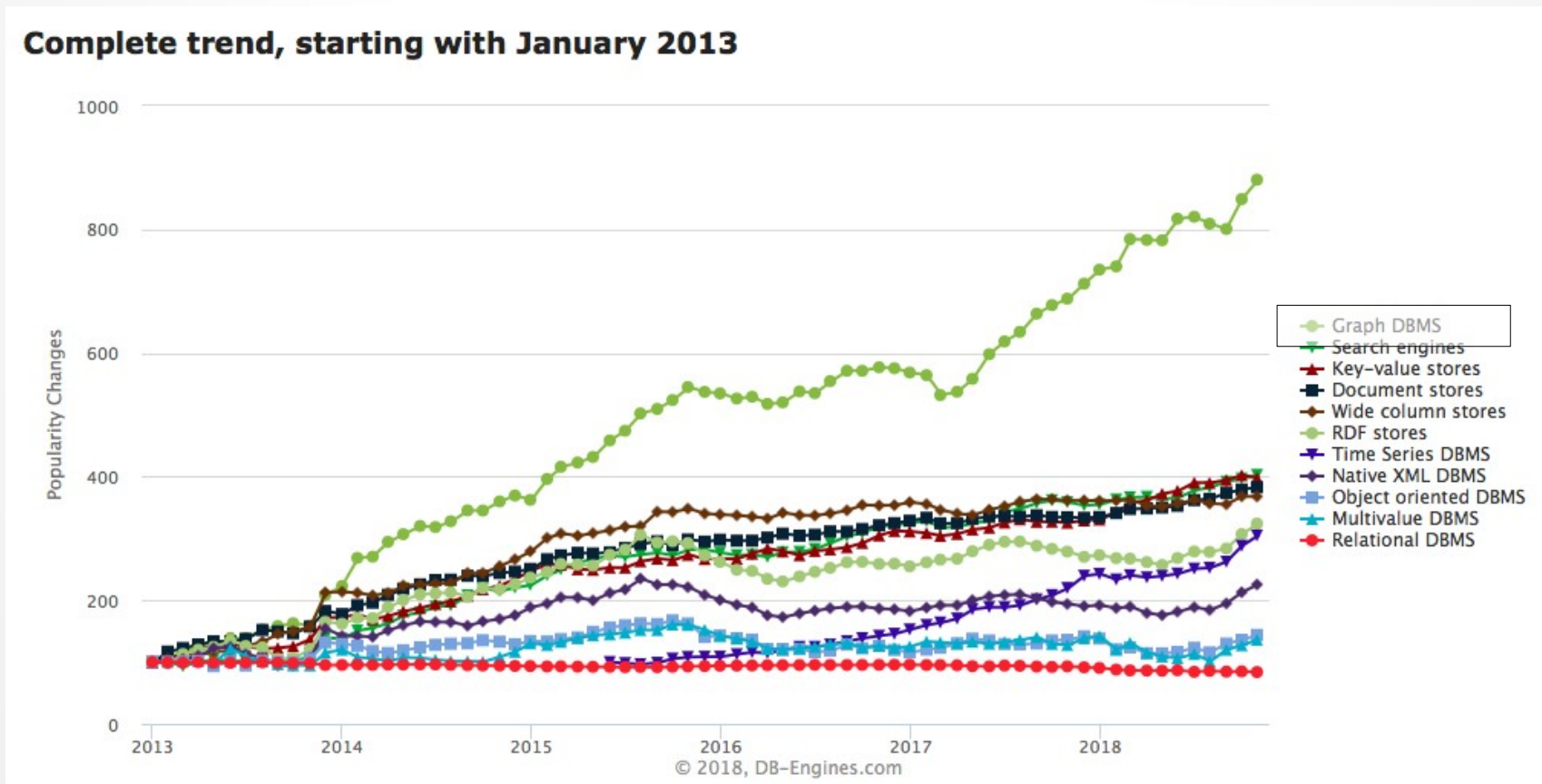
2-1: Trend (Complete ranking)

- DB-Engines Ranking
 - All DBMS



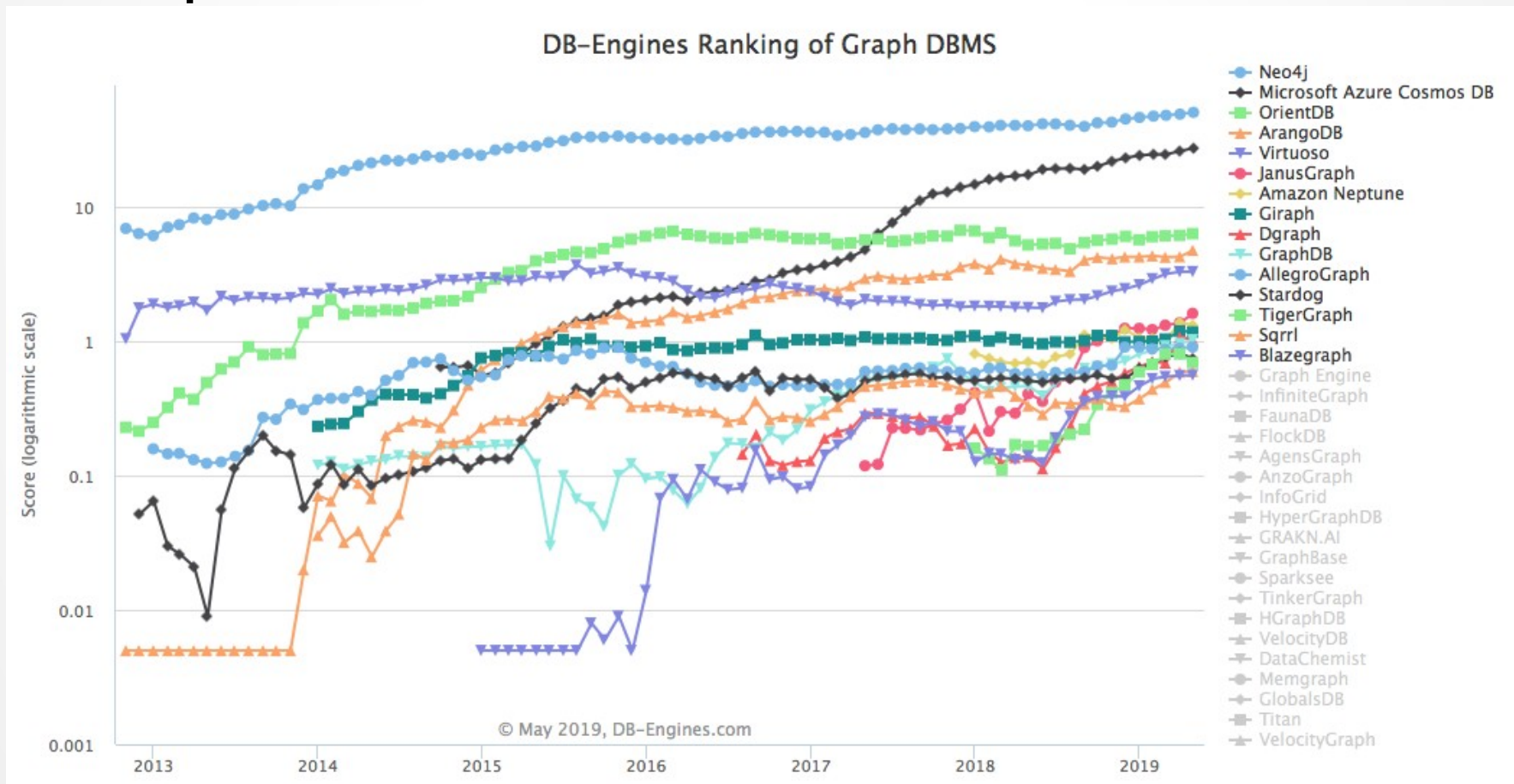
Trend (Ranking by database model)

- Complete Trend
 - Database Model



2-1: Trend (Graph DBMS)

- DB-Engines Ranking - Graph DBMS

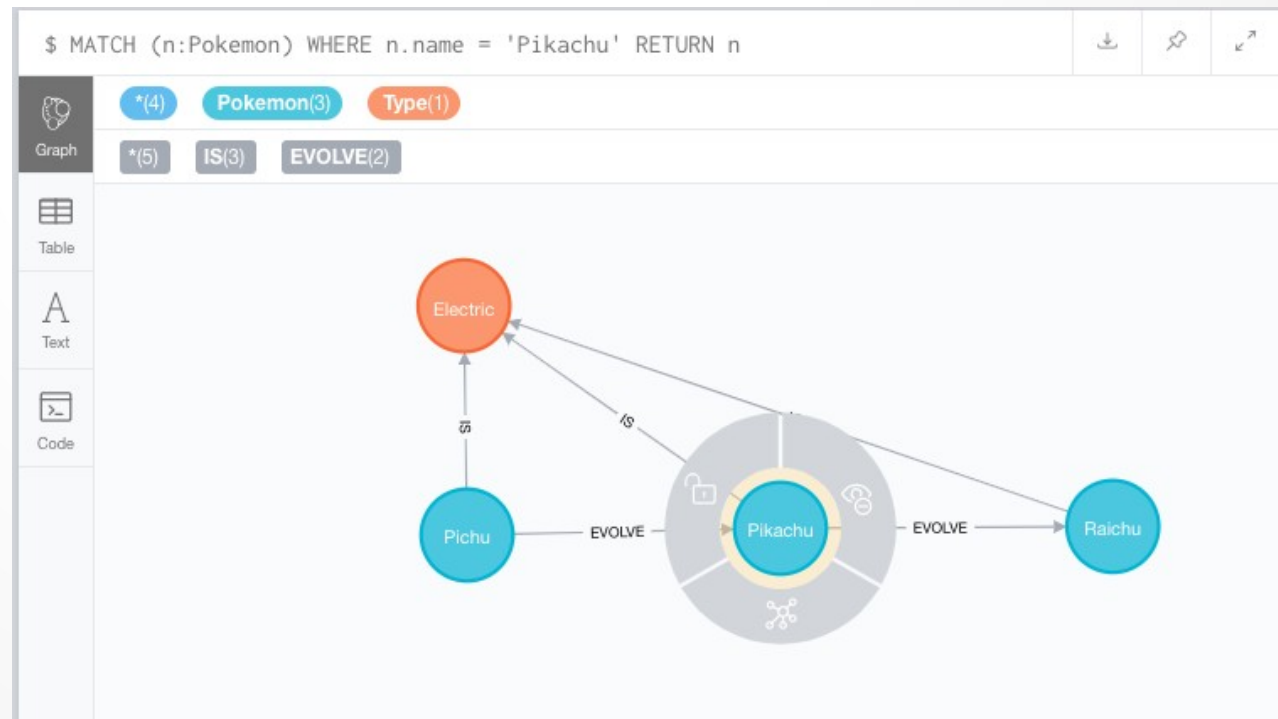


放送大学 教養学部

- 放送大学 教養学部 (The Open University of Japan)
- データベース ('17)
 - 14回、データベースの発展技術 (2)
 - グラフデータベース
 - リレーショナルモデルとは大きく異なるデータモデルを採用するグラフデータベースについて説明する。データモデルや問い合わせ言語について述べる。
- データ構造とプログラミング ('18)
 - 15回、グラフ
 - グラフのデータ構造について学習する。グラフに関する用語と意味、そして、コンピュータにおけるグラフの表現方法、深さ優先探索 (DFS)、幅優先探索 (BFS) 等のグラフ探索アルゴリズムについて学ぶ。

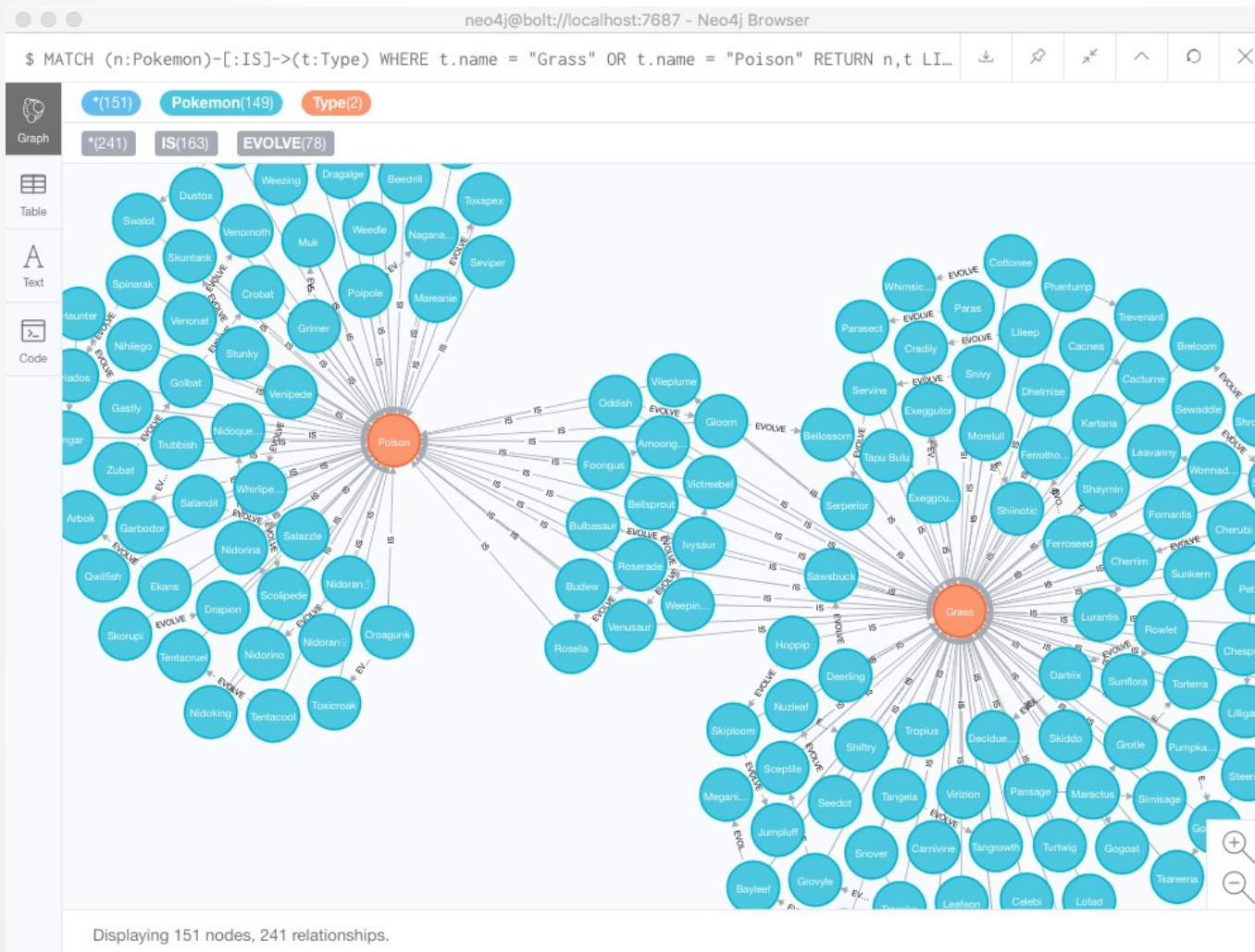
Pokemon

- ピカチュウ(Pikachu)の進化
 - でんきタイプ(Electric)
 - 進化
 - ピチュウ(Pichu)、ピカチュウ(Pikachu)、ライチュウ(Raichu)



Pokemon

- どくタイプ(Poison)、くさタイプ(Grass)

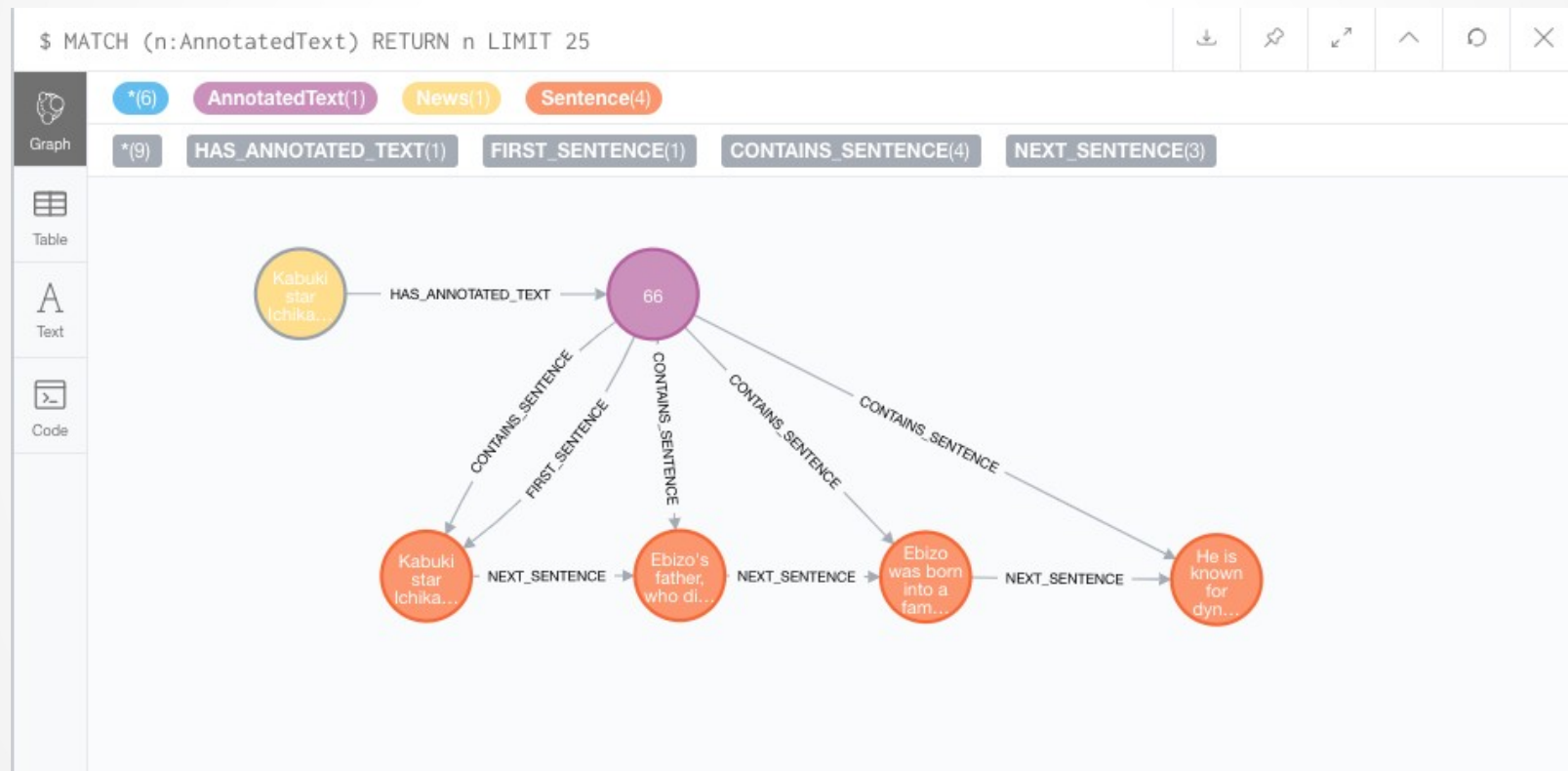


GraphのNLP

- GraphでNLP(自然言語処理)はどうするのか？
- テキストをノードに登録 (NHK World Japanのテキスト)
 - CREATE (n:News)
SET n.text = "Kabuki star Ichikawa Ebizo says he will assume his family's prestigious stage name of Danjuro next year.Ebizo told reporters at the Kabukiza theater in Tokyo on Monday that he will become the 13th actor to use the name Ichikawa Danjuro, beginning with a performance in May next year.The stage name has been handed down through generations since the 17th century. Ebizo's father, who died 6 years ago, was Ichikawa Danjuro the 12th.Ebizo was born into a famous family of kabuki actors. He is known for dynamic performances across a wide range of roles.Ebizo said his 5-year-old son, Kangen Horikoshi, will take the name of Ichikawa Shinnosuke the 8th.";

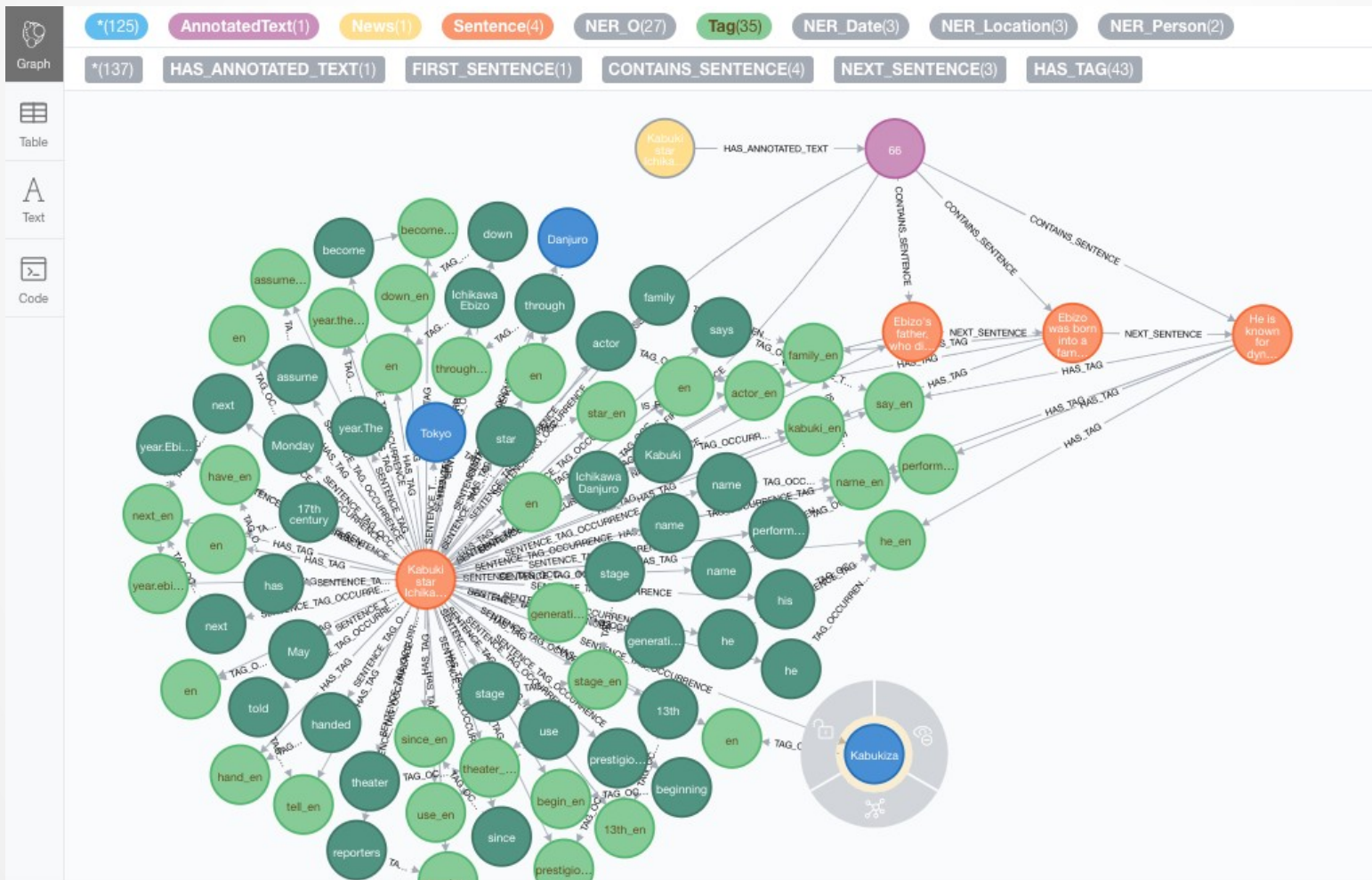
GraphのNLP

- 4つのセンテンスに別れている



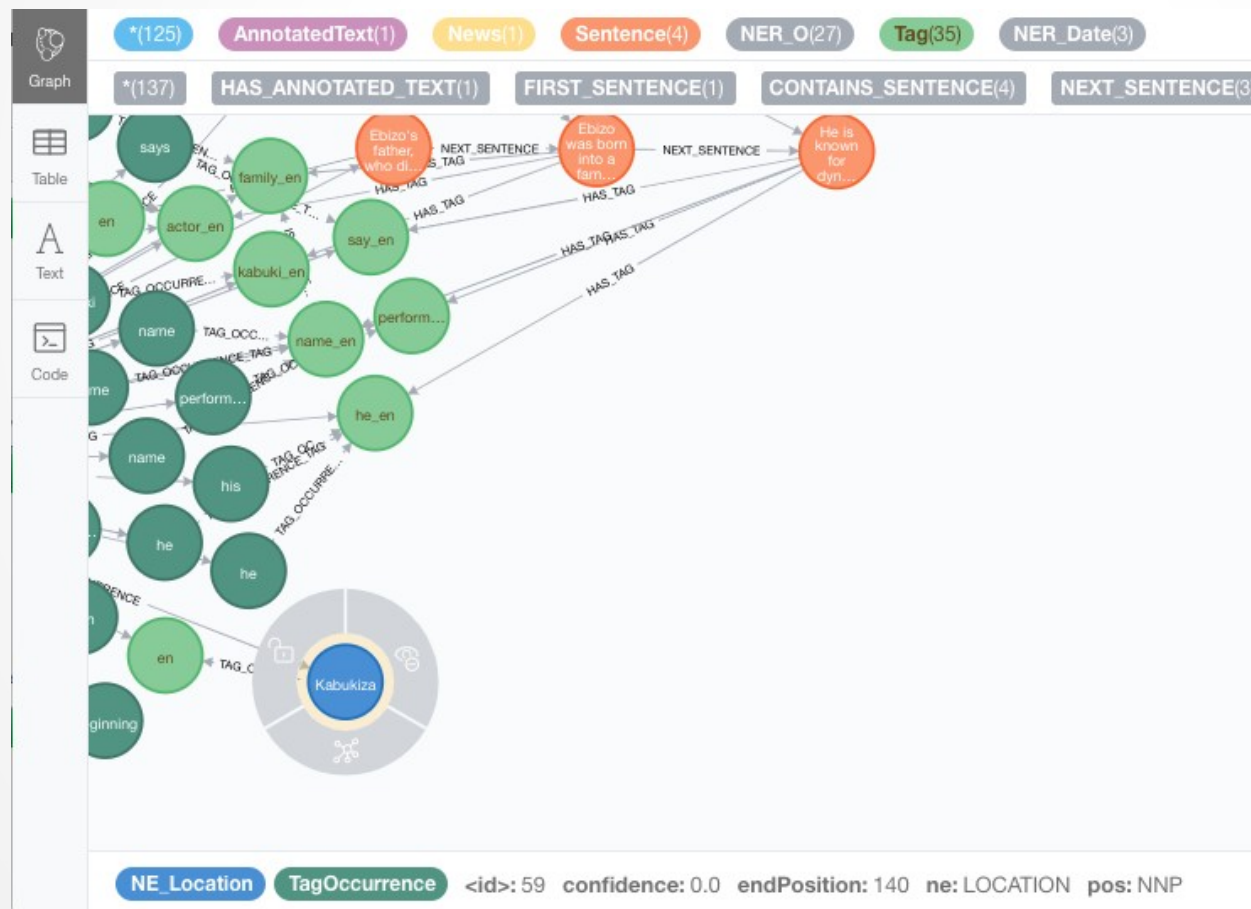
GraphのNLP

- 最初のセンテンスを見ると多くの単語が見える



歌舞伎座は場所

- KabukizaはLocationのラベルが付いている



翻訳を取得する

- LibreOffice master – UI / cui / messages.po

	A	B
1	English	Japanese
34	AutoCorrect	オートコレクト
35	Writing aids	文章作成支援
36	New Menu %n	新規メニュー %n
37	New Toolbar %n	新規ツールバー %n
38	Move Menu	メニューを移動
39	Add Submenu	サブメニューの追加
40	Submenu name	サブメニュー名
41	Are you sure to delete the image?	イメージを削除します。よろしいですか？
42	The icon %ICONNAME is already contained in, Would you like to replace the existing icon?	アイコン %ICONNAME はすでにイメージリストに含まれてい 既存のアイコンを置き換えますか？
43	Confirm Icon Replacement	アイコンの置換を確認
44	Yes to All	すべて「はい」
45	There are no more commands on the toolbar.	ツールバーにはもうコマンドがありません。ツールバーを消し
46	The menu configuration for %SAVE IN SELEC	%SAVE IN SELECTION% のメニュー設定は、初期設定にリ
47	The toolbar configuration for %SAVE IN SELEC	%SAVE IN SELECTION% のツールバー設定は、初期設定に

CalcからCSVに変換

- 漢字コードをUTF-8で書き込む
 - CalcはUTF-8が楽に扱えるのでExcelよりオススメ



```
libreofficepootle.csv
"English","Japanese"
"All","すべて"
"User-defined","ユーザー定義"
"Number","数"
"Percent","パーセント"
"Currency","通貨"
"Date","日付"
"Time","時刻"
"Scientific","指数表記"
"Fraction","分数"
"Boolean Value","ブール値"
"Text","テキスト"
"Configuration","設定"
"My Documents","マイドキュメント"
"Images","イメージ"
"Icons","アイコン"
"Palettes","パレット"
"Backups","バックアップ"
```

CSVのデータをNeo4jに

- Cypherはこんな感じ
 - CREATE (:En {text: line.English})-[:TRANSLATE]->(:Ja {text: line.Japanese})

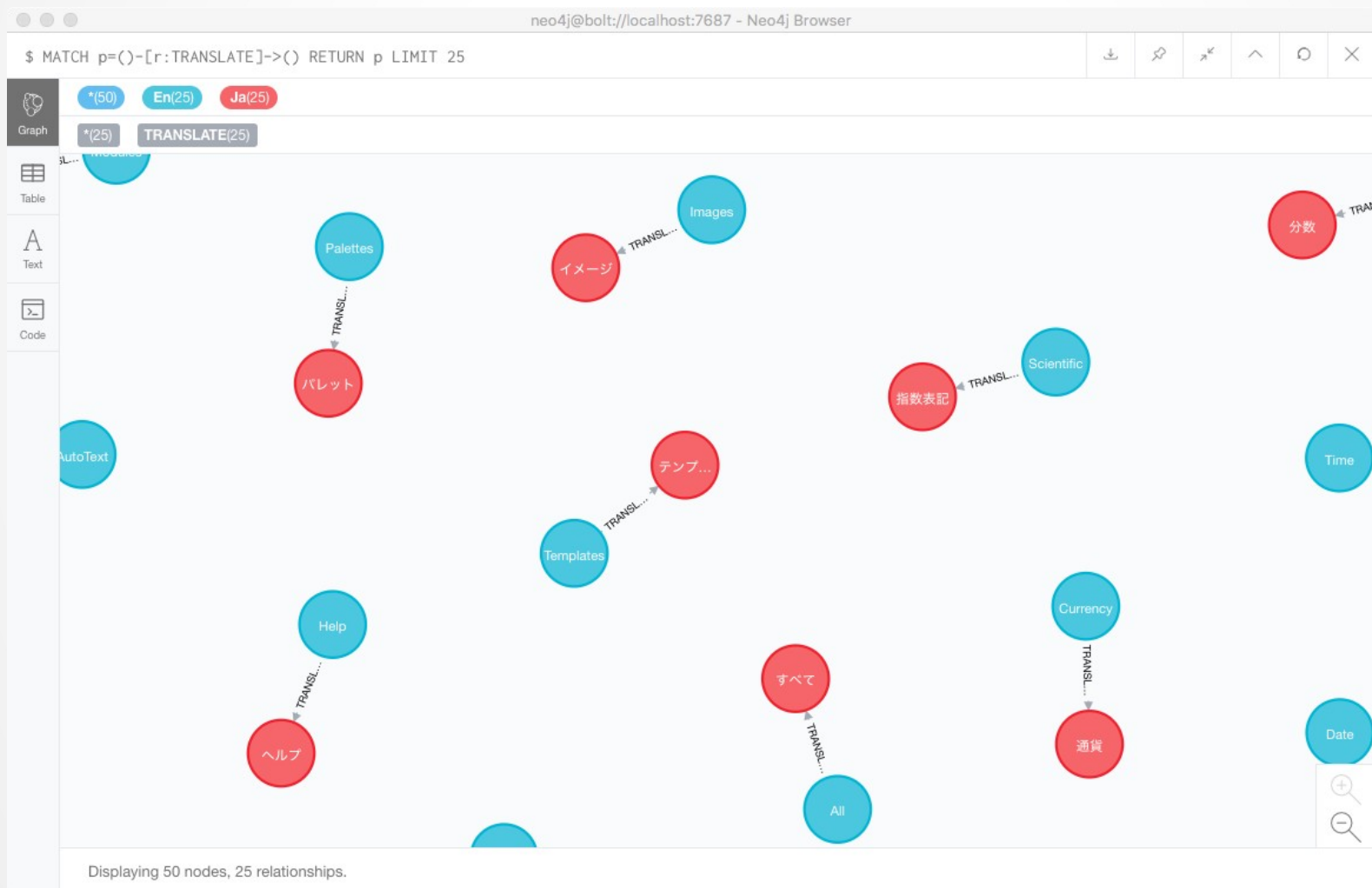
```
1 USING PERIODIC COMMIT 500
2 LOAD CSV WITH HEADERS FROM 'file:///libreofficepootle.csv' AS line
3 CREATE (:En {text: line.English})-[:TRANSLATE]->(:Ja {text: line.Japanese})
```

```
$ USING PERIODIC COMMIT 500 LOAD CSV WITH HEADERS FROM 'file:///libreofficepootle.csv' AS line CREATE ...
```

Added 432 labels, created 432 nodes, set 432 properties, created 216 relationships, completed after 1253 ms.

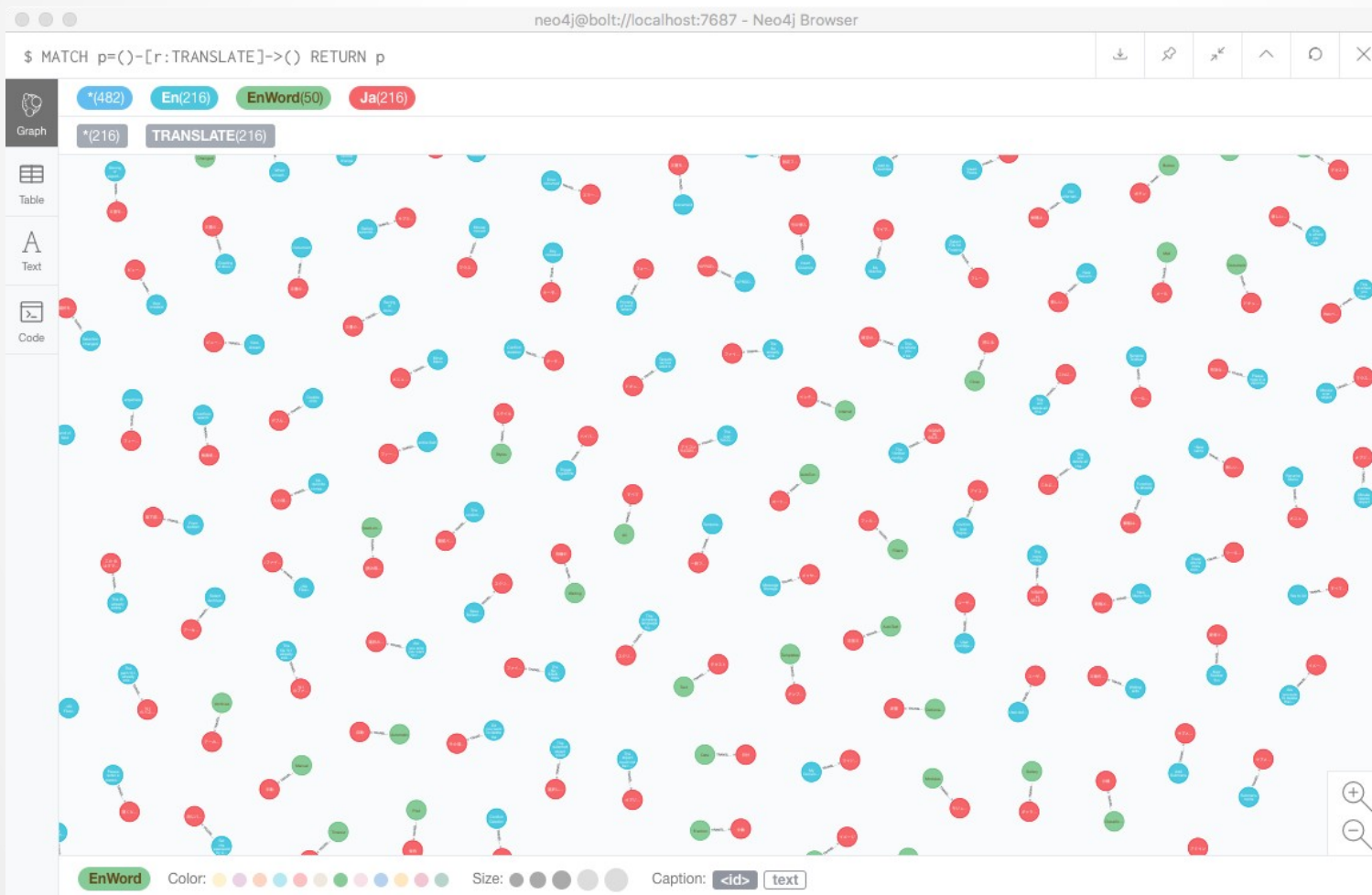
表示してみる

- 英語と日本語の関係が登録されている



単語を見分ける

- 登録されている英語には単語と文章があるので単語に “:EnWord” のラベルをつける



英単語を使っている文章を探す

- e.text CONTAINS w.text

The screenshot shows the Neo4j Browser interface. At the top, the browser title is "neo4j@bolt://localhost:7687 - Neo4j Browser". The query editor contains the following Cypher query:

```
1 MATCH (e:En),(w:EnWord)
2 WHERE e.text CONTAINS w.text
3 AND NOT e:EnWord
4 RETURN e.text,w.text
```

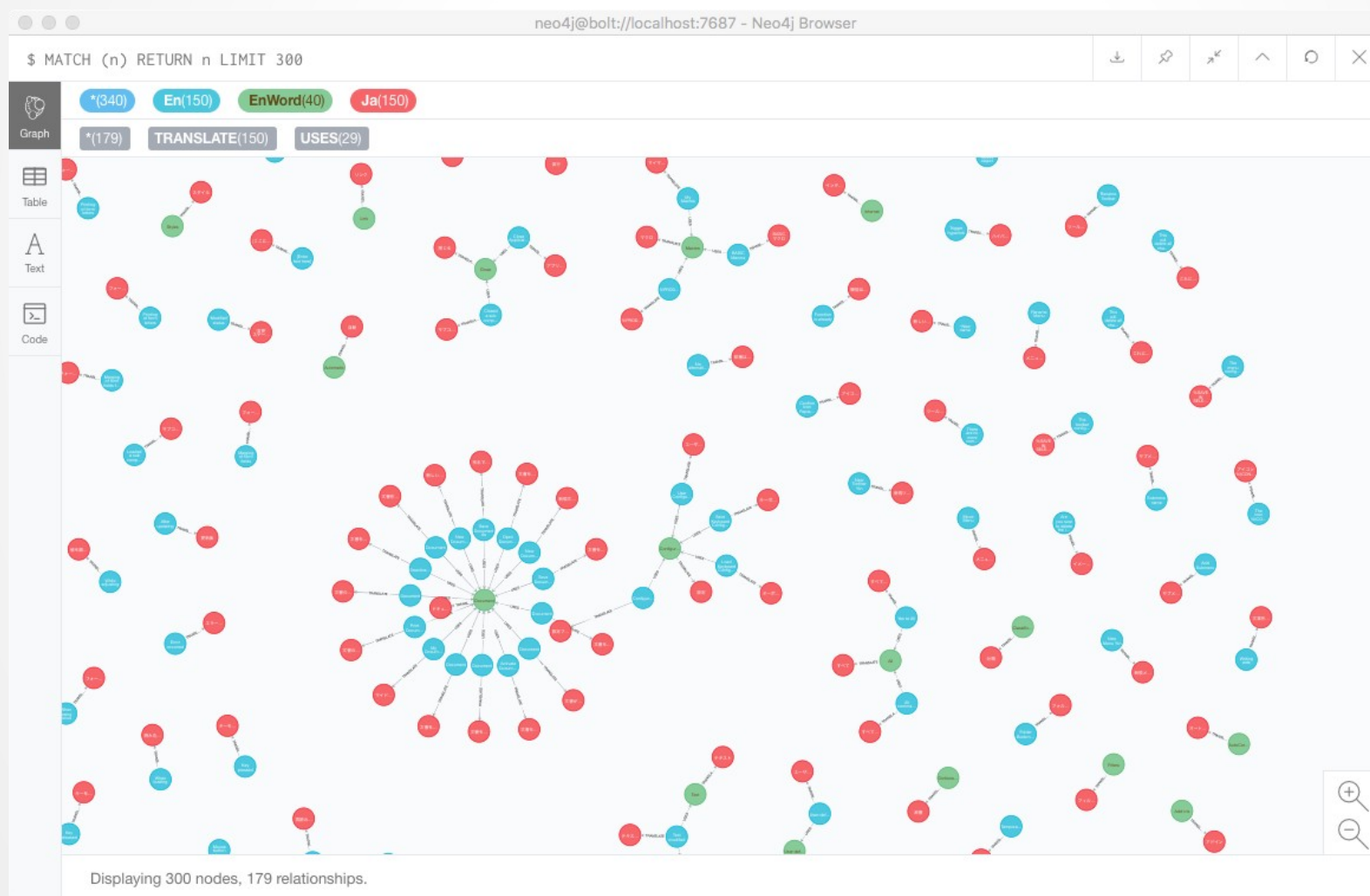
Below the query editor, the query is executed, and the results are displayed in a table view. The table has two columns: "e.text" and "w.text". The results are as follows:

e.text	w.text
"My Documents"	"Document"
"User Configuration"	"Configuration"
"User-defined dictionaries"	"User-defined"
"Yes to All"	"All"
"Yes to All"	"Yes"
"New Document"	"Document"
"No alternatives found."	"No"
"All commands"	"All"
"My Macros"	"Macros"
"%PRODUCTNAME Macros"	"Macros"
"Load Keyboard Configuration"	"Configuration"
"Save Keyboard Configuration"	"Configuration"
"Configuration (*.cfg)"	"Configuration"
"BASIC Macros"	"Macros"
"Close Application"	"Close"
"New Document"	"Document"

At the bottom of the table, a status message reads: "Started streaming 38 records after 8 ms and completed after 361 ms."

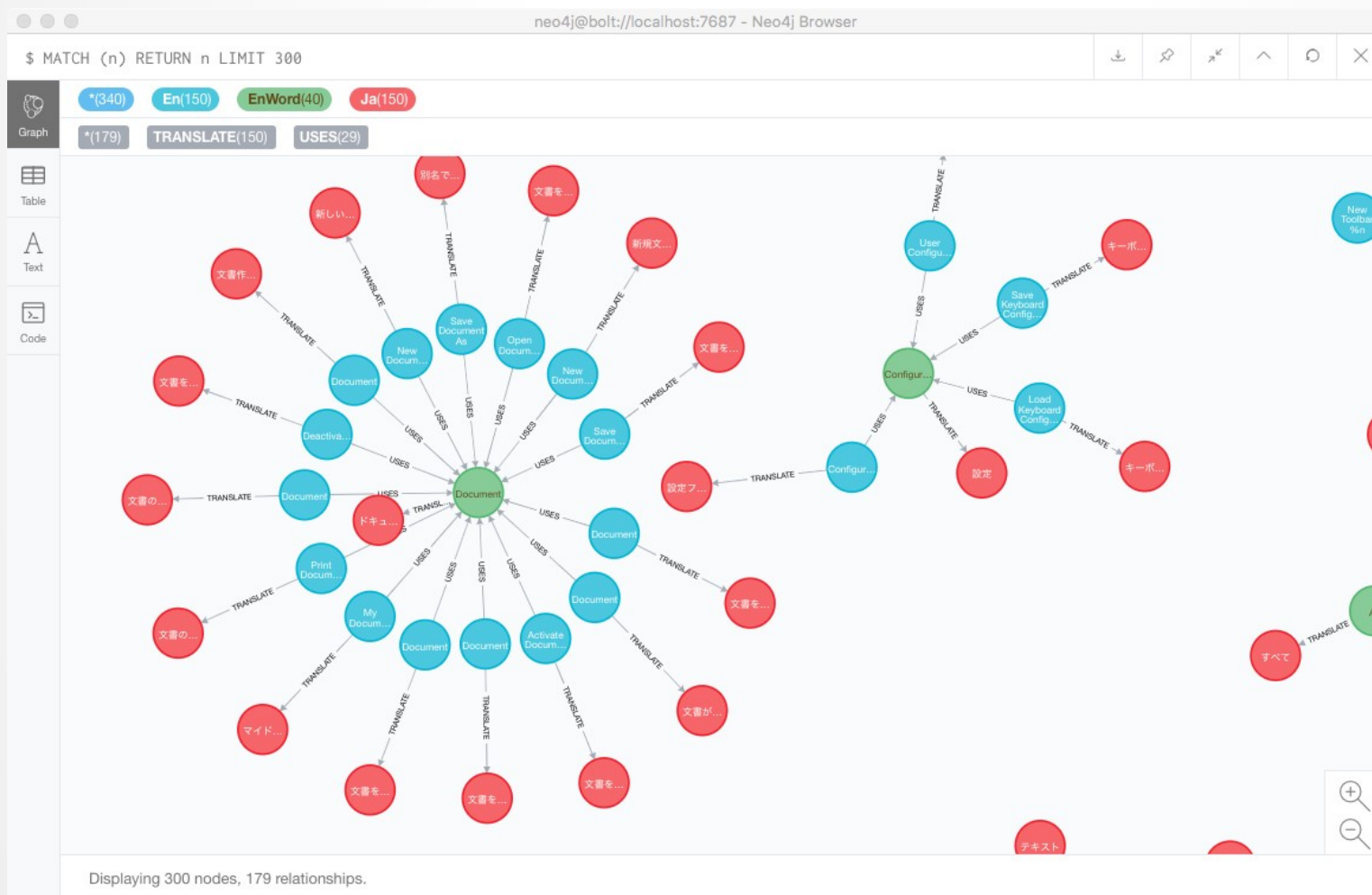
英単語と使っている文章をつなぐ

- (:En)-[:USES]->(:EnWord)



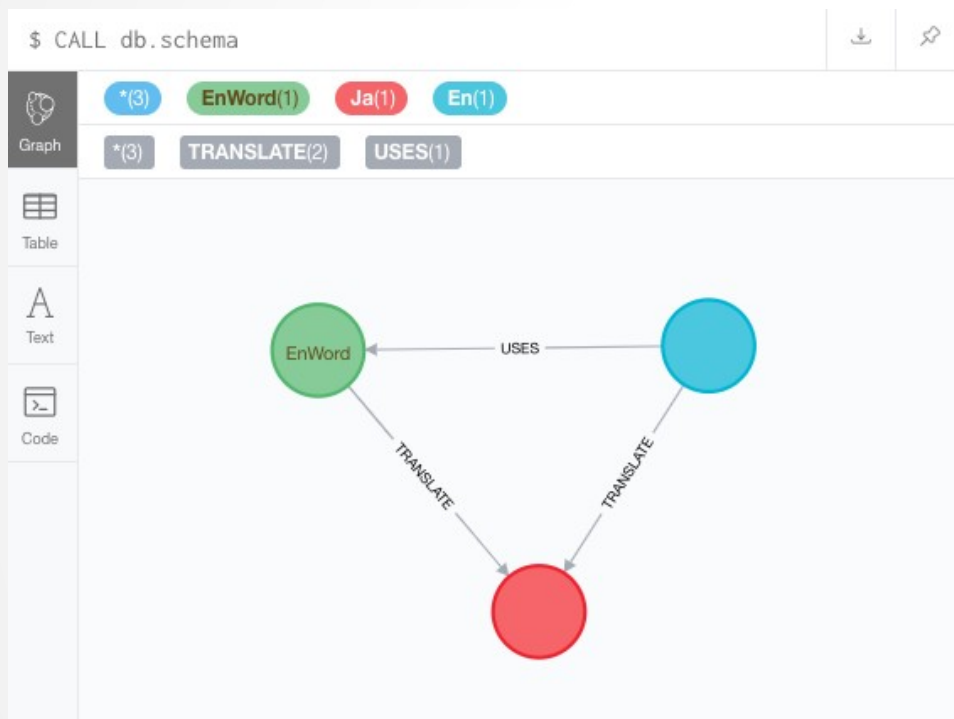
英単語と使っている文章をつなぐ

- (:En)-[:USES]->(:EnWord)



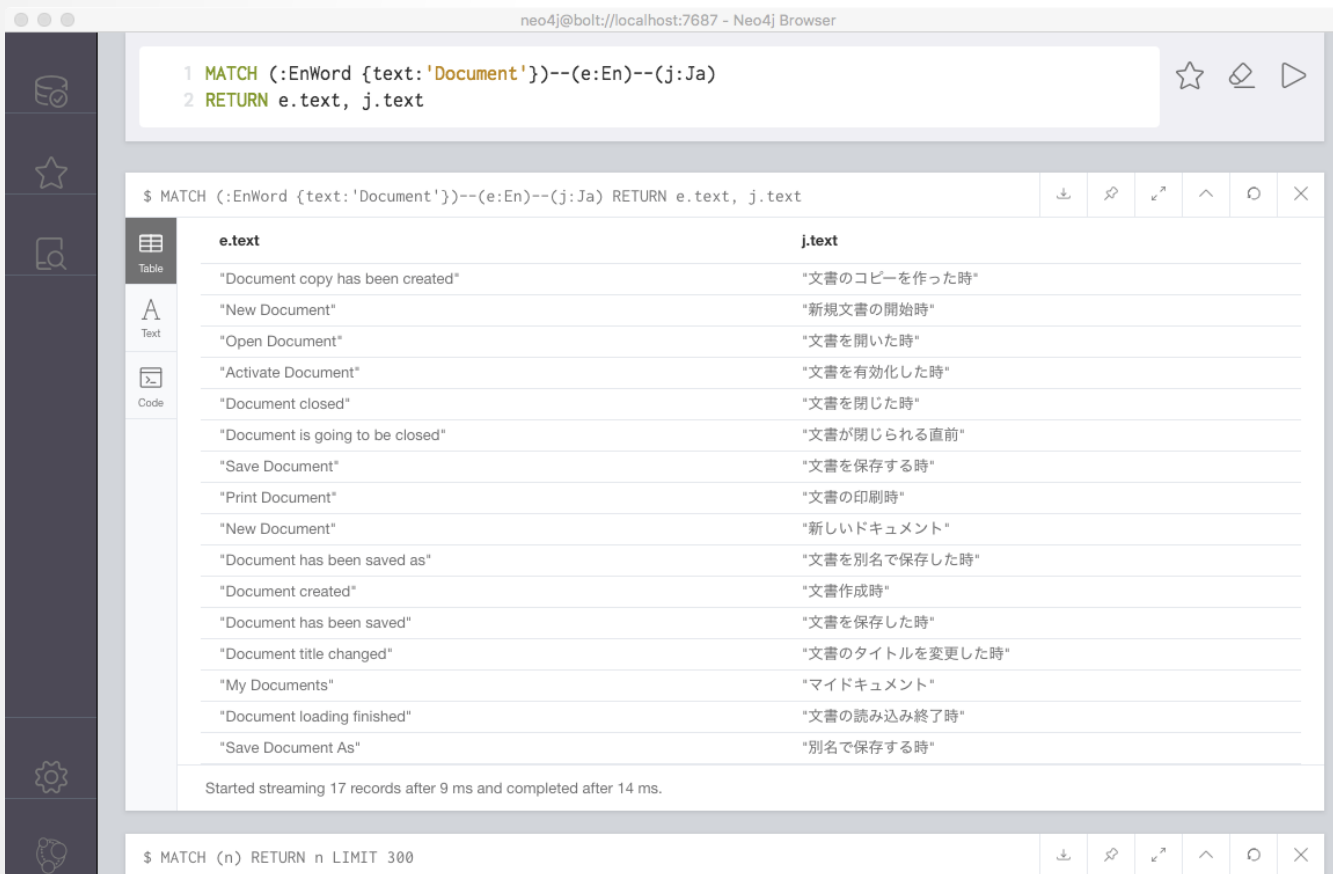
スキーマ

- CALL db.schema



Document と言う単語を見ると

- 同じ Document の英単語
 - `MATCH (:EnWord {text:'Document'})--(e:En)--(j:Ja)`
`RETURN e.text, j.text`



The screenshot shows the Neo4j Browser interface. At the top, the browser title is "neo4j@bolt://localhost:7687 - Neo4j Browser". The query editor contains the following Cypher query:

```
1 MATCH (:EnWord {text:'Document'})--(e:En)--(j:Ja)
2 RETURN e.text, j.text
```

The results are displayed in a table with two columns: "e.text" and "j.text". The table contains 17 rows of data, each representing a document-related action in English and its corresponding Japanese translation.

e.text	j.text
"Document copy has been created"	"文書のコピーを作った時"
"New Document"	"新規文書の開始時"
"Open Document"	"文書を開いた時"
"Activate Document"	"文書を有効化した時"
"Document closed"	"文書を閉じた時"
"Document is going to be closed"	"文書が閉じられる直前"
"Save Document"	"文書を保存する時"
"Print Document"	"文書の印刷時"
"New Document"	"新しいドキュメント"
"Document has been saved as"	"文書を別名で保存した時"
"Document created"	"文書作成時"
"Document has been saved"	"文書を保存した時"
"Document title changed"	"文書のタイトルを変更した時"
"My Documents"	"マイドキュメント"
"Document loading finished"	"文書の読み込み終了時"
"Save Document As"	"別名で保存する時"

At the bottom of the interface, the status bar shows: "\$ MATCH (n) RETURN n LIMIT 300".

Document

- 同じDocumentの英単語
 - 日本語の翻訳は3種類
 - 文書
 - "Document copy has been created", "文書のコピーを作った時"
 - "New Document", "新規文書の開始時"
 - ドキュメント
 - "New Document", "新しいドキュメント"
 - "My Documents", "マイドキュメント"
 - 使っていない
 - "Save Document As", "別名で保存する時"

翻訳のゆらぎを考える

- ゆらぎを考える

- 「**文書**」に統一するとどうなるか？

- "新しいドキュメント"は、"新しい**文書**"

- "マイドキュメント"は、"マイ**文書**" (これは微妙)

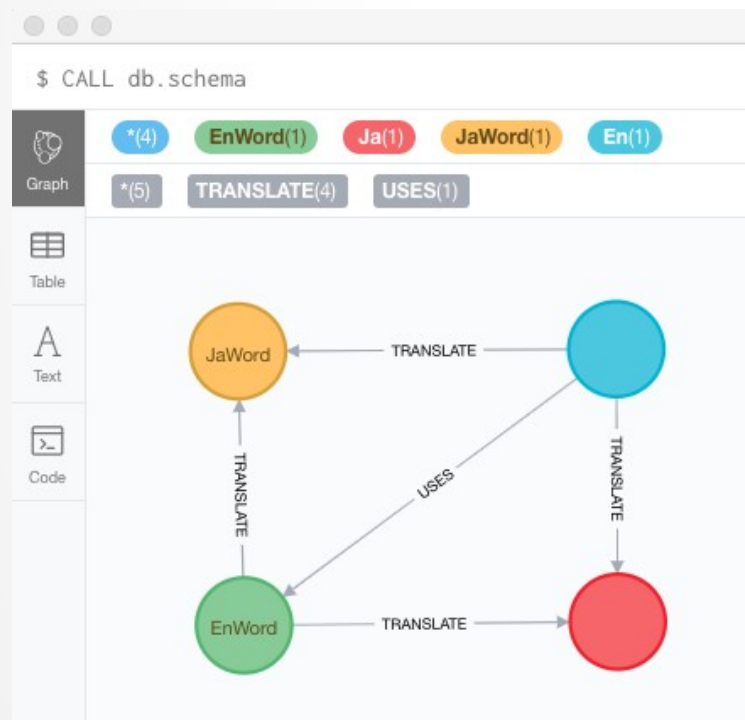
- 「**ドキュメント**」に統一するとどうなるか？

- "文書のコピーを作った時"は、"**ドキュメント**のコピーを作った時"

- "新規文書の開始時"は、"新規**ドキュメント**の開始時"

日本語も単語と文章に分ける

- MATCH (:EnWord)-[:TRANSLATE]->(j:Ja)
SET j:JaWord



日本語の単語と文章もつなく

- (:Ja)-[:USES]->(:JaWord)

The screenshot shows the Neo4j Browser interface. At the top, the browser title is "neo4j@bolt://localhost:7687 - Neo4j Browser". The query editor contains the following Cypher query:

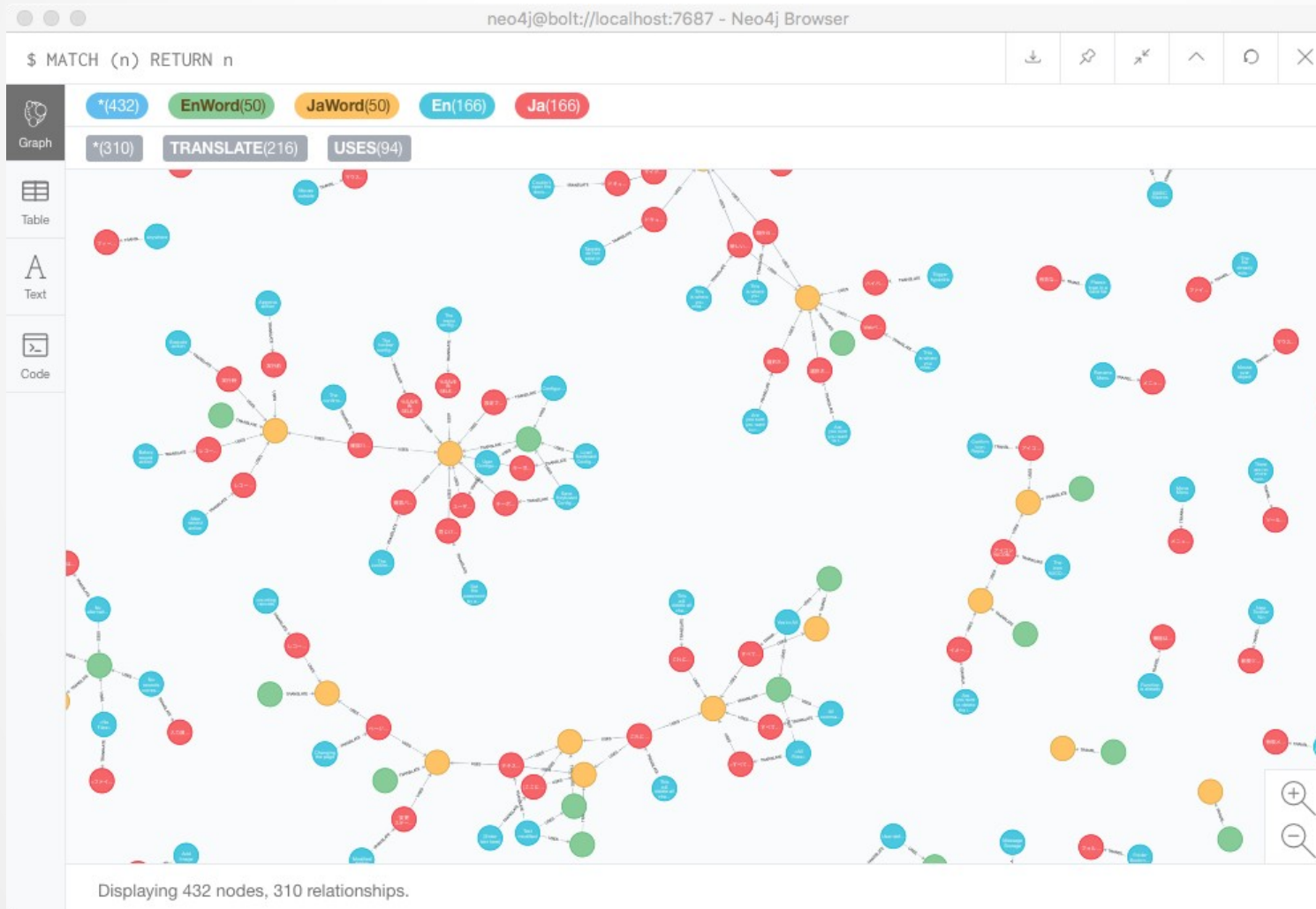
```
1 MATCH (e:Ja),(w:JaWord)
2 WHERE e.text CONTAINS w.text
3 RETURN e.text, w.text
```

Below the query editor, the query is repeated: "\$ MATCH (e:Ja),(w:JaWord) WHERE e.text CONTAINS w.text RETURN e.text, w.text". The results are displayed in a table view with two columns: "e.text" and "w.text".

e.text	w.text
"すべて「はい」"	"すべて"
"これにより、このツールバーに以前加えられたすべての変更が削除されます。ツールバーをリセットしますか?"	"すべて"
"これにより、このコンテキストメニューに以前加えられたすべての変更が削除されます。コンテキストメニューをリセットしますか?"	"すべて"
"すべてのコマンド"	"すべて"
"<すべてのファイル>"	"すべて"
"ユーザー定義辞書"	"ユーザー定義"
"ページ数の変更時"	"数"
"レコードを数える"	"数"
"これにより、このコンテキストメニューに以前加えられたすべての変更が削除されます。コンテキストメニューをリセットしますか?"	"テキスト"
"[ここにテキストを入力]"	"テキスト"
"テキストの変更時"	"テキスト"
"ユーザー設定"	"設定"

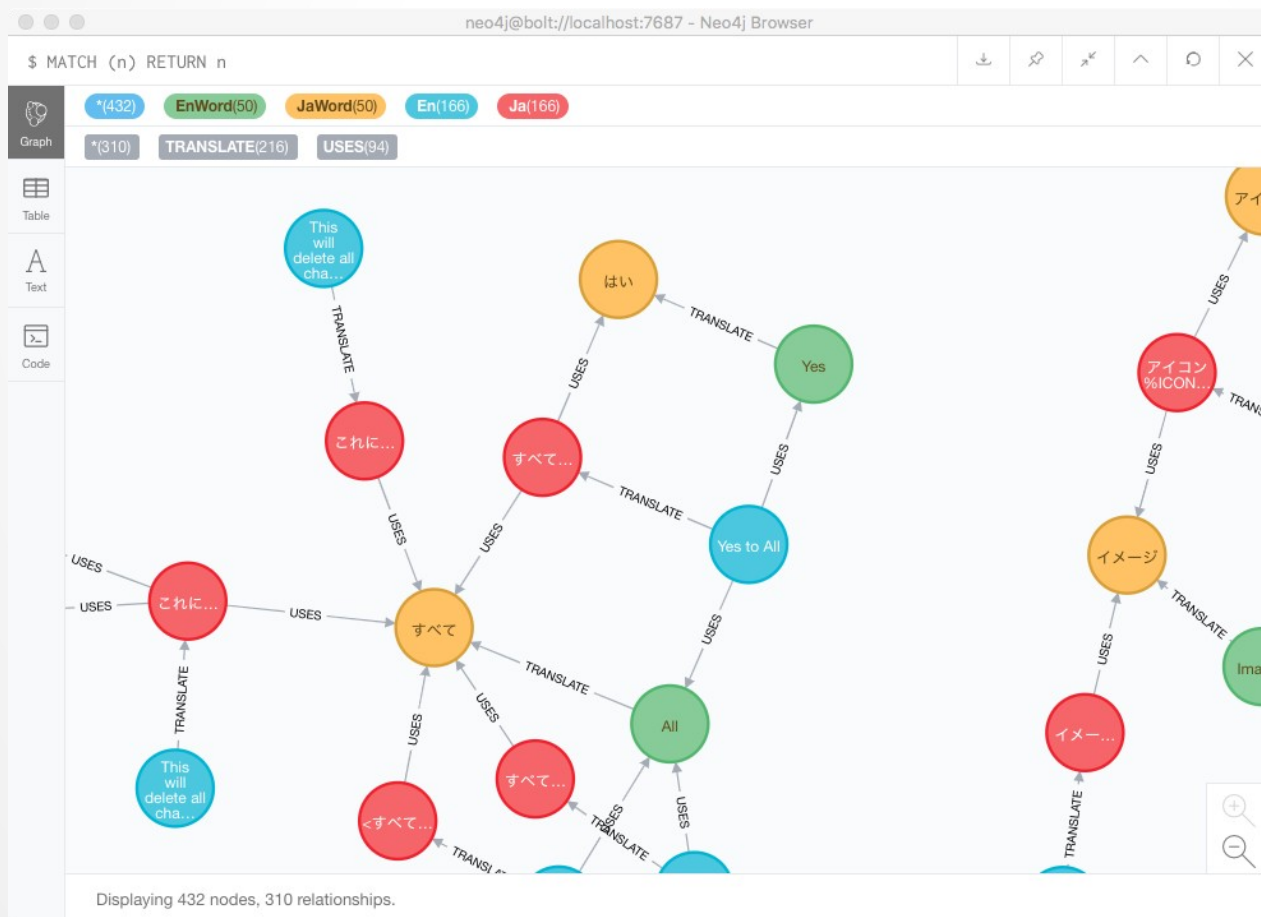
日本語と英語の関係

- 面白そうなグラフになりました



同じ単語が日本語と英語で使用





- 四角にループしているところは同じ訳語
 - (英語文書)--(英単語)--(日本語単語)--(日本語文書)



単語が整合性があるもの

neo4j@bolt://localhost:7687 - Neo4j Browser

```
1 MATCH (e:En)--(ew:EnWord)--(jw:JaWord)--(j:Ja)--(e)
2 RETURN e.text,ew.text,jw.text,j.text
```

\$ MATCH (e:En)--(ew:EnWord)--(jw:JaWord)--(j:Ja)--(e) RETURN e.text,ew.text...      

	e.text	ew.text	jw.text	j.text
Table	"Yes to All"	"All"	"すべて"	"すべて「はい」"
Text	"<All Files>"	"All"	"すべて"	"<すべてのファイル>"
	"All commands"	"All"	"すべて"	"すべてのコマンド"
Code	"User-defined dictionaries"	"User-defined"	"ユーザー定義"	"ユーザー定義辞書"
	"Text modified"	"Text"	"テキスト"	"テキストの変更時"
	"Configuration (*.cfg)"	"Configuration"	"設定"	"設定ファイル (*.cfg)"
	"User Configuration"	"Configuration"	"設定"	"ユーザー設定"
	"Load Keyboard Configuration"	"Configuration"	"設定"	"キーボード設定を読み込む"
	"Save Keyboard Configuration"	"Configuration"	"設定"	"キーボード設定を保存"
	"New Document"	"Document"	"ドキュメント"	"新しいドキュメント"
	"My Documents"	"Document"	"ドキュメント"	"マイドキュメント"
	"Text modified"	"Text"	"テキスト"	"テキストの変更時"
	"BASIC Macros"	"Macros"	"マクロ"	"BASIC マクロ"
	"My Macros"	"Macros"	"マクロ"	"マイマクロ"
	"%PRODUCTNAME Macros"	"Macros"	"マクロ"	"%PRODUCTNAME のマクロ"
	"Select Archives"	"Archives"	"アーカイブ"	"アーカイブの選択"

Started streaming 17 records after 2 ms and completed after 14 ms.

Demo

- デモを行います

今後の予定 (What's Next?)

- データ量
 - 全てのpoデータを使う
- 熟語・慣用句
 - 単語だけでなく熟語・慣用句も抽出
- 時間
 - 翻訳の変更が多いもの
 - 翻訳が古いままのもの
- 多言語
 - 中国語
 - 韓国語
 - etc

ありがとうございました

- 何か質問がありましたら
If you have any questions?
 - koji@annoura.com