ヨーロッパにおける日本語学習者の日本語作文のテキストマイニング

―その言語的特徴に関するパイロットスタディー―

2017年5月24日（水）

語学研究所　定例研究会

東京外国語大学大学院国際日本学研究院

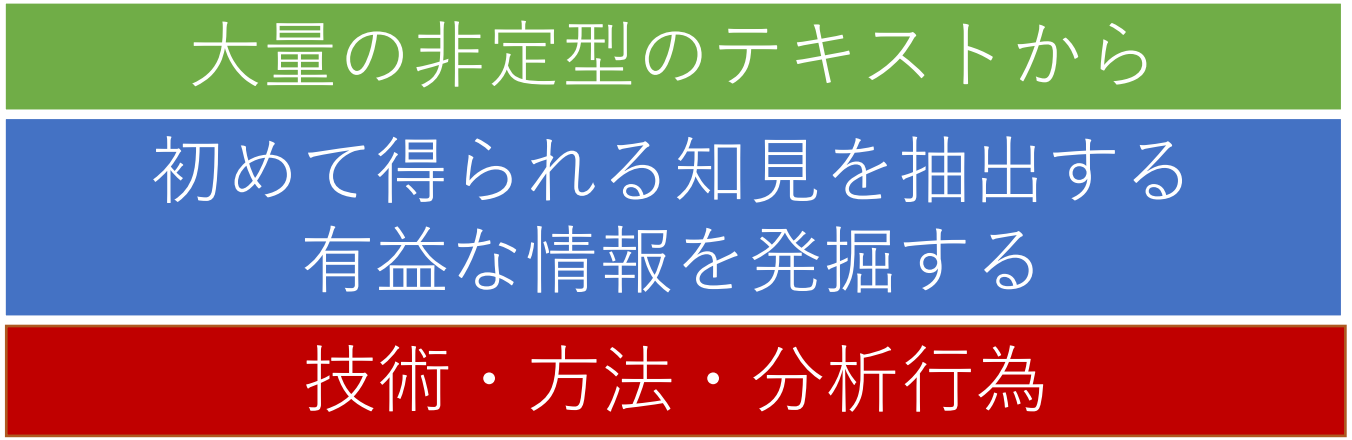
准教授　阿部 新

１．目的

- ヨーロッパにおける日本語学習者の日本語作文の言語的特徴を
- テキストマイニングによって探索する

1．目的：テキストマイニングによる探索

- テキストマイニングとは（阿部2014）
  - 複数の文書データの内容を総合的にとらえることで初めて得られる知見を抽出するための内容分析の技術（那須川2006: 1）
  - コンピュータを使って大量のテキストの中から有益な情報を探し出す技術（石田2008: 1）
  - 定型化されていない文章（テキスト）から有益な情報を発掘（マイニング）するための方法論および分析行為（内田2010: 102）



1．目的：テキストマイニングによる探索

- コーパス言語学との比較（阿部2014）

	コーパス言語学（石川2013）	**テキストマイニング（那須川2006）**
データ	現実の言語から網羅的・代表的に収集されたテキストデータ	**分析したい内容を含む**ように収集されたテキストデータ
目的	言語の記述 言語に関する仮説の検証	初めて得られる**知見の探索** 有益な**情報の探索**
技術	主に検索	**検索・分類整理・分析**

1．目的：テキストマイニングによる探索

- テキストマイニングの適用分野の広がり
  - 定量的調査の調査環境の悪化（サンプリング実施難，回収率の悪さ）
    →定性的調査への関心の高まり（大隅・Lubart 2000:340）

  - マーケティング分野での活用事例（那須川2006:66-68）
    - コールセンターの文字化データ：消費者からの苦情・質問・要望
    - 営業活動報告・アンケート調査の自由回答：企業から消費者への評判把握

２．先行研究

- 人文科学分野・社会科学分野での活用事例
  - 石田2008
  - 松村・三浦2014
  - 岸江・阿部・石田・西尾2011
  - 岸江・田畑編2014

  - アンケート自由回答の分析
  - 電子掲示板，ブログの口コミ情報，メーリングリスト，議事録の分析
  - 新聞記事の自動分類
  - 文学作品の書き手判別

２．先行研究（アンケート自由回答分析）

- アンケート自由回答の分析（阿部2014）
  - 言語表現の自然さに関するアンケートの自由回答を分析

- 評価基準の探索（森2016）
  - 評価コメントから評価基準を探索的に選び出す

- 方向性の探索（阿部・嵐・須藤2016）
  - 音声教育に対するコメント（アンケートの自由回答）から，教師の特徴と教育に対する考え方を探索的に取り出し，今後の音声教育の方向性についての示唆を得る

２．先行研究（第二言語としての日本語産出データ）

- 森ほか（2012）
  - 外国人集住地域の小学校に通う「外国にルーツを持つ児童」（外国人児童＝80％は日本で生育）と日本人児童の日本語作文（複数のテーマ）を収集し，両者をテキストマイニングによって分析し，両者の作文の特性を探った。
  - 語彙数，出現回数の分布，出現上位語の分析，品詞別・作文テーマ別の出現語彙の傾向，共起ネットワーク分析等を行った。

2．先行研究（第二言語としての日本語産出データ）

- 森ほか（2012）
  - 語彙量や品詞分布では日本人児童と外国人児童では差がないが，語彙の豊かさには差があった。
  - 語の選択や表記には違いが見られ，語同士の関連付けにも違いがあった。

２．先行研究（第二言語としての日本語産出データ）

- 松田ほか（2013）
  - ACTFL(The American Council on the Teaching of Foreign Languages)によるOPI(Oral Proficiency Interview)によるレベル判定で超級レベルと判定された日本語学習者の口頭産出の談話的特徴を，上級レベル学習者と比較して，テキストマイニングによって探った。
  - KYコーパスというレベル別の学習者OPIが収録されたコーパスのデータを使用した。（中・英・韓の超級話者3名分，上級話者3名分）

2．先行研究（第二言語としての日本語産出データ）

- 松田ほか（2013）
  - 結果として，超級話者は，談話の結束性を示す表現，聞き手配慮に関わる表現を上級話者より使うことが分かった。
    - 文脈指示の「コソア」のうち「コ」系表現（複数段落における結束性を高める）
    - 発話緩和や発話の埋め合わせ機能を持つ「ね」
    - 多様なフィラー（あの，あのー，まあ，なんか，もう，いや，と，その，こう）
    - 言葉を選んでいることを示す「試行的提示」（「ていうか」「ていうんですか」）

３．本研究のデータ

- 科研「**日本語ライティング評価の支援ツール開発：「人間」と「機械」による評価の統合的活用**」（研究課題番号：26284074）

- 2014年11月～2016年3月
- 11か国の日本語学習者
  - クロアチア，フランス，ドイツ，ハンガリー，イタリア，オランダ，セルビア，スロベニア，スペイン，ロシア，アメリカ
- 国籍は16種類
  - ベルギー，クロアチア，フランス，ドイツ，香港，ハンガリー，イタリア，韓国，ポーランド，ロシア，セルビア，スロベニア，スペイン，オランダ，アメリカ，ベトナム

３．本研究のデータ

- 5種類のプロンプト
  - 英語，スペイン語，フランス語，ハンガリー語で準備
  - 辞書使用可，1時間程度で書くように指示。

  - A1/A2：比較対照の説明・論証→**こちらを分析（A1:136編，A2:153編）**
  - B1/B2/B3：ナラティブ（留学生にあなたの街を紹介，留学生に居住方法の説明，「忘れられない出来事」）

  - Aから最低1つ，Bから最低1つ選んで書く（一人が最低2つの作文を書く）

3．本研究のデータ

- A1　あなたは以下の作文コンテストのポスターを見ました。そして，この作文コンクールに応募することにしました。

「個人旅行」と「パック旅行」

知り合いのいない国を 1 週間旅行するとしたら，個人で準備する「個人旅行」と，旅行会社が準備してくれる「パック旅行」と，どちらで行きますか。

それぞれのプラス面とマイナス面を挙げて比較し，旅行についてのあなたの意見を 600 字～800 字で書いてください。

入賞された方には，日本への往復航空券（１カ月有効）をプレゼントいたします。

日本さくら旅行

3．本研究のデータ

- A2　あなたは以下の作文コンテストのポスターを見ました。そして，この作文コンクールに応募することにしました。

あなたは「外食派」？ それとも「自炊派」？

「外食」と「自炊」，それぞれのプラス面とマイナス面を挙げて比較し，「食生活」についてのあなたの意見を 600 字～800 字で書いてください。

応募者の中から抽選で 20 名様に，弊社のレストラン★★のランチ券（2 名様分）または弊社の自炊グッズ（フライパンと鍋）を差し上げます。

★★食品会社マーケティング部

3．本研究のデータ

国・プロンプト	A-1	A-2	A合計	B-1	B-2	B-3	B合計	計
ベルギー	14	17	31	16	1	12	29	60
クロアチア		4	4	3		1	4	8
フランス	16	16	32	16	3	13	32	64
ドイツ	12	10	22	15	4	3	22	44
香港	1		1				0	1
ハンガリー	35	37	72	34	6	32	72	144
イタリア	2	5	7	6			6	13
韓国	1	1	2	1		1	2	4
ポーランド	1		1	1			1	2
ロシア	14	10	24	8	2	14	24	48
セルビア	6	5	11	9		2	11	22
スロベニア	3	9	12	9		3	12	24
スペイン	12	14	26	22	3	2	27	53
オランダ		1	1	1			1	2
アメリカ	18	24	42	22	14	8	44	86
ベトナム	1		1			1	1	2
計	136	153	289	163	33	92	288	577

３．本研究のデータ

SPOT：自然な話速度の読み上げ文を聞き，解答用紙に書かれた同文のひらがな1文字分1か所の空欄（文法項目）を穴埋め（選択式）するテスト。（小林2015）30問×3部＝90問

例：隣の人（　）教えてもらったんです。

- 執筆前に個人情報
  - 年齢，性別，専攻分野，国籍，母語，移住経験，言語4技能の流暢さ，日本での学習歴，日本語能力試験受験経験，日本語学習歴，日本語の試験の受験経験，日本語能力（SPOT (Simple Performance-Oriented Test)というテストのスコア）
- 執筆後にアンケート
  - 母語でアカデミックライティングの授業を受けたことがあるか
  - いつ受けたか
  - そこで何を勉強したか
  - アカデミックライティングを書く際の日本語と母語の違い（自由回答）
  - 今回，どういうことに注意して書いたか
    - 10択で複数選択可（例：全体の構成，段落つけ，文の結束性…など）

３．本研究のデータ

	A1旅行　136編	A2料理　153編
性別	男44人，女89人	男43人，女106人
日本語学習歴	3年未満26人 3～5年73人 6年以上34人	3年未満32人 3～5年96人 6年以上21人
日本での学習歴	あり48人，なし85人	あり39人，なし110人
日本語能力試験受験経験	あり68人，なし65人	あり64人，なし85人
SPOT（全体　90点満点）	平均65.5点（中級前半）	平均64.2点（中級前半）
母語の作文授業受講経験	あり66人，なし55人	あり71人，なし67人
今回執筆時の注意（10択）	3つ以下23人 4～6つ75人 7つ以上28人	3つ以下34人 4～6つ74人 7つ以上34人

４．分析方法

- KH Coder（フリーソフト）を使用　http://khc.sourceforge.net/
  - 「テキスト型（文章型）データを統計的に分析するためのフリーソフトウェアです。アンケートの自由記述・インタビュー記録・新聞記事など、さまざまな社会調査データを分析するために制作しました。「計量テキスト分析」または「テキストマイニング」と呼ばれる方法に対応。」

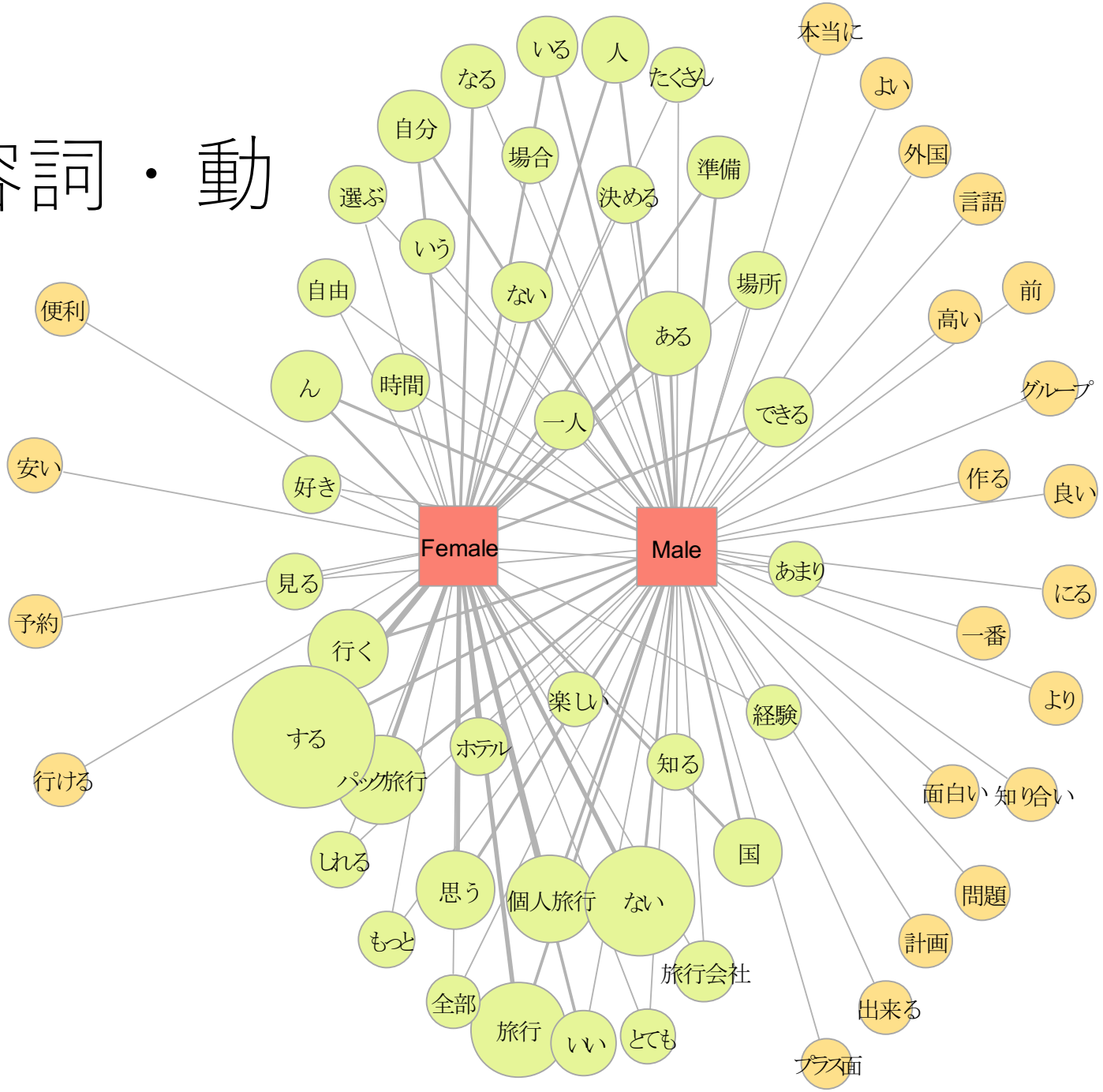
- 大まかに2つの段階からなる分析手順
  - **段階1**：データ中から語を自動的に取り出して、その結果を集計・解析します。これによって、分析者の予断をなるべく交えずに、データの特徴をさぐったり、データを要約したりします。
  - **段階2**：分析者が「こういう表現があれば、コンセプトAが出現していたと見なす」といった指定（コーディングルール作成）を積極的かつ明示的に行い、データ中からコンセプトを取り出します。その結果を集計・解析することで、分析を深めます。

5．結果（語数）

	A1旅行	A2料理
総抽出語数	50,610	56,325
異なり語数	2,777	3,013
作文数	136	153
1作文あたり総抽出語数	372.13	368.13
1作文あたり異なり語数	20.42	19.69

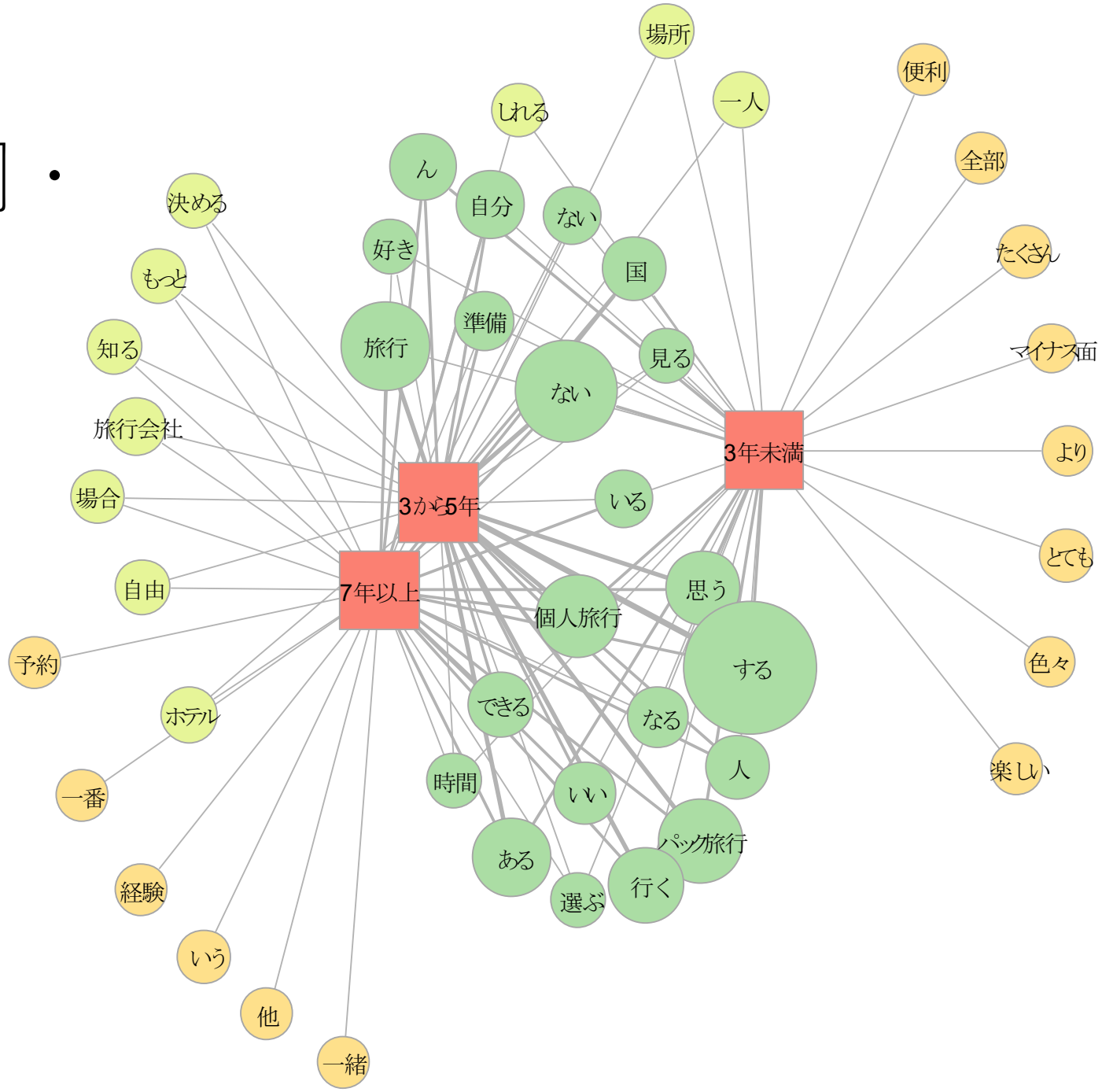
共起ネットワーク
A1旅行　名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全157語のうち,
  100語を描画
- 性別
- 女性：安い，便利
- 男性：外国，言語，
  高い，グループ，
  面白い，知り合い，
  問題，計画…



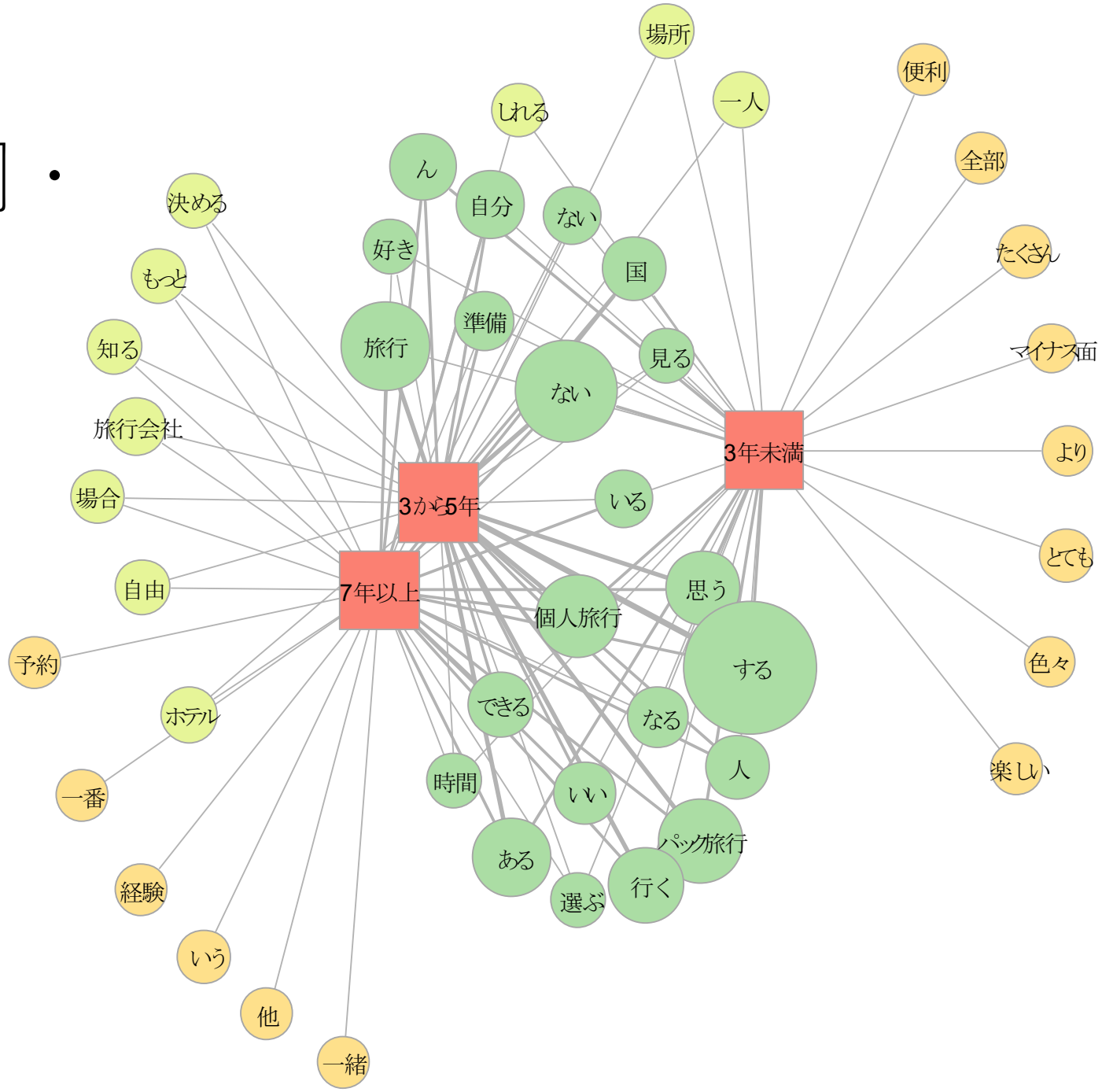
A1旅行
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全157語のうち,
  100語を描画
- 日本語学習歴



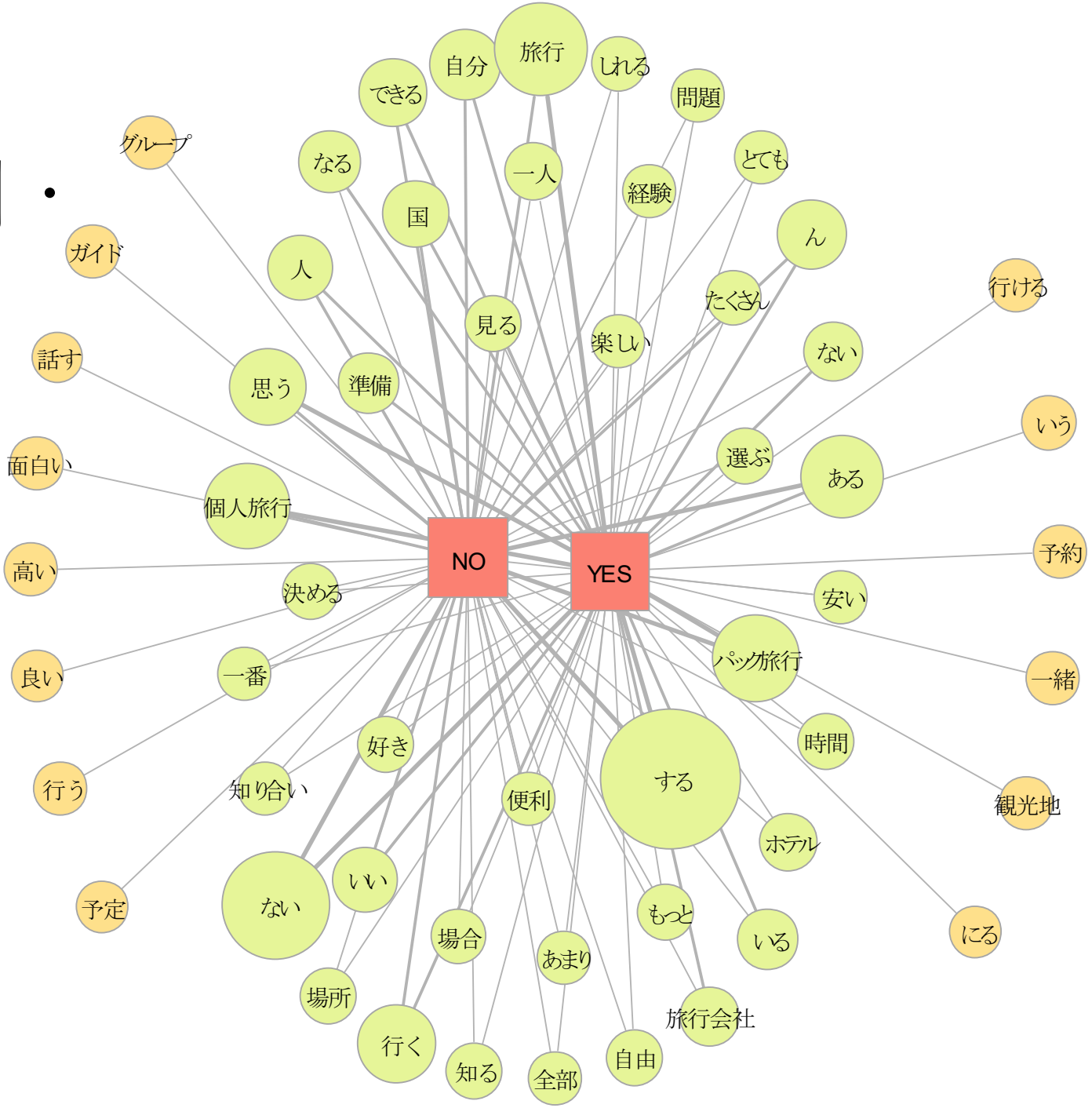
A1旅行
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全157語のうち,
  100語を描画
- 日本での学習経験
- ある方が多様な語



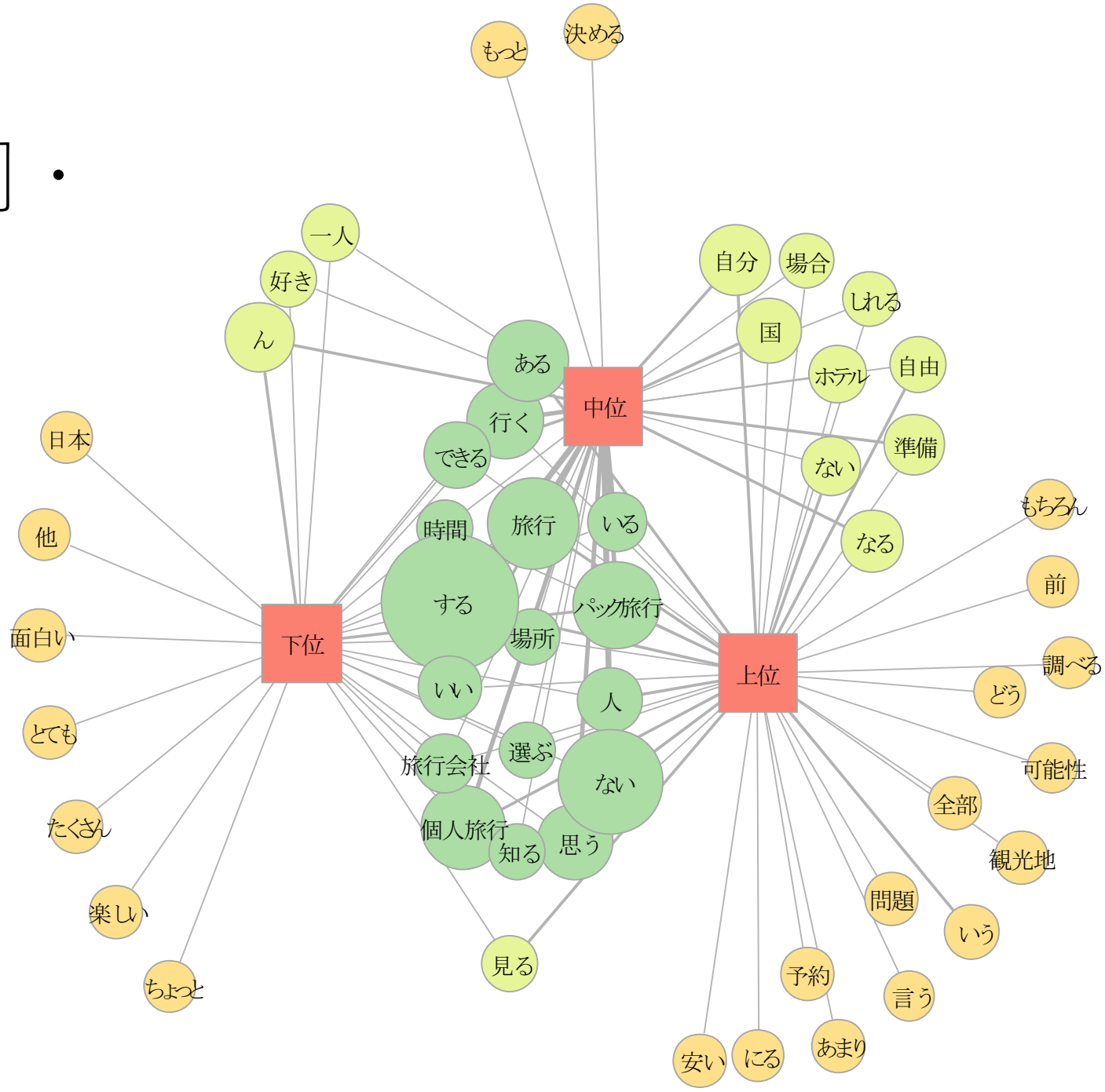
A1旅行
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全157語のうち,
  100語を描画
- 日本語能力試験の
  受験の経験



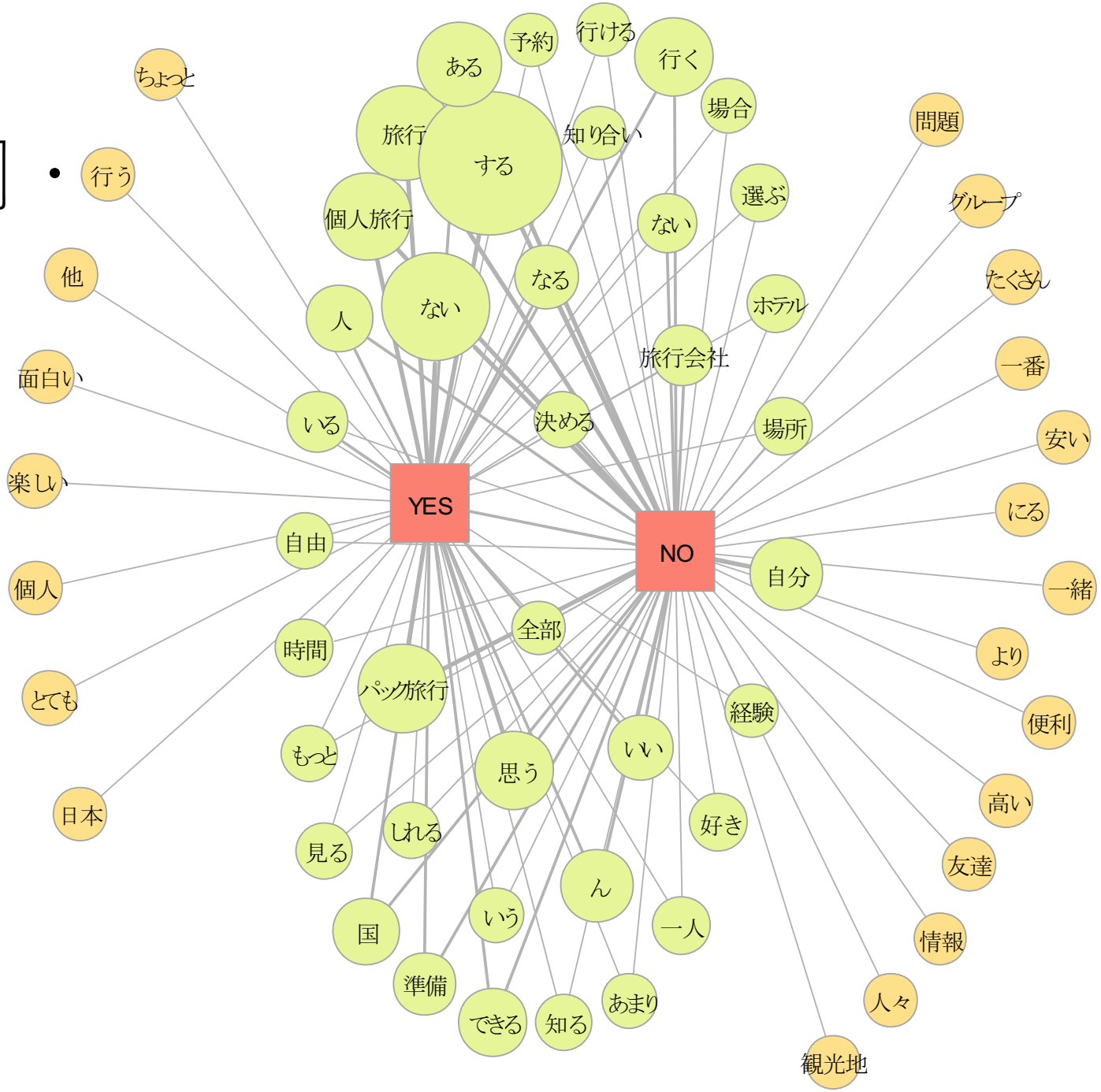
A1旅行
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全157語のうち,
  100語を描画
- SPOT(パート３)の得点
- 上位のほうが多様な語



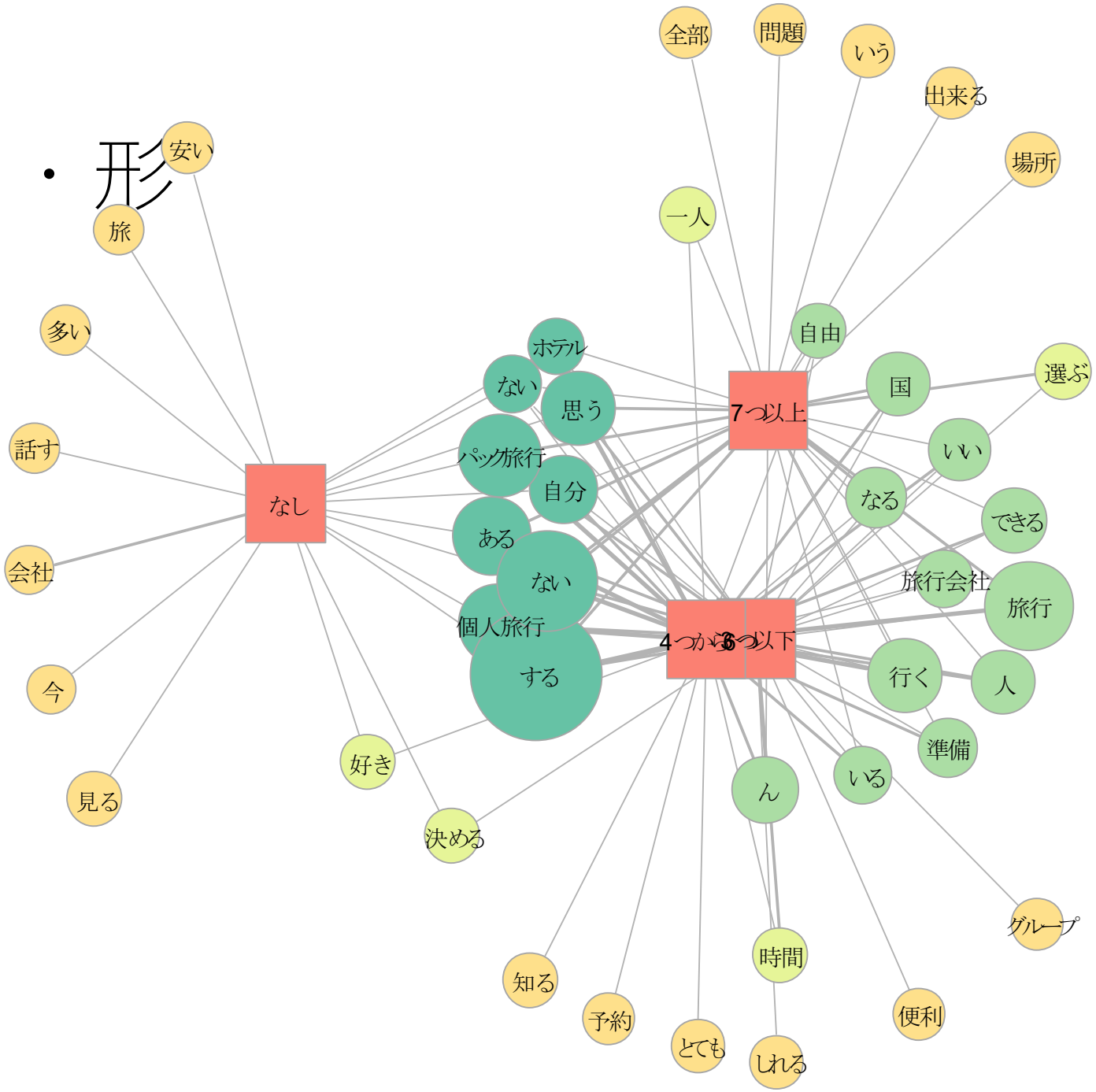
A1旅行
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全157語のうち,
  100語を描画
- 作文学習経験
- ない方が多様な語



A1旅行
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全157語のうち,
  100語を描画
- 今回の作文執筆時の注意

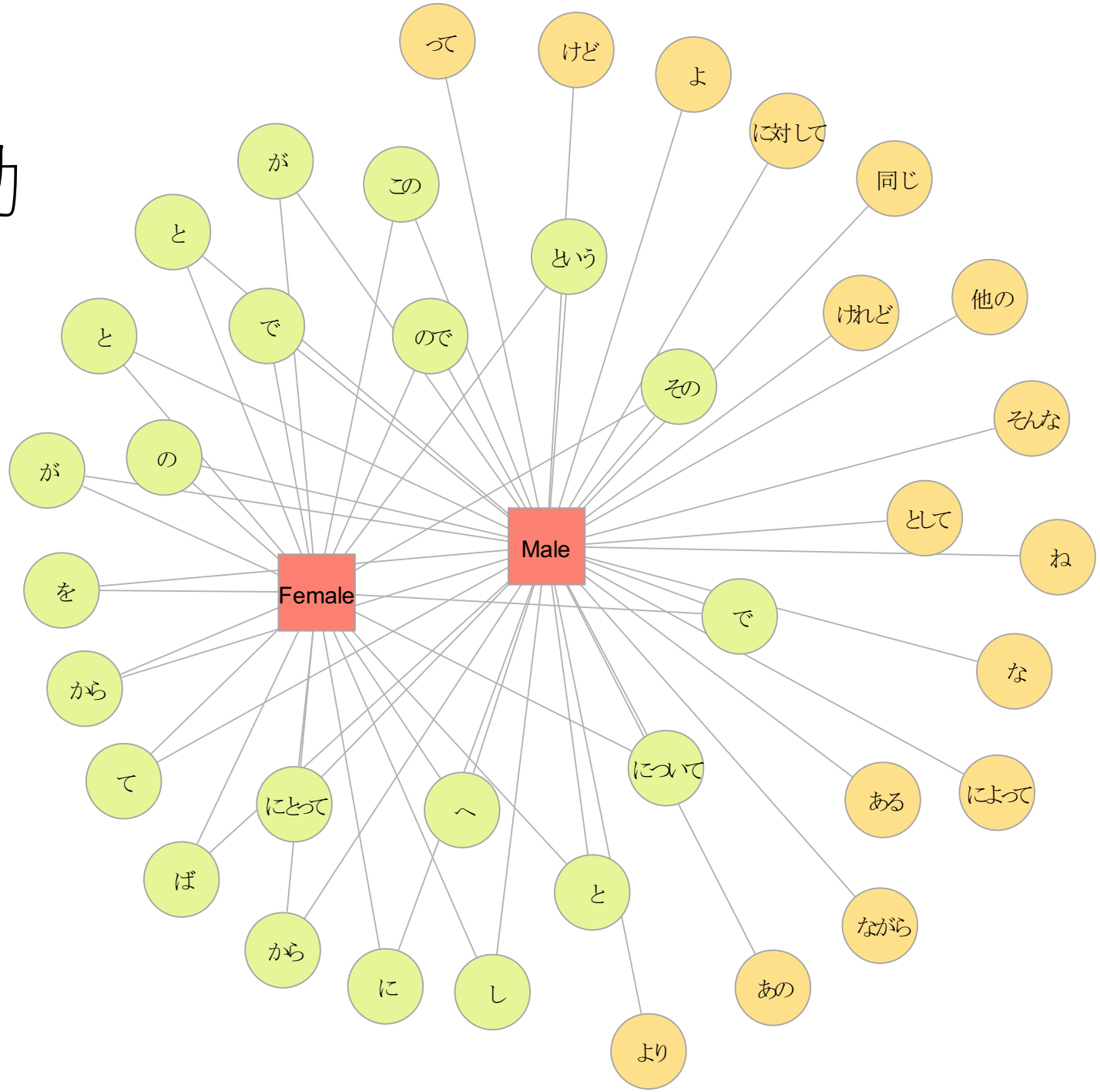


A1旅行
名詞・形容詞・動詞・形容動詞など

- 女性：安い，便利
- 男性：外国，言語，高い，グループ，面白い，知り合い，問題，計画…
- SPOTの点が良いほうが多様な語の使用。

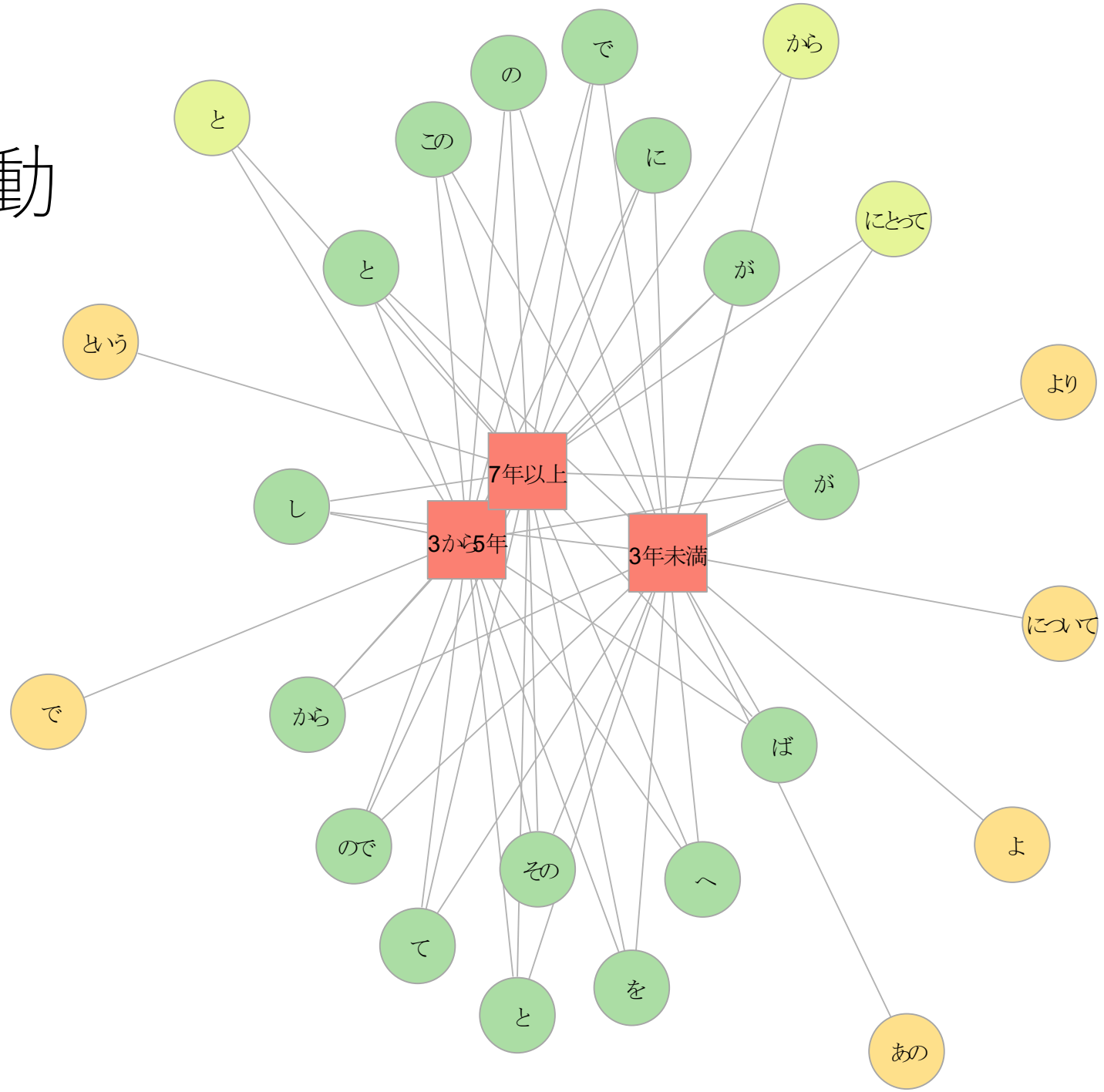
A1旅行
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全50語を描画
- 性別



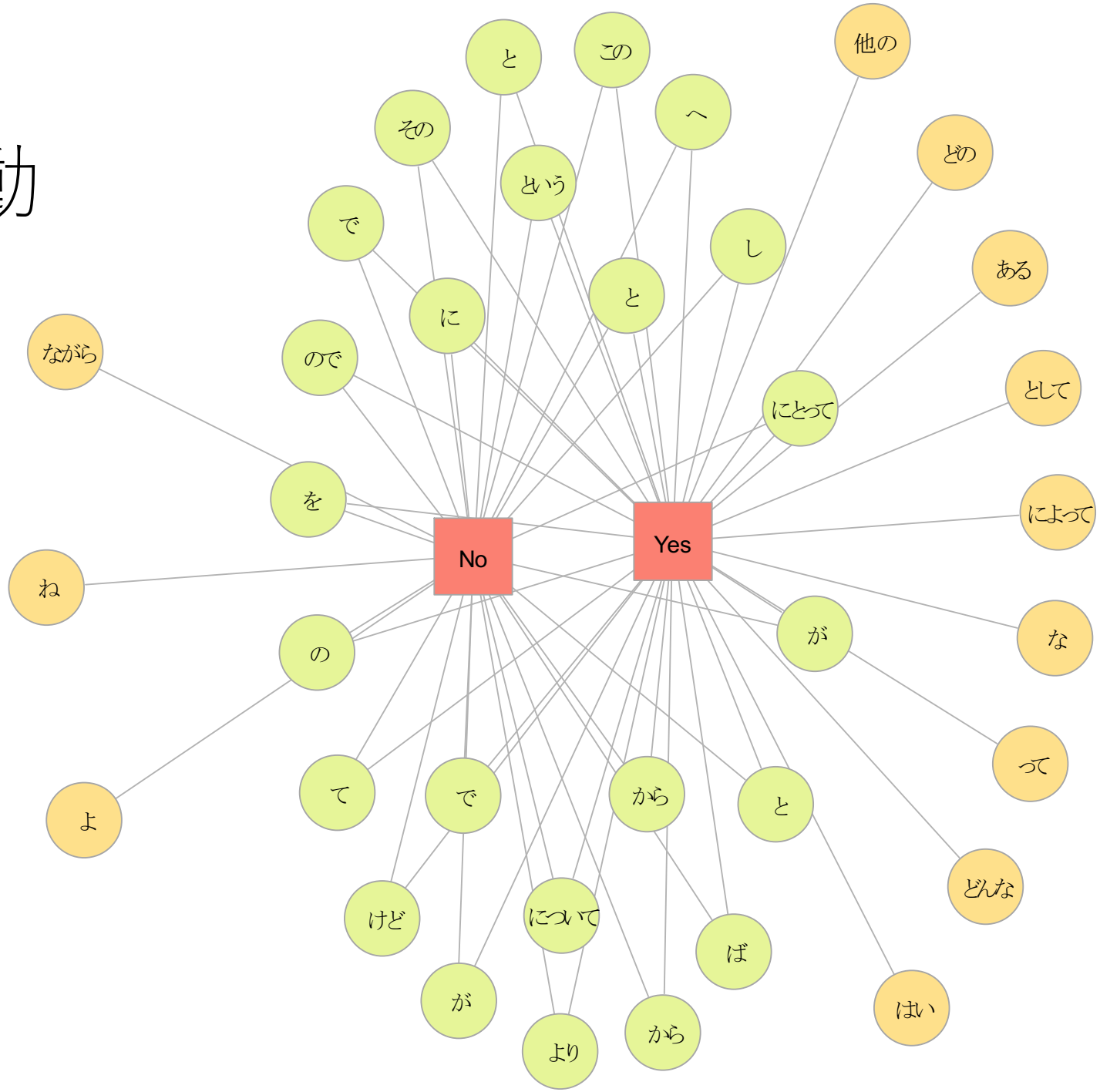
A1旅行
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全50語を描画
- 日本語学習歴
- 短いと「よ」



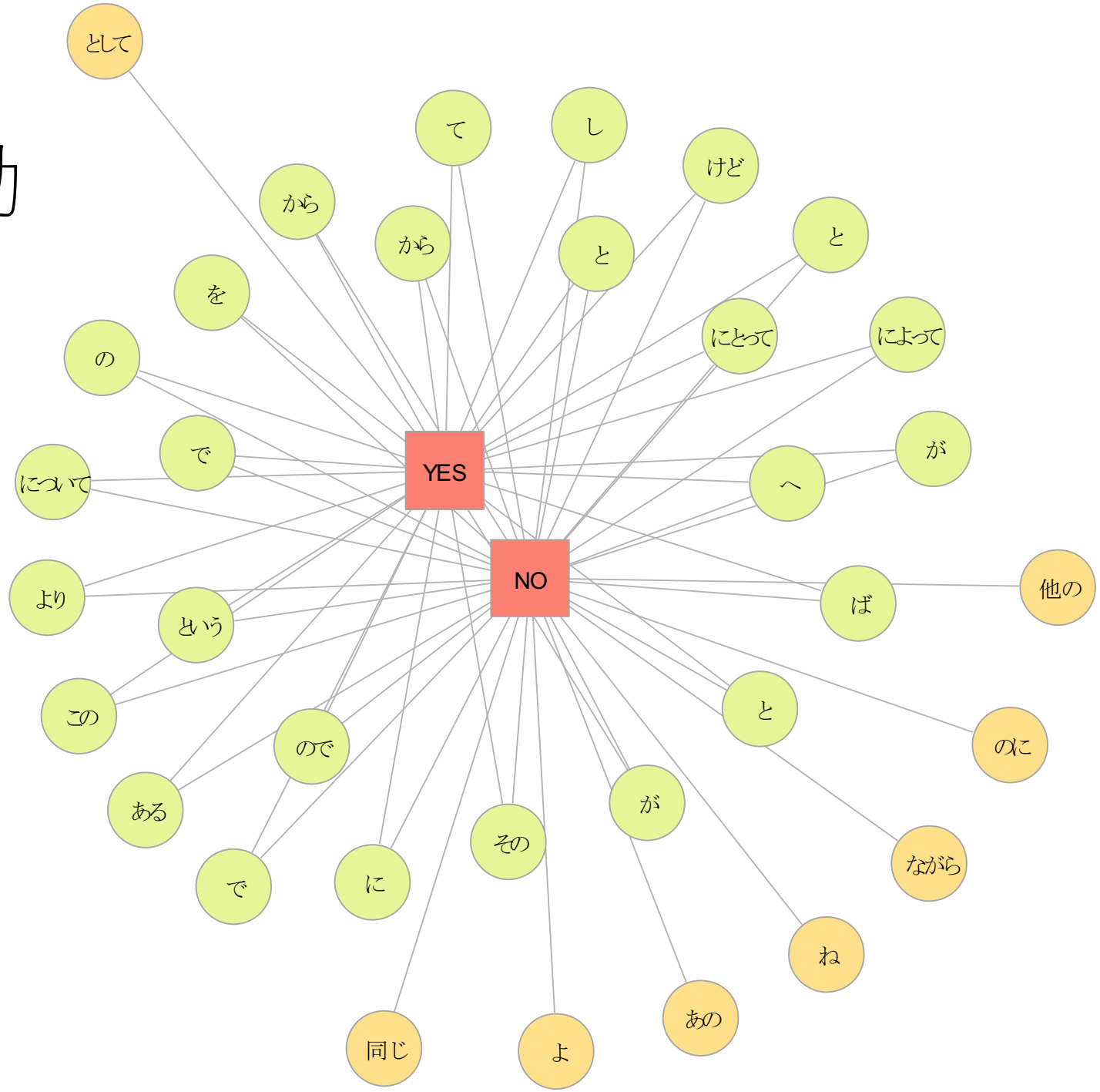
A1旅行
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全50語を描画
- 日本での学習経験
- 経験あると多様
- ないと「よ」
  「ね」



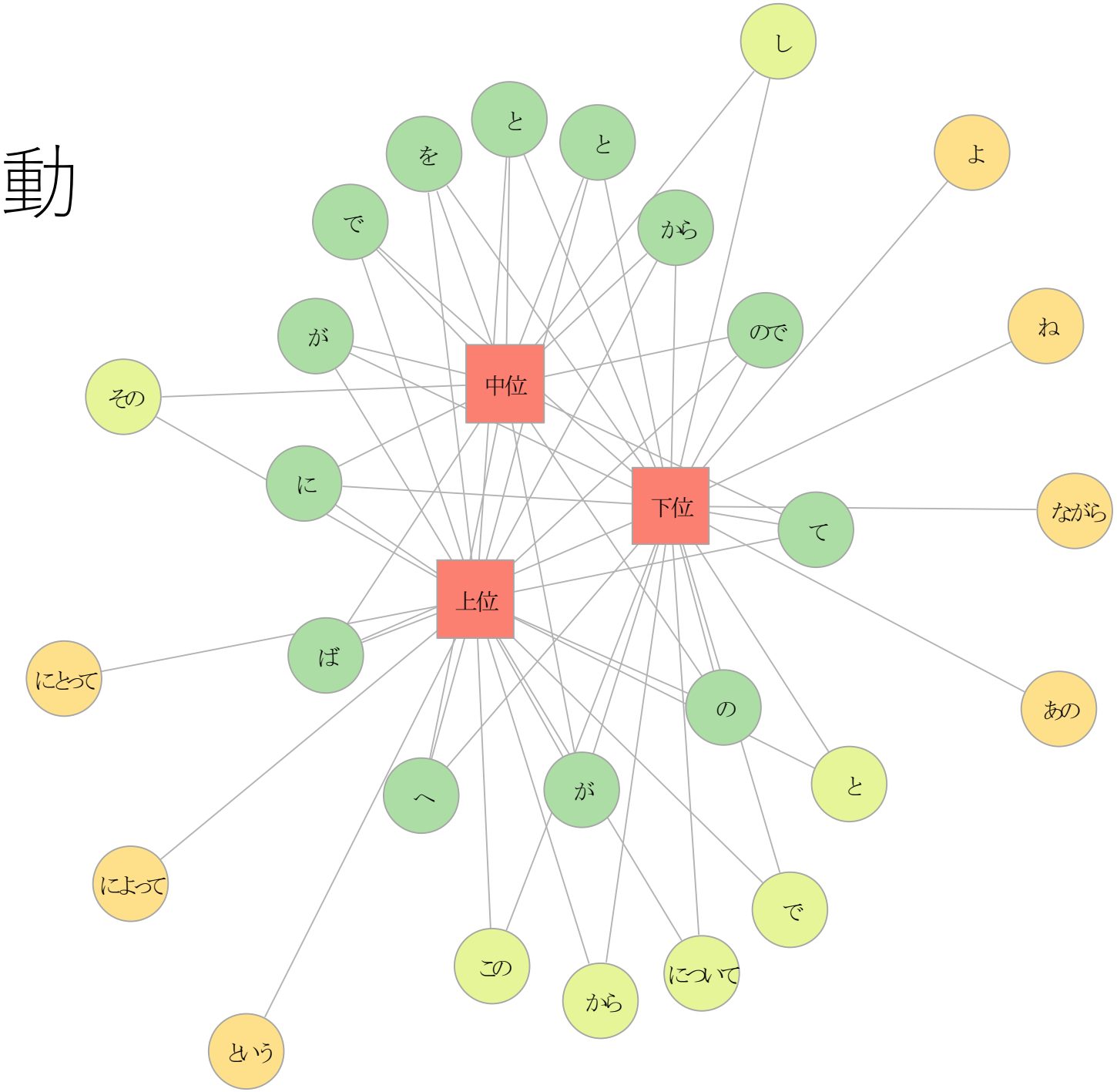
A1旅行
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全50語を描画
- 日本語能力試験の
  受験の経験
- 経験がないと
  「よ」「ね」



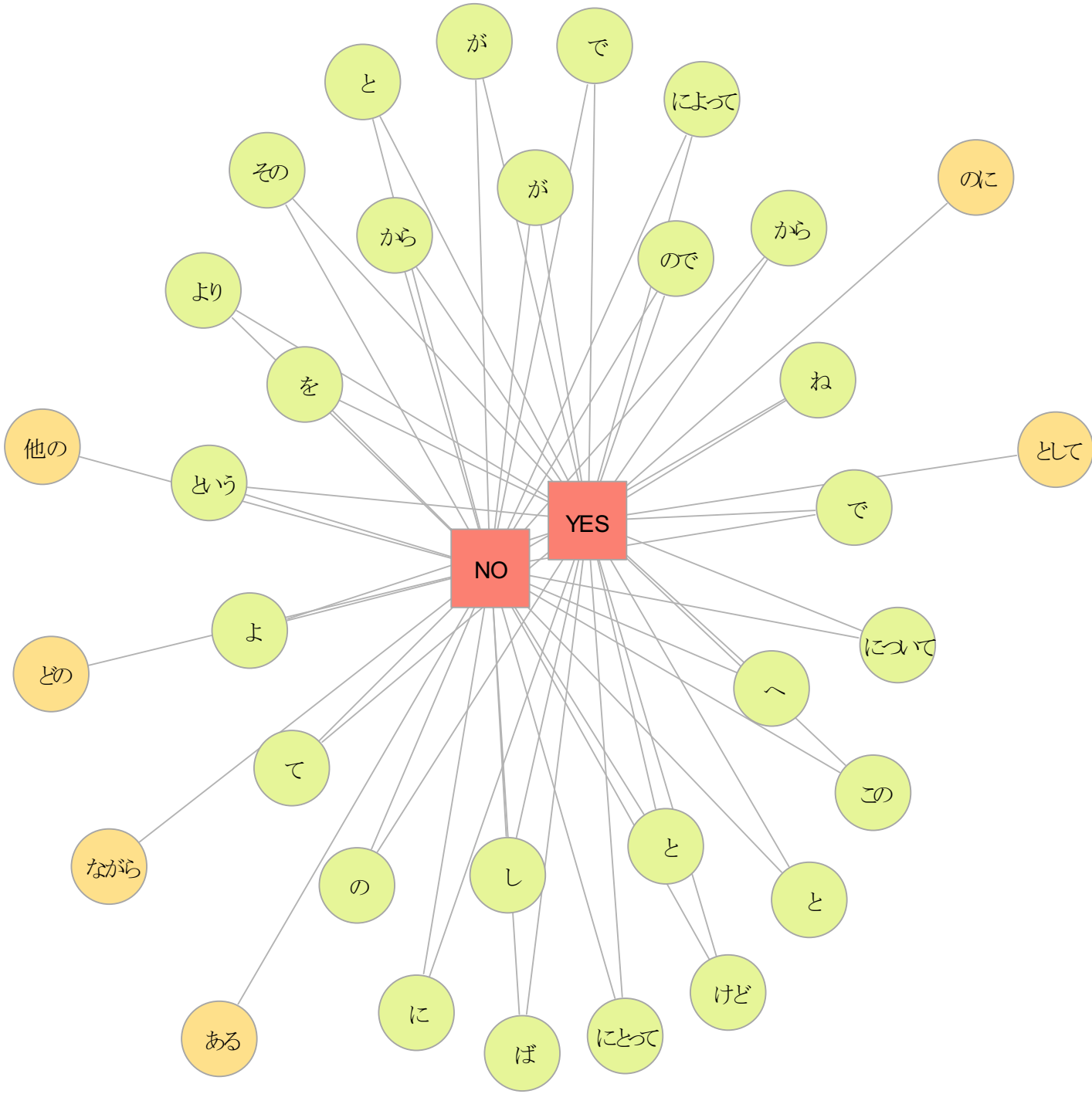
A1旅行
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全50語を描画
- SPOT(パート３)
  の得点
- 下位で「よ」
  「ね」



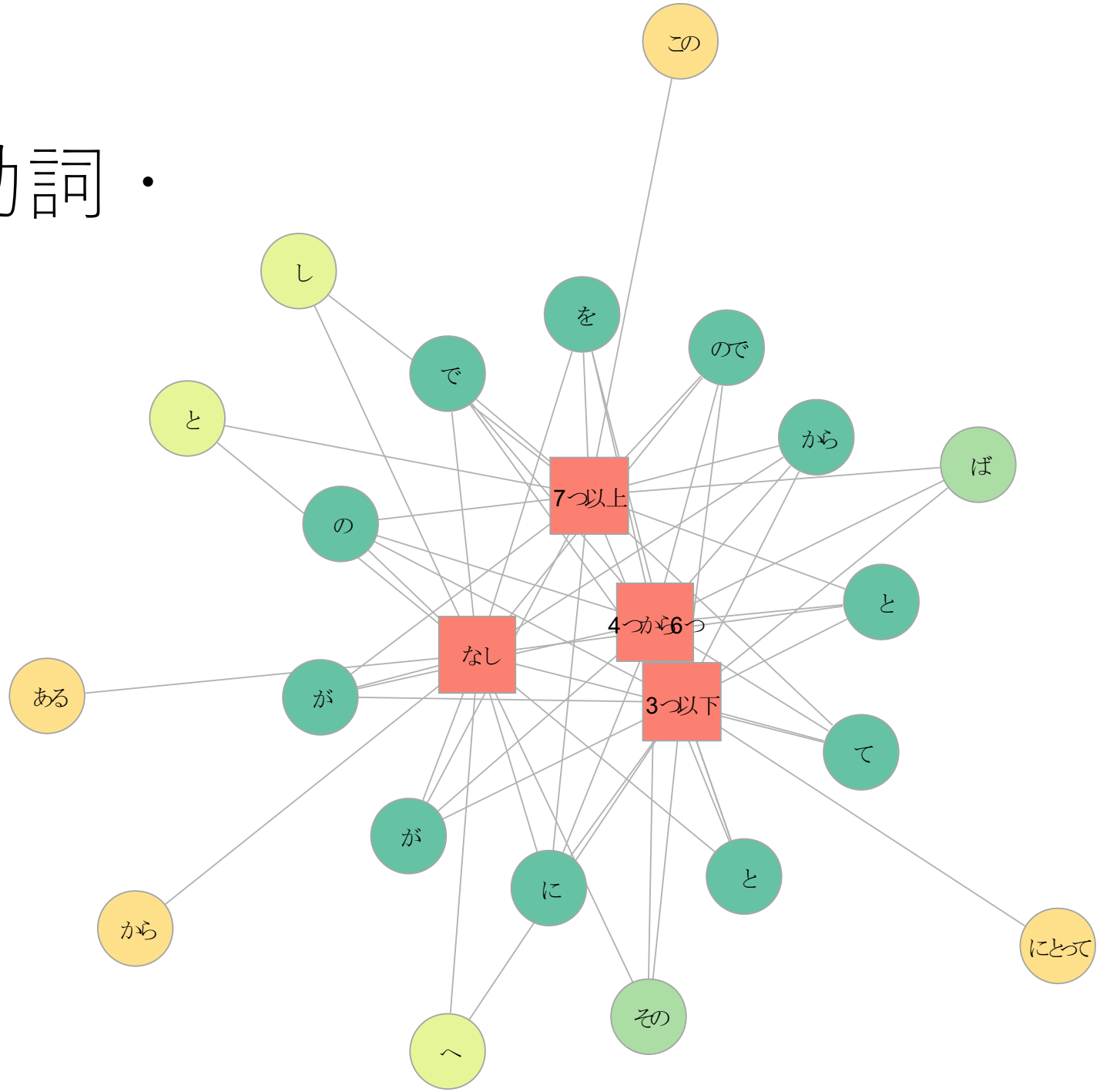
A1旅行
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全50語を描画
- 作文学習経験



A1旅行
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全50語を描画
- 今回の作文執筆時の注意



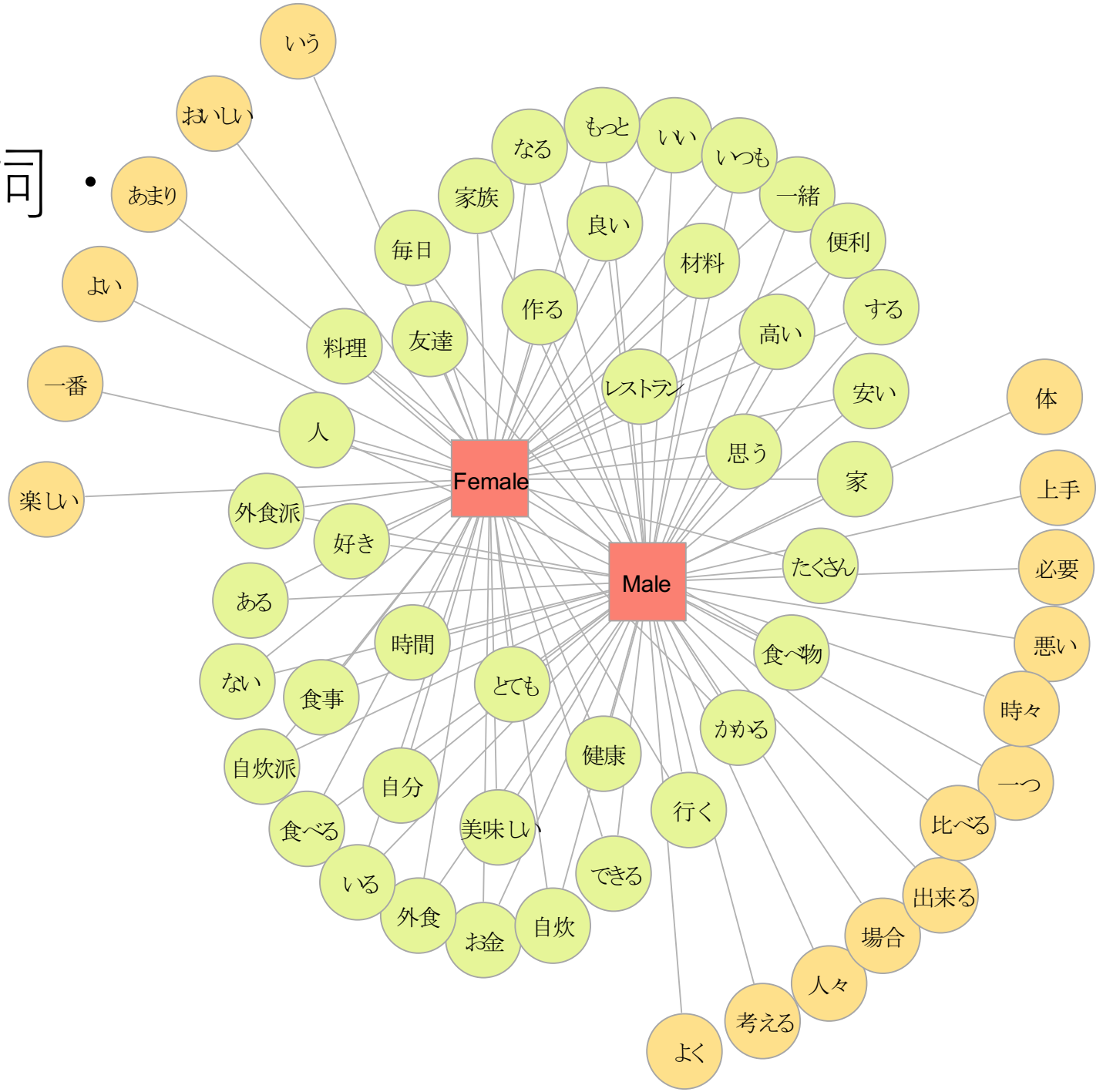
A1旅行
連体詞・助詞・助動詞・感動詞など

- 日本語学習歴が短かったり，日本での学習経験，日本語能力試験の受験がない人，SPOTの点が低い人

  →文章なのに「よ」「ね」を使う。
- 日本語学習歴があると多様な表現を使う。

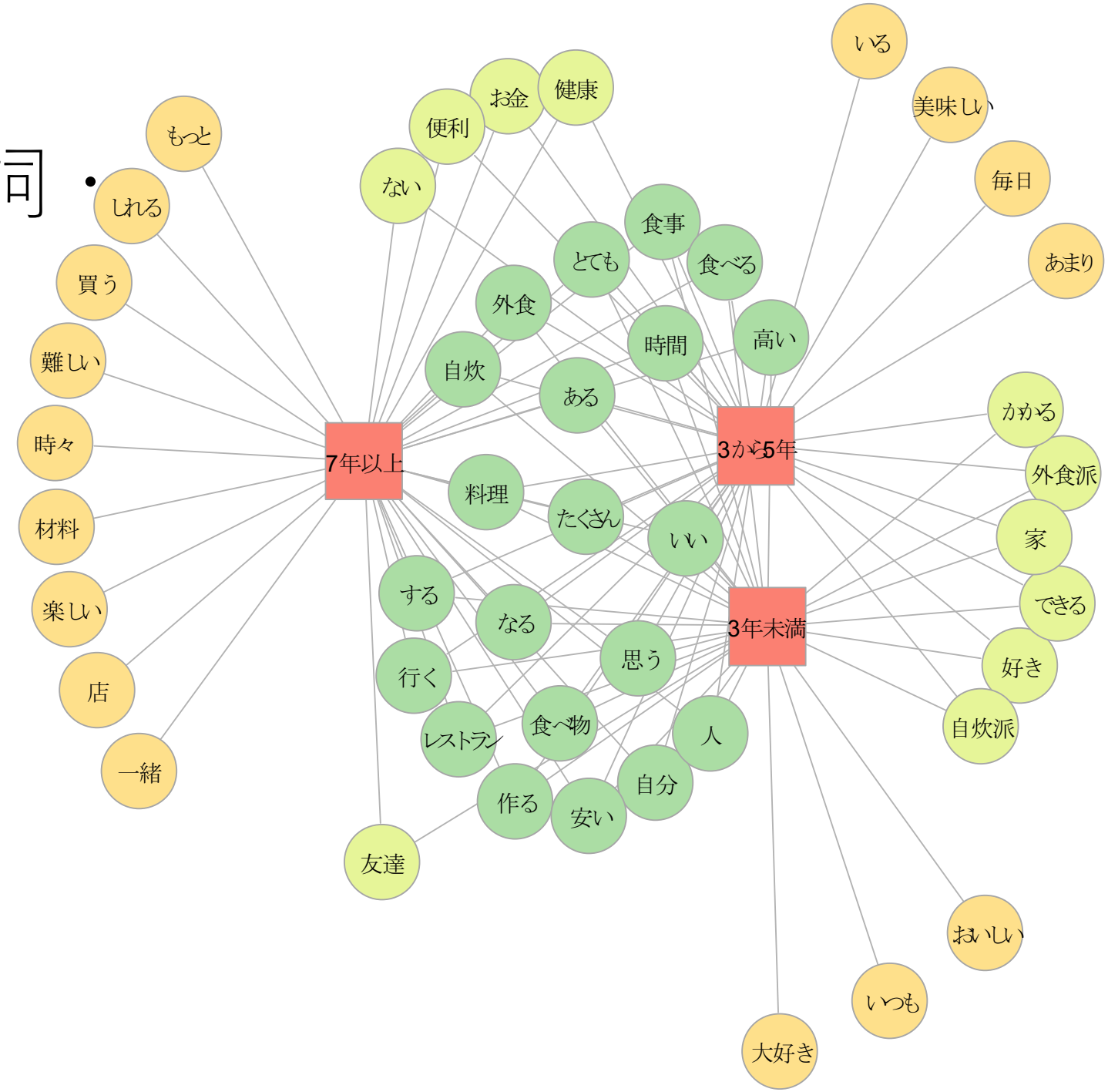
A2料理
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上，全155語のうち，100語を描画
- 性別
- 女性：おいしい，楽しい
- 男性：体，必要，比べる，考える



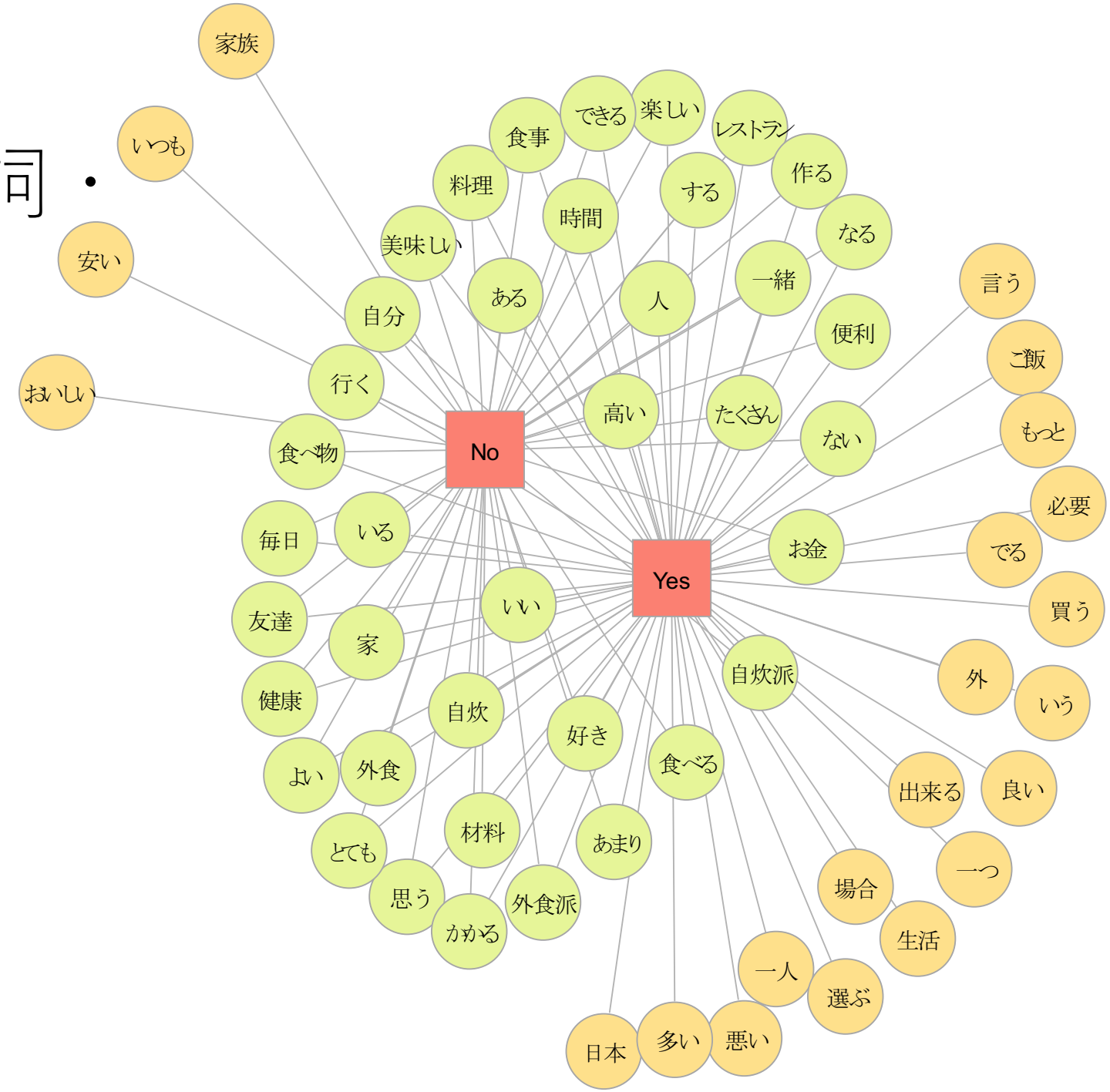
A2料理
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全155語のうち,
  100語を描画
- 日本語学習歴
- 長い方が多様な語



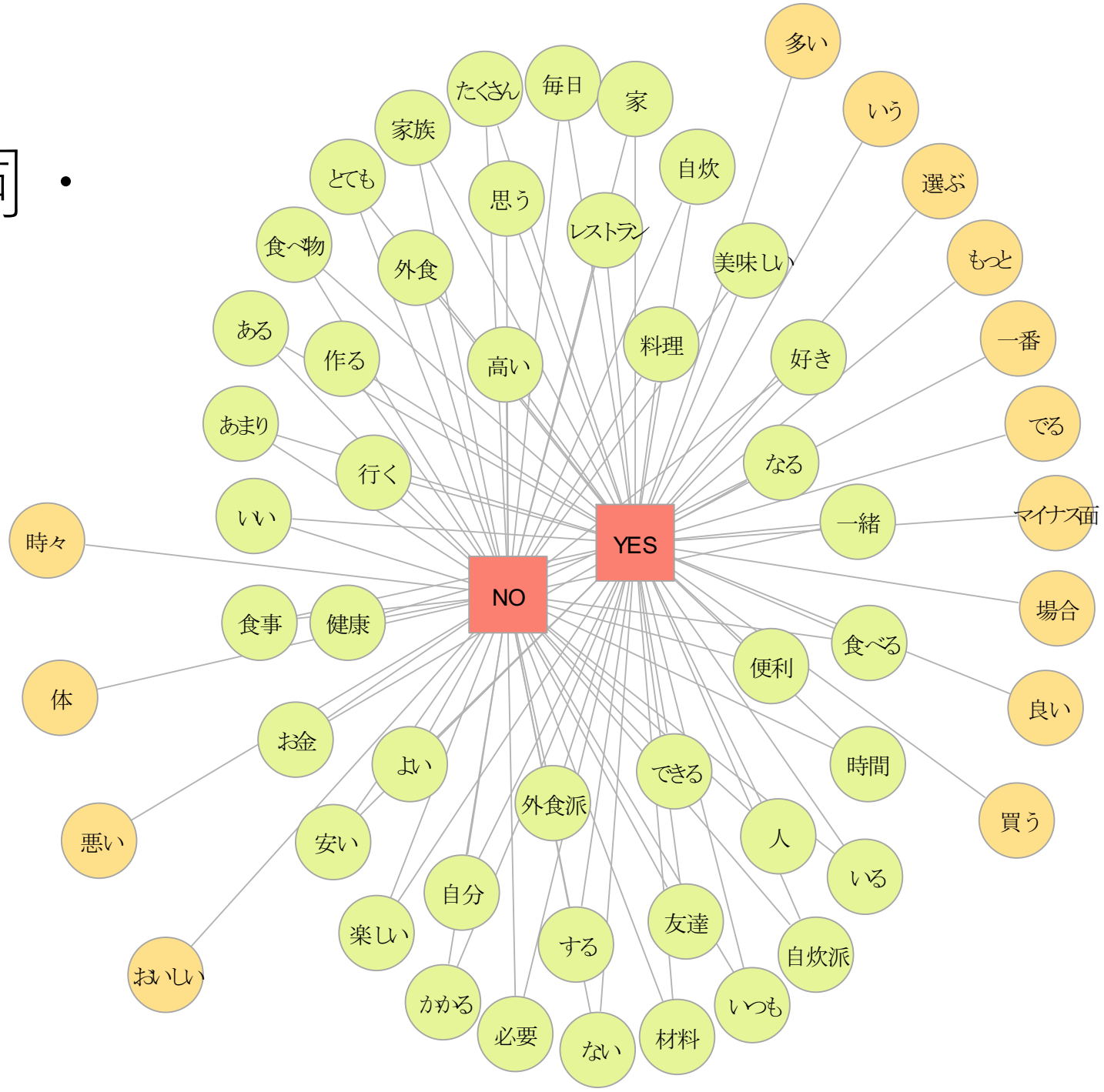
A2料理
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全155語のうち,
  100語を描画
- 日本での学習経験
- ある方が多様な語



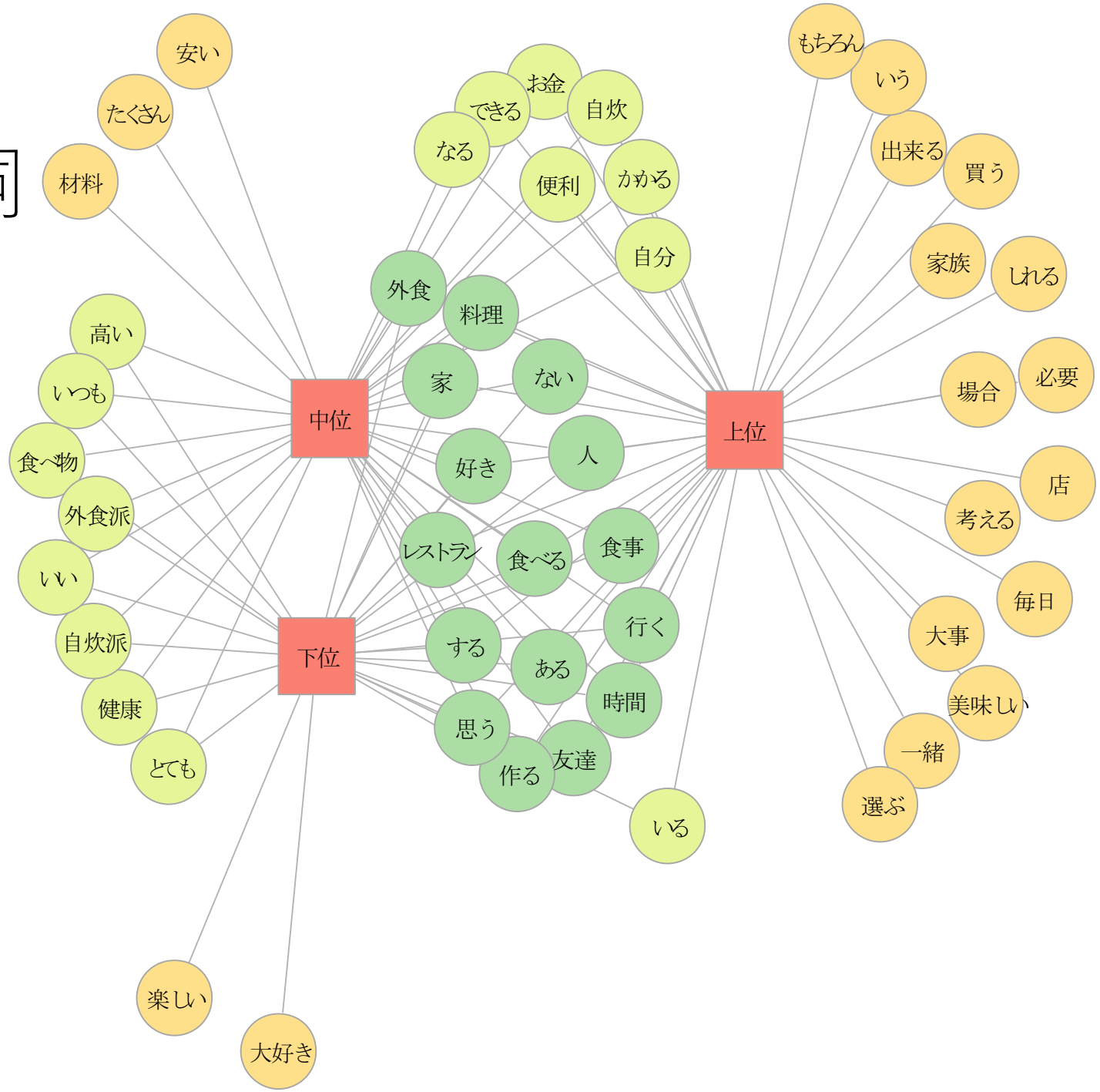
A2料理
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全155語のうち,
  100語を描画
- 日本語能力試験の受験の経験
- ある方が多様な語



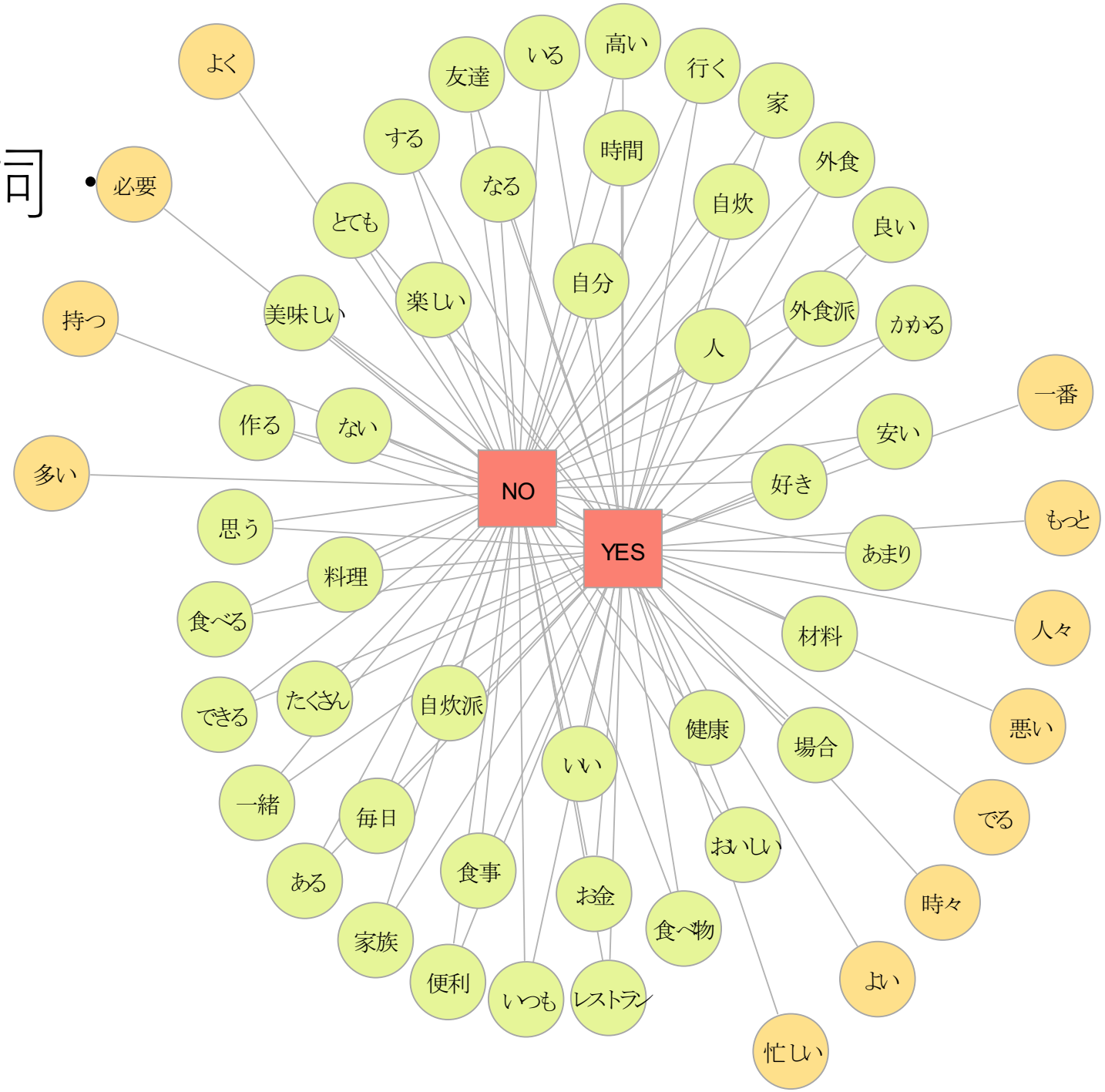
A2料理
名詞・形容詞・動詞
形容動詞など

- 出現回数20回以上,
  全155語のうち,
  100語を描画
- SPOT(パート３)
  の得点
- 上位のほうが多様
  な語



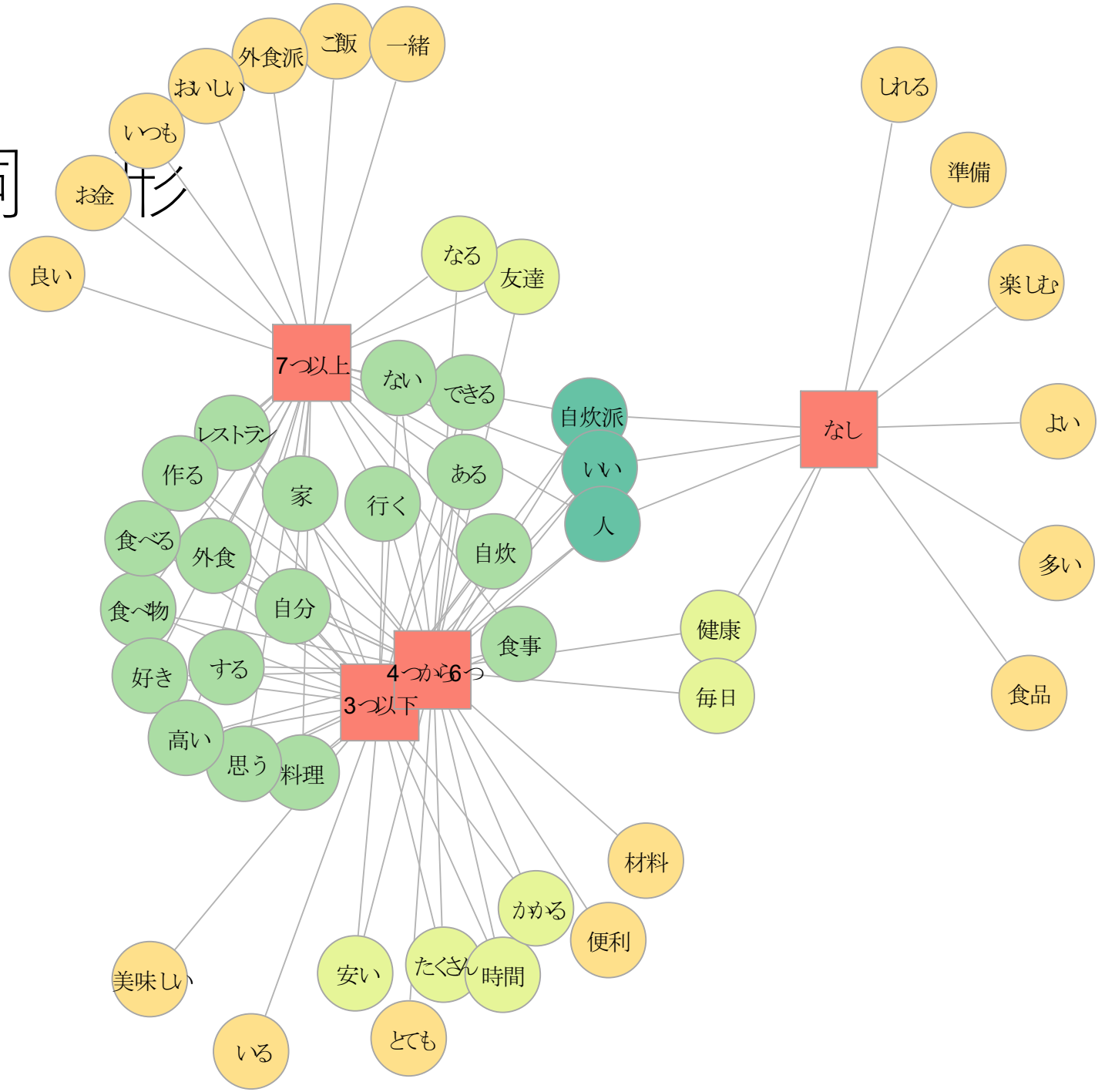
A2料理
名詞・形容詞・動詞・形容動詞など

- 出現回数20回以上,
  全155語のうち,
  100語を描画
- 作文学習経験
- ある方が多様な語



A2料理
名詞・形容詞・動詞形容動詞など

- 出現回数20回以上，全155語のうち，100語を描画
- 今回の作文執筆時の注意

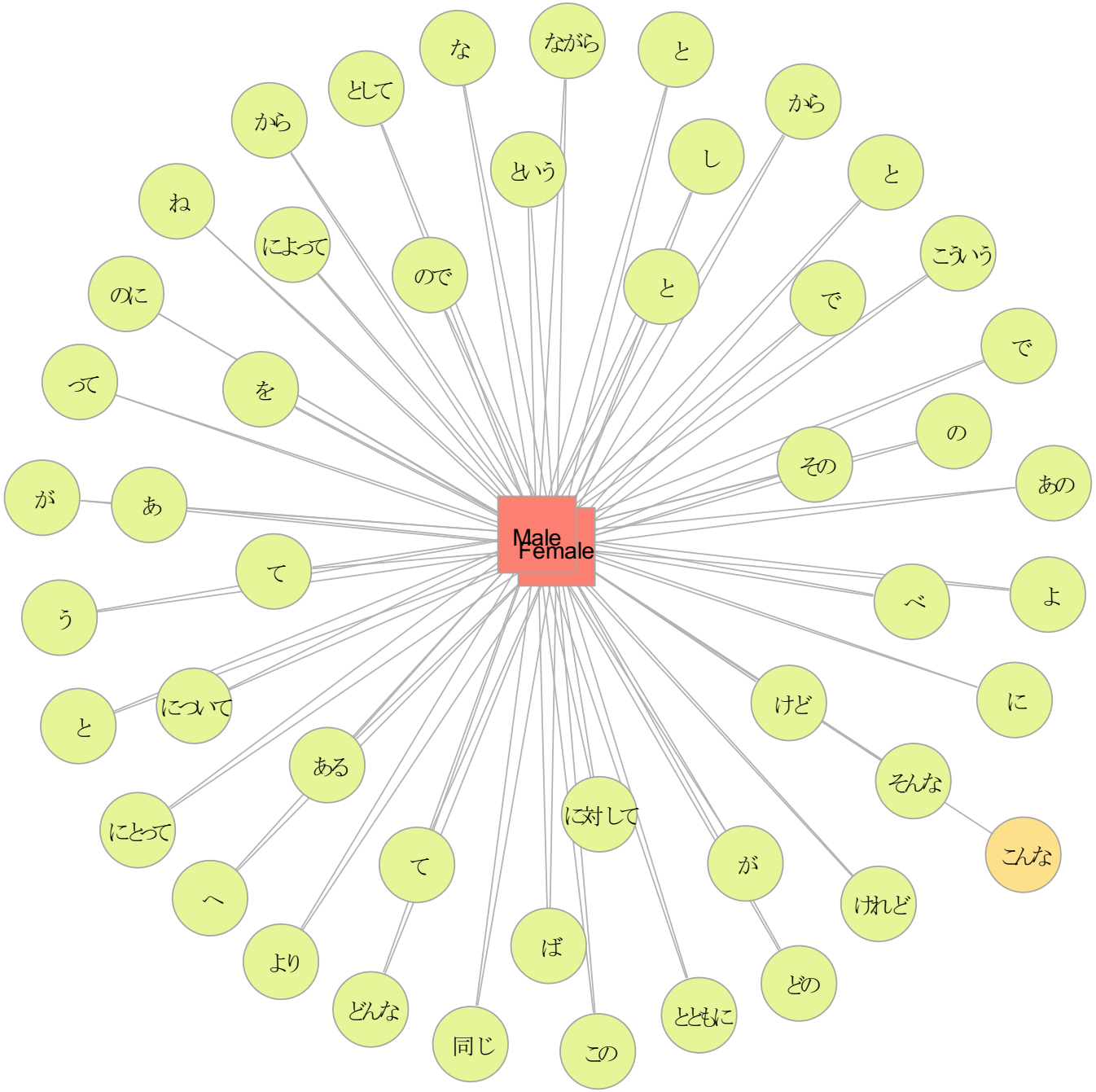


A2料理
名詞・形容詞・動詞・形容動詞など

- 女性：おいしい，楽しい
- 男性：体，必要，比べる，考える
- 学習歴などが長い・経験が多いほうが，多様な語の使用。

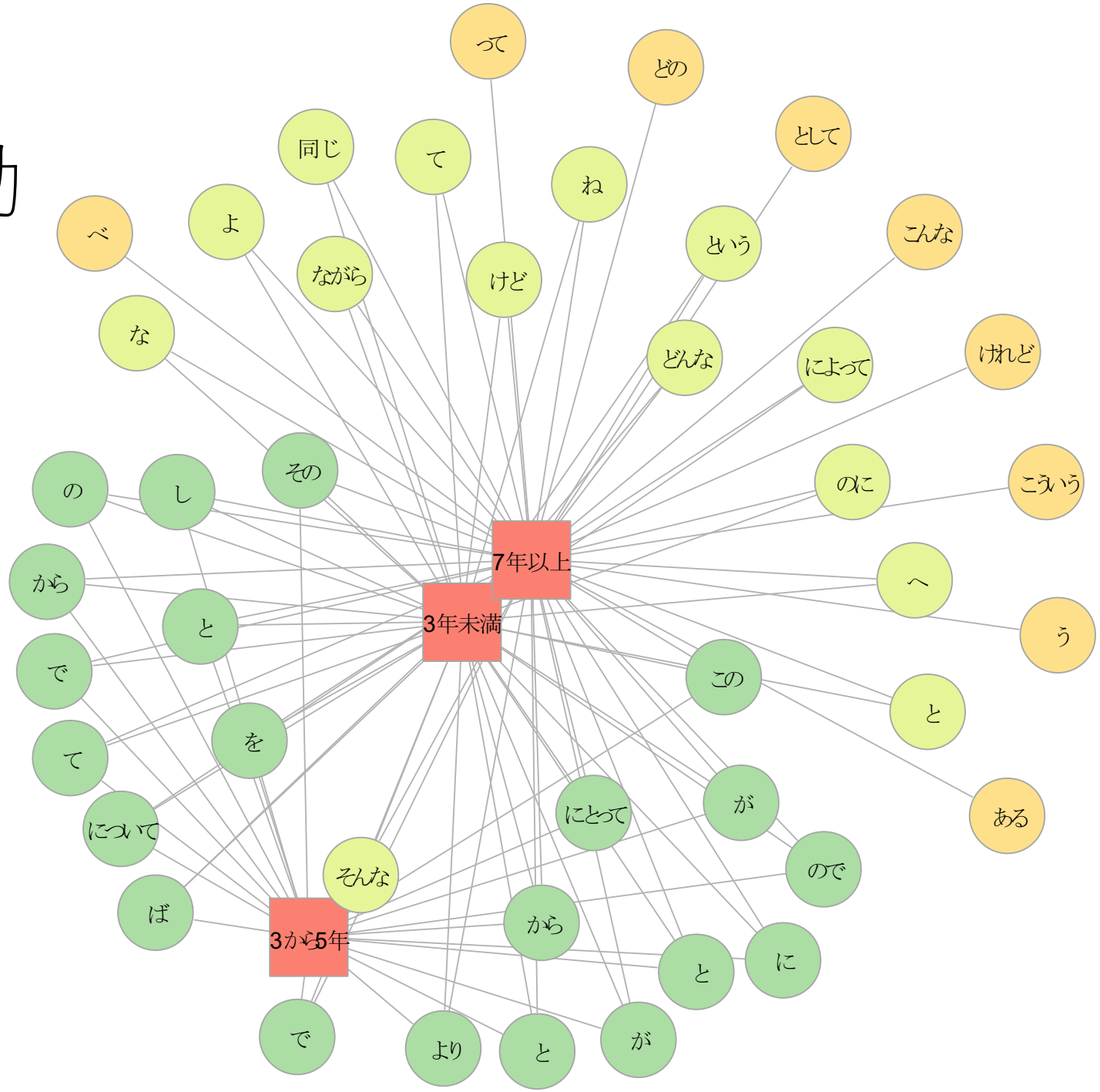
A2料理
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全48語を描画
- 性別



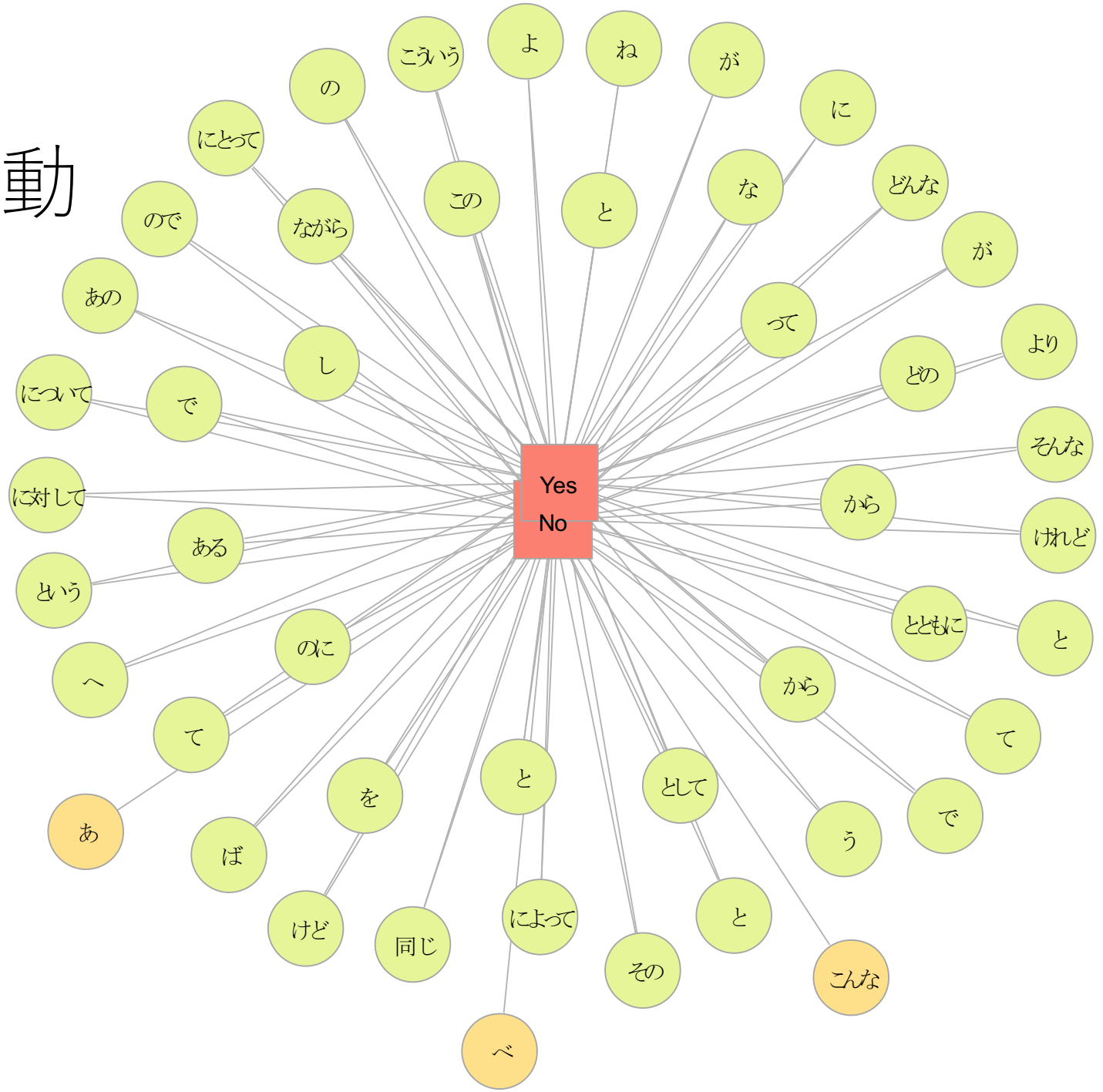
A2料理
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全48語を描画
- 日本語学習歴



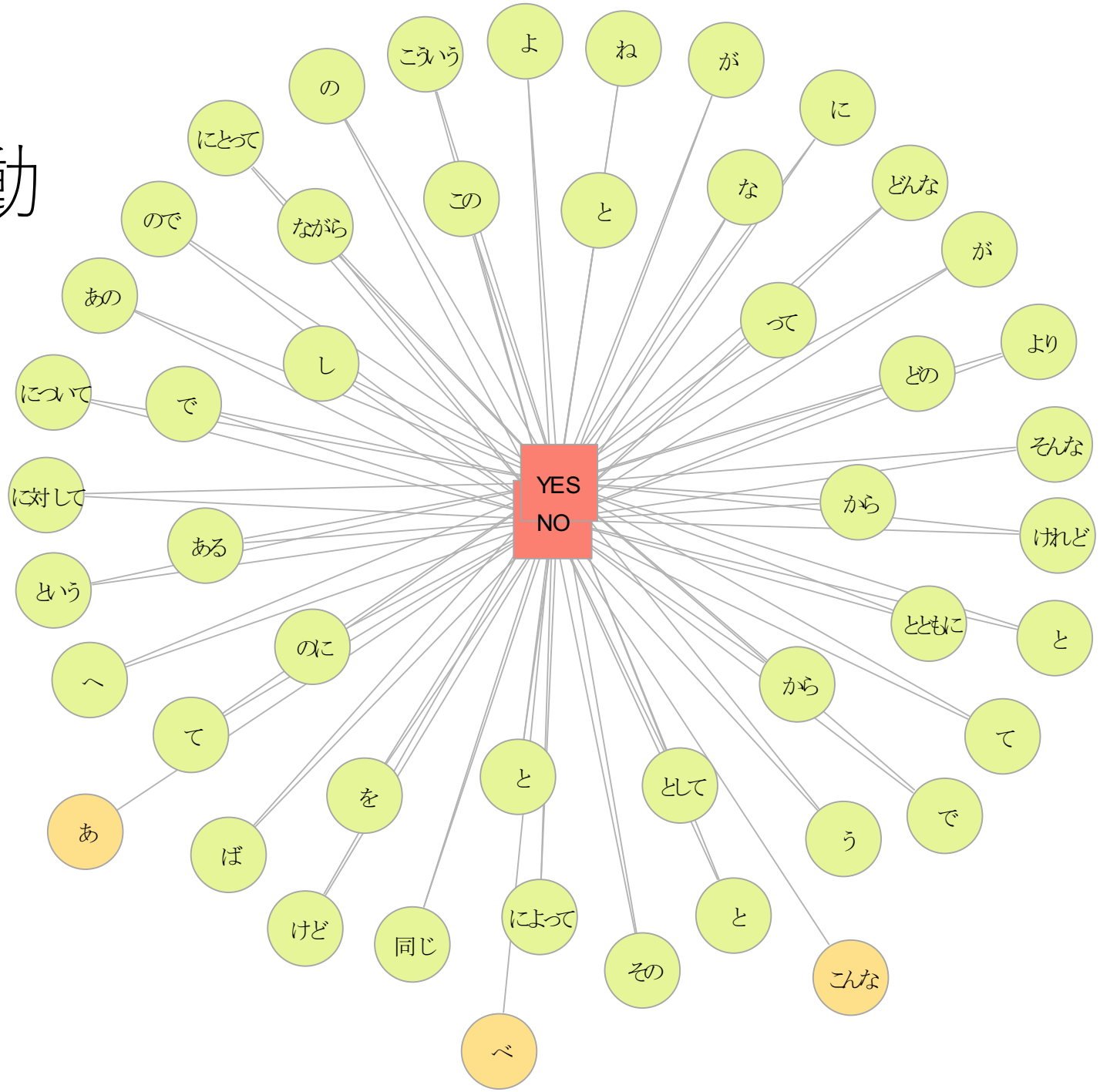
A2料理
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全48語を描画
- 日本での学習経験



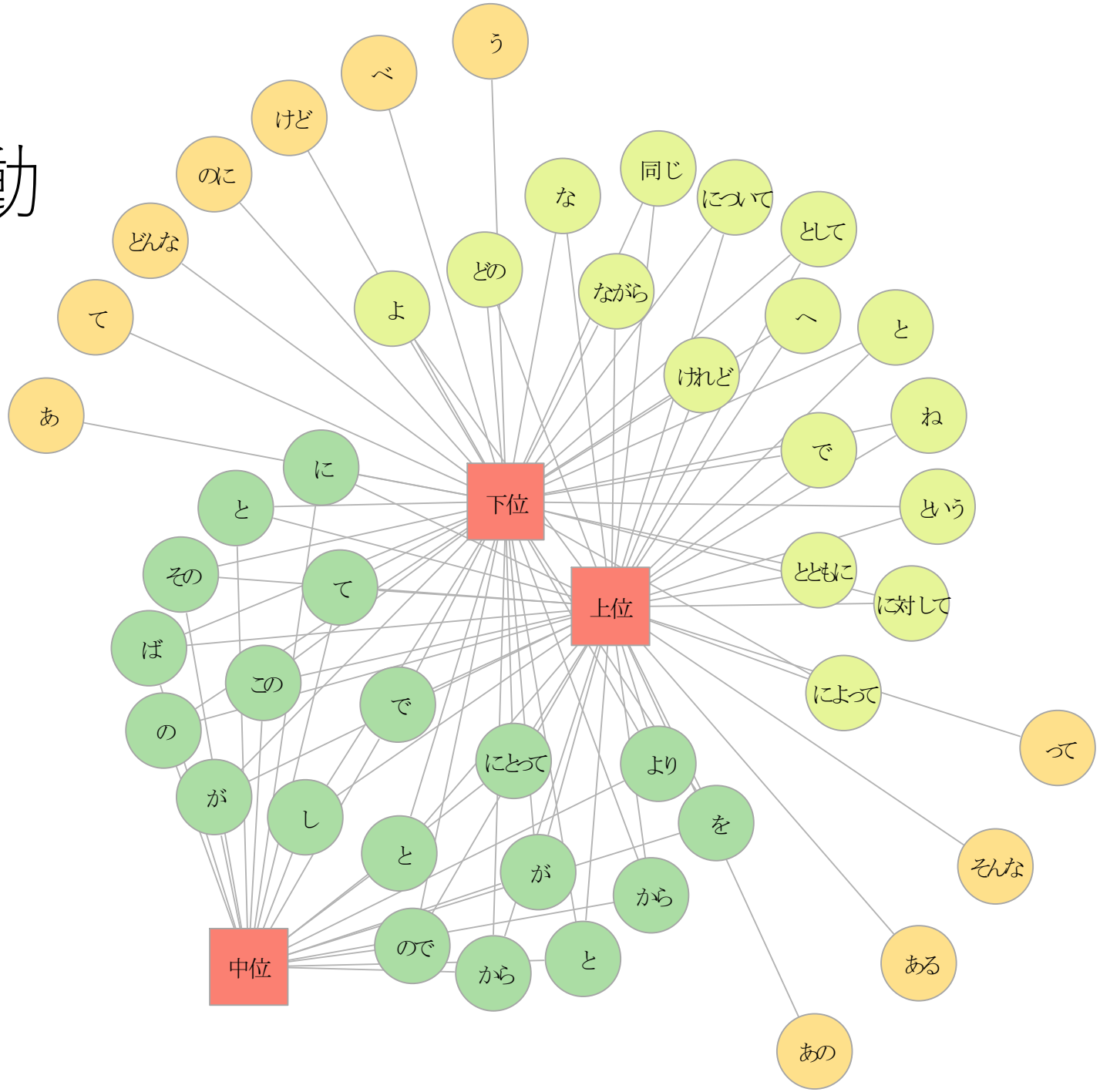
A2料理
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全48語を描画
- 日本語能力試験の
  受験の経験



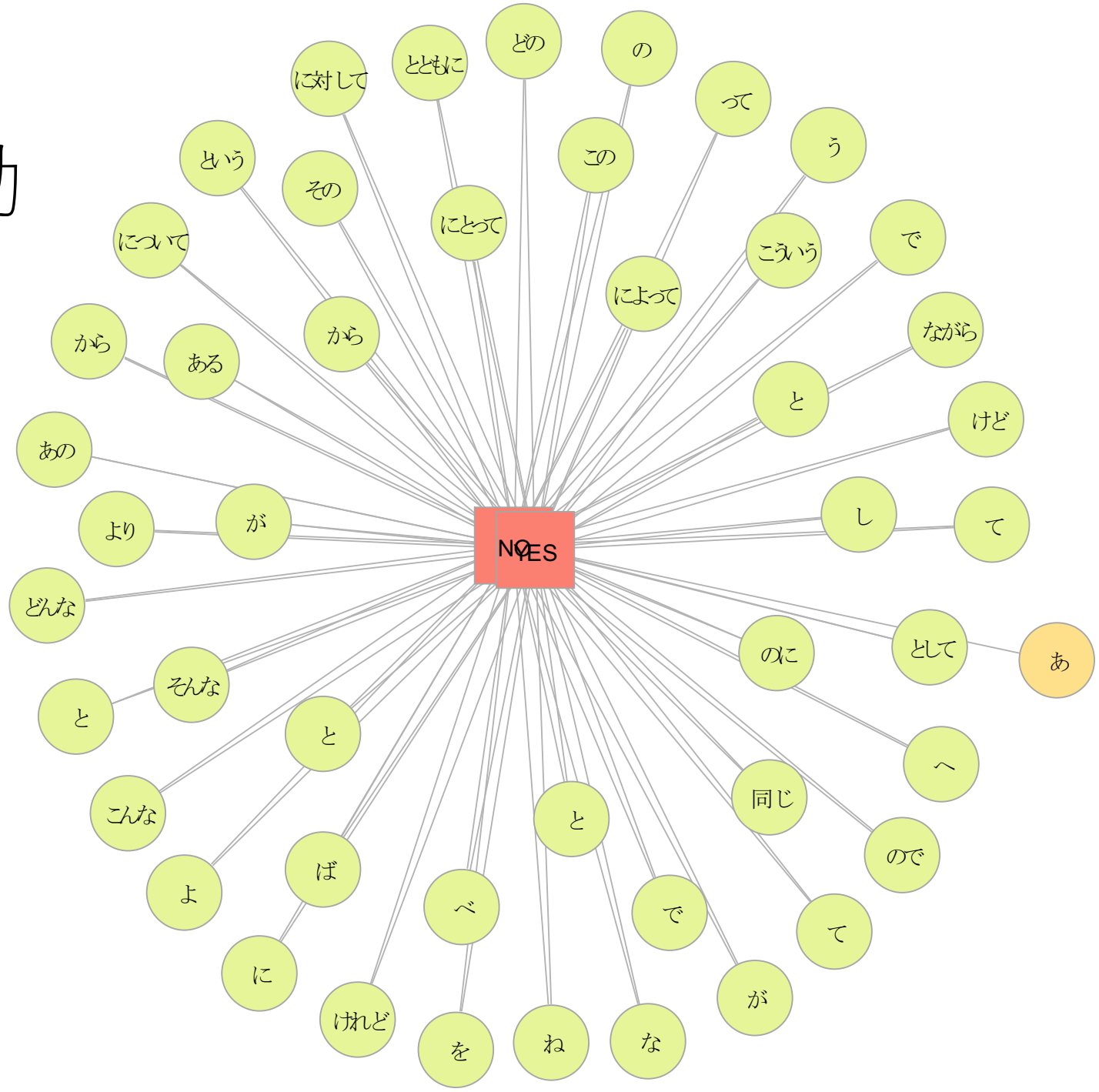
A2料理
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全48語を描画
- SPOT(パート３)
  の得点



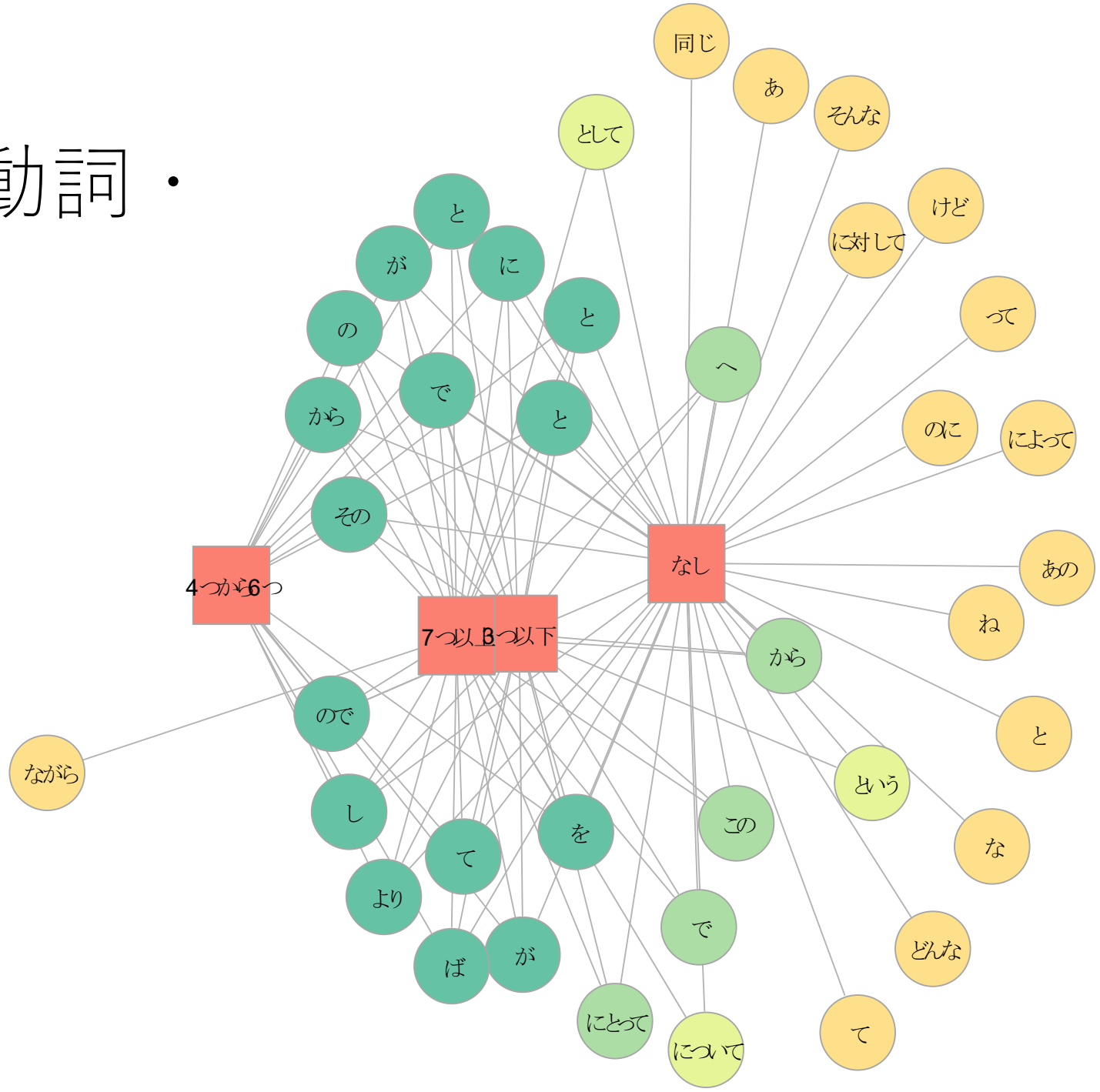
A2料理
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全48語を描画
- 作文学習経験



A2料理
連体詞・助詞・助動詞・感動詞など

- 出現回数5回以上,
  全48語を描画
- 今回の作文執筆時の注意



A2料理
連体詞・助詞・助動詞・感動詞など

- 学習歴が長いと多様なものが使われる。

名詞・形容詞・動詞・形容動詞など

- A1旅行
  - 女性：安い，便利
  - 男性：外国，言語，高い，グループ，面白い，知り合い，問題，計画…
  - SPOTの点が良いほうが多様な語の使用。

- A2料理
  - 女性：おいしい，楽しい
  - 男性：体，必要，比べる，考える…
  - 学習歴などが長い・経験が多いほうが，多様な語の使用。

連体詞・助詞・助動詞・感動詞など

- A1旅行
  - 日本語学習歴が短かったり，日本での学習経験，日本語能力試験の受験がない人，SPOTの点が低い人
    →文章なのに「よ」「ね」を使う。
  - 日本語学習歴が長いと多様な表現を使う。

- A2料理
  - 日本語学習歴が長いと多様なものが使われる。

６．検討

- 男女の違いが比較的見られた。
- 経験が多い人は表現が多様である。

- 文章で文末に「よ」「ね」を使う人がいる。
  - 日本語学習歴が短い
  - 日本での学習経験，日本語能力試験の受験がない
  - SPOTの点が低い

    →文章の書き方についての指導が必要・重要

参考文献

- 阿部新（2014）「表現の「自然さ」の判断基準を探るテキストマイニング―「よろしかったですか」等の表現の「自然さ」についてのアンケートにおける自由回答を例として―」岸江信介・田畑智司編『テキストマイニングによる言語研究』ひつじ書房　pp.59-80.
- 阿部新・嵐洋子・須藤潤（2016）「日本語音声教育の方向性の探索―音声教育に対する日本語教師のビリーフの自由回答をデータとして―」宇佐美洋編『「評価」を持って街に出よう』くろしお出版　pp.270-290.
- 石川慎一郎（2012）『ベーシック　コーパス言語学』ひつじ書房
- 石田基広（2008）『Rによるテキストマイニング入門』森北出版
- 内田治（2010）『数量化理論とテキストマイニング』日科技連出版社
- 大隅昇・Lubart, Ludovic（2000）「調査における自由回答データの解析－InfoMinerによる探索的テキスト型データ解析―」『統計数理』48巻2号　pp.339-376.
- 岸江信介・阿部貴人・石田基広・西尾純二（2011）「ワークショップ　地域言語のデータ処理の批判的検討と新展開」『日本語学会2011年度春季大会予稿集』pp.41-58.
- 小林典子（2015）「SPOT (Simple Performance-Oriented Test)」李在鎬編『日本語教師のための言語テストガイドブック』くろしお出版　pp.110-126.
- 那須川哲哉（2006）『テキストマイニングを使う技術／作る技術　基礎事例と適用事例から導く本質と活用法』東京電機大学出版会
- 樋口耕一（2014）『社会調査のための計量テキスト分析　内容分析の継承と発展を目指して』ナカニシヤ出版
- 松田真希子・宮永愛子・庵功雄（2013）「超級日本語話者の談話特性―テキストマイニングを用いた分析―」『国立国語研究所論集』第5号　pp.43-63.
- 松村真宏・三浦麻子（2014）『人文・社会科学のためのテキストマイニング［改訂新版］』誠信書房
- 森篤嗣（2016）「小学生の話し合い活動に対する評価基準策定のための評価表現の帰納的探索」宇佐美洋編『「評価」を持って街に出よう』くろしお出版　pp.240-254.
- 森篤嗣・齋藤ひろみ・陳楠・フルゲン-マリア-クラウディア-ワカ・嶌田陽子（2012）「テキストマイニングによる外国人児童の作文語彙の分析―日本人児童の作文との比較から―」『社会言語科学会　第30回大会発表論文集』pp.24-27.