

Санкт-Петербургский государственный университет

**ХАЧАТУРОВА Элеонора Артуровна**

**Выпускная квалификационная работа**

**Анализ структуры семантических полей (на материале групп французских синонимов)**

Уровень образования: бакалавриат

Направление 45.03.02 «Лингвистика»

Основная образовательная программа СВ.5055. «Иностранные языки»

Профиль «Французский язык»

Научный руководитель:

д.ф.н., профессор,

Кафедра романской филологии

Марусенко

Михаил Александрович

Рецензент:

д.ф.н., профессор

ФГБОУ «Российский государственный педагогический университет им. А.А.Герцена»

Беляева

Лариса Николаевна

Санкт-Петербург

2022

## **Содержание**

ВВЕДЕНИЕ	2
ГЛАВА 1	5
1.1 АКТУАЛЬНОЕ ПОЛОЖЕНИЕ СЕМАНТИКИ	5
1.2 ЗНАЧЕНИЕ И ЕГО ИНТЕРПРЕТАЦИЯ	7
1.3 ТЕОРИЯ ПОЛЯ	15
1.4 СТРУКТУРА ПОЛЯ	17
ГЛАВА 2	21
2.1 ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА	21
2.1 ИЗМЕРЕНИЕ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ	25
2.2 ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ	27
2.2.1 ТАБЛИЦЫ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ	27
2.2.2 ПОСТРОЕНИЕ ГРАФОВ	30
ЗАКЛЮЧЕНИЕ	35
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ	38

## **Введение**

Настоящая работа посвящена изучению структуры семантических полей прилагательных французского языка с использованием методов обработки естественного языка.

Предметом лингвистической семантики является языковое мыслительное содержание, то есть содержание, заложенное в словах, морфемах, грамматических формах, синтаксических конструкциях естественного языка (Васильев 1990). Многих исследователей волновал вопрос существования в языке системы, связывающей между собой содержание разных понятий, и ее внутреннего устройства. В связи с этим было предложено немалое количество теорий, что на протяжении долгого времени провоцировало споры в научном сообществе, многие из которых и по сей день остаются неразрешенными.

Из необходимости структурировать лексический состав языка родилась теория поля, впервые описанная Йостом Триером в 1939 году, предлагавшая полевой подход к лингвистическому анализу. Теория не раз подвергалась критике, многие исследователи предлагали свой собственный взгляд на поле как на структуру в принципе, велись споры вокруг критериев выделения самых разнообразных полей на уровне грамматики, фонетики, лексики и синтаксиса.

Основная проблема теории поля заключается в отсутствии единого взгляда на основные положения теории и, в сущности, незрелости теории как таковой. Несмотря на большой интерес к теории поля, особенно отмечавшийся во второй половине 20 века, до сих пор отсутствует единый подход к систематизации и формализации лексики, не существует общепринятых критериев выделения полей лексическом уровне. Ввиду особого характера семантики как области изучения и ее тесной взаимосвязи с философией и психолингвистикой, семантика изучена исследователями-формалистами в наименьшей степени.

Между тем, если исследования значения лексических единиц в последнее время все больше сводятся к культурологии и анализу концепта в социуме, то сфера обработки естественного языка стремительно развивается, в том числе в области семантического анализа. Многие машинные методы базируются на

разработанных учеными теориях и успешно адаптируют методы анализа, синтезируя лингвистические принципы с математическими и получая внушительные результаты, уже широко применяемые на практике. Более детальная проработка теории поля могла бы стимулировать разнообразные исследования семантических связей, создание баз данных и разработку новых инструментов для анализа и воспроизведения языка и речи, а также развитие области машинного перевода.

Стоит, однако, отметить, что чаще методы машинной обработки естественного языка применяются на практике с целью извлечения некоей пользы: создания приложений, виртуальных помощников, инструментов для работы с текстом. Довольно обширно данные методы используются в маркетинге, например, при анализе пользовательских запросов в поисковой строке. Гораздо реже методы и модели применяются при непосредственно лингвистических исследованиях в отношении языка как такового в качестве объекта изучения. Учитывая высокую эффективность и мощность существующих разработок, необходимо уделять больше внимания возможному их применению в языкознании.

Применительно к наиболее распространённым языкам, современные методы обработки естественного языка позволяют осуществлять детальный анализ на уровне фонетики, синтаксиса, морфологии, лексики и семантики. Основное внимание в данной работе уделяется именно семантическому анализу отдельных лексем, формирующих семантическое поле одного понятия, что обусловлено направленностью исследования.

Структура работы представляет собой две главы. Первая глава посвящена истокам теории поля, то есть предшествующим ей методам семантического анализа, берущим свое начало в аналитической философии 19 века, и их теоретическому обоснованию, а также последующему развитию теории поля и ее корректировкам. Вторая глава посвящена практической части работы. Она представляет собой обзор актуальных методов машинного анализа естественного языка, необходимых для дальнейшего построения поля: изучается

внутреннее устройство моделей skip-gram и continuous bag of words, на основе которых была разработана открытая библиотека FastText. С помощью данной библиотеки были получены векторные репрезентации синонимов прилагательного «intéressant», затем было измерено семантическое расстояние между данными векторами, и на основе результатов были построены графы, иллюстрирующие семантическое поле понятия.

Основная задача исследования заключается в детальном изучении структуры семантических полей синонимов с помощью синтеза теоретической лингвистики и машинных методов семантического анализа. Немаловажно также произведенное сравнение двух метрик, применяемых для измерения семантического расстояния, поскольку метрики на данный момент изучены недостаточно, при том что широко применяются в разработках последних лет.

## ГЛАВА 1

### 1.1 Актуальное положение семантики

Как отмечает Л.М. Васильев, в последнее время успешно развиваются новые направления семантики: структурная семантика, интерпретирующая семантика, порождающая (генеративная) семантика, концептуальная семантика, коммуникативная семантика, психолингвистическая семантика и прочие направления (Васильев 1990). В настоящей работе необходим обзор наиболее значимых для понимания истоков теории поля направлений.

Несмотря на то, что в своих исследованиях Фердинанд де Соссюр еще не употребляет термин «поле», его идеи послужили основой для большинства дальнейших работ в лингвистике и впоследствии подтолкнули многих авторов к изучению языка с новой, на тот момент революционной точки зрения. Ключевым аспектом его подхода было представление о языке как о системе связанных элементов, заложенное затем в основу структурализма – доминирующей идеи в науке 19го века, адаптированной многими учеными (например, Шарлем Балли, Альбером Сеше и другими членами женеvской лингвистической школы), подтолкнувшей немалое количество ученых к дальнейшим исследованиям, в частности в области семантики.

Наиболее активно структурная семантика начала развиваться в 1960е годы, когда лингвистика как наука существовала в тесном сотрудничестве с философией и логикой. Возросший интерес ученых к данной области был вызван популярностью формального подхода во всех научных направлениях в принципе, а также развитием генеративной грамматики, базировавшейся на успешном формальном представлении фонетического, грамматического и синтаксического строя языка. Формальная репрезентация естественного языка стремилась к поиску универсальных формул и разработке логико-математического аппарата, способного с учетом логических связей передавать отношения, существующие внутри языковой системы.<sup>1</sup>

---

<sup>1</sup> См. Апресян Ю.Д. Исследования по семантике и лексикографии. Том. 1: Парадигматика – М.: Языки славянских культур, 2009. 568 с.

Важно отметить, что идея о переложении структуры естественного языка на язык логики зародилась еще в 19 веке в трудах Готтлоба Фреге, и хоть исследования ученых-логиков в области языка фокусируются по большей части на логических и философских аспектах языка и речи и не были призваны отвечать на вопросы лингвистики, они во многом влияют на развитие различных подходов к изучению языка и часто ложатся в основу различных лингвистических теорий (см. Витгенштейн, Фреге).

Современные методы компьютерного анализа естественного языка во многом базируются на формальном анализе, так как без формализации в том или ином виде машинная обработка языка представляется невозможной, поэтому актуальные разработки в данной области так или иначе основываются на существующих теориях формалистов. Область компьютерных технологий развивается заметно активнее, чем теоретическая семантика как раздел лингвистики, что приводит к тому, что многие машинные методы уходят слишком далеко от лингвистики и не имеют под собой теоретического обоснования. Исследование теорий, лежащих в основе данных методов, позволяет улучшать методы и модели, а также подходить осознанно к полученным с их помощью результатам.

Отметим, что, предположительно, именно семантика труднее всего поддается формализации, поскольку значение лексических единиц чрезвычайно обширно. Этой особенностью лексического состава языка обуславливается тот факт, что долгое время семантика конкретных лексических единиц оставалась за пределами внимания формалистов, и до сих пор отсутствует математический аппарат, способный оперировать значениями слов и в полной мере передать их с помощью формул, в то время как, например, как отмечалось ранее, генеративная грамматика имела немалый успех уже в 80е годы. Вокруг теорий и методов семантической формализации до сих пор ведутся споры среди ученых, и не существует единого общепринятого подхода, подкрепленного стройной

---

Немаловажными оказались также труды Ричарда Монтегю.

теоретической базой. Причина, вероятно, кроется в особенностях языка как явления; некоторые исследователи полагают, что слово само по себе не имеет значения, а наделяется им непосредственно в предложении, поэтому большинство формалистов оперируют именно законченными синтаксическими единицами, а не отдельно взятыми лексемами.

Из генеративной грамматики берет свое начало интерпретативная семантика, о которой шла речь выше. Она возникла как необходимый дополнительный элемент в теории порождающей грамматики и была призвана связать синтаксические конструкции с некоей “глубинной структурой”, содержащей в себе непосредственно семантику предложения (Барбара Парти 2012, Васильев 1990:18-19). Однако же и у этого подхода отмечались свои недостатки, спровоцировавшие немало споров в научном сообществе, остающихся нерешенными по сей день (см. Katz, Хомский, Парти, Васильев).

Показательно также, что первые модели генерации естественного языка базировались именно на синтаксических связях ввиду большей адаптивности синтаксиса к формализации, и даже разработанные в последние десятилетия методы машинной обработки языка, включающие в себя семантический анализ (например, GloVe, Word2Vec), основываются на статистических моделях и оперируют векторными представлениями слов, выходящими далеко за пределы трехмерного пространства и следовательно не поддающимися в полной степени пониманию человеком.

## **1.2 Значение и его интерпретация**

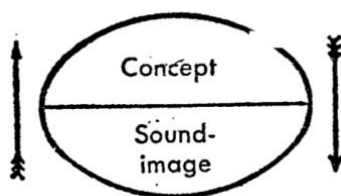
Продолжительное время наблюдались разные взгляды на само явление семантического значения ввиду его многогранности и контекстуальной зависимости. В отечественной лингвистике влиятельными оказались идеи А.А. Потебни, Л.В. Щербы, А.А. Уфимцевой и А.М. Кузнецова, высказавшими сомнения в конструктивности отождествления значения слова с его словарным определением и указавшими на многогранность лексического значения и различные способы его реализации в контексте. Во многом благодаря развитию



этой мысли на настоящий момент устоялось представление о семантике слова как о сложном системно-структурном объединении элементов смысла – сем, разнообразных по манифестационным, комбинаторным, информативно-содержательным свойствам семемы и глубине залегания в структуре (Шипицина 2005).

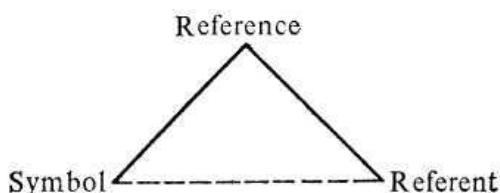
Основные идеи о чрезвычайно важной роли контекста (языкового и внеязыкового) принадлежат Людвигу Витгенштейну, автору так называемой «языковой игры» – системы конвенциональных правил, которыми обуславливается значение в той или иной ситуации. Подробнее о роли контекста в работах Витгенштейна речь пойдет ниже.

Согласно Соссюру, автору знаменитой идеи о двойственном языковом знаке, языковой знак соединяет не объект и его «название», но понятие (или концепт) и звуковую форму слова. Связь между двумя элементами языкового знака обеспечивается наличием в сознании акторов речи некоего общего психологического представления, ассоциации или их набора.



*Рис. 1 Языковой знак*

Свое толкование предлагают Огден и Ричардс, предложившие в 1923 году следующую схему:



*Рис. 2 Семантический треугольник*

«Reference» (референция) в данном понимании есть само понятие, идея, связывающая в сознании говорящего и слушающего референт и его языковое выражение (Palmer 1981). Изучение непосредственно механизма, обеспечивающего данную связь в сознании человека, - задача из области нейролингвистики, психоллингвистики и философии (см Derrida “Структура, знак и игра”, Клод Леви-Штраус)<sup>2</sup>. Для семантики же вышеописанные концепции важны с той точки зрения, что указывают на некую обособленность лексического значения, на его нематериальный, абстрактный характер, и вместе с тем на теснейшую связь понятия с языковой действительностью и звуковым обликом.

Основополагающим утверждением Соссюра является следующее: язык – семиологическая система, все элементы которой неразрывно связаны между собой четкими правилами, где связь между элементами обеспечивает баланс системы. Идентификация слова в данной системе целиком определяется его семантическим значением и неотделима от него (Saussure 1963:110). Аналогичный взгляд встречается у А.А. Потебни: «Язык – система есть нечто упорядоченное, всякое явление его находится в связи с другими. Задача языкознания и состоит именно в уловлении этой связи, которая лишь в немногих случаях очевидна» (Потебня 1973: 243).

Подобно Соссюру датский лингвист Луи Ельмслев (Hjelmslev) в своем подходе исходил из положения о том, что языковые знаки не существуют в изоляции и связаны между собой различными отношениями. Пытаясь структурировать лексический состав языка, в 60е годы исследователь предложил классификацию на основе сопоставления объединенных неким признаком слов, таких как, например, наименования цветов, в разных языках (Palmer 1976:69). Подобные структуры, хоть и критиковались впоследствии разными учеными как слишком упрощенные и не отражающие языковую действительность, также

---

<sup>2</sup> Semiotics and structuralism (Saussure, Levi-Strauss, Derrida) URL: <https://partiallyexaminedlife.com/2012/02/01/now-taking-questions-on-semiotics-and-structuralism-saussure-levi-strauss-derrida/>

непосредственно повлияли на развитие теории поля. В некоторых исследованиях можно встретить трактовку групп, основанных на общности референтов, в качестве полей.

Многогранности лексического значения также посвящено немало исследований. Фундаментальные идеи в данной сфере принадлежат, в первую очередь, Юджину Найда (Eugene Nida) – основоположнику компонентного анализа (1951 г). Исходя из положения о том, что значение представляет собой совокупность элементов, автор предлагает изучать значение посредством анализа его компонентов. Элементы значения (семы) внутри данной совокупности неравнозначны: выделяются доминантные семы и второстепенные, подчиняемые доминантной и организованные как бы вокруг нее. Немаловажно, что данная идея очень тесно перекликается с концепцией семантического поля, за тем только исключением, что семантическое поле обычно стремится упорядочить непосредственно лексическую систему языка, в то время как компонентный анализ упорядочивает сами семы, то есть элементы значения слов (Wunderlich, Nida). Не исключено, что разумно было бы понимать под семантическим полем некоего слова совокупность «микро–полей» его составляющих, то есть ветвящуюся структуру, подобную нейронной сети.

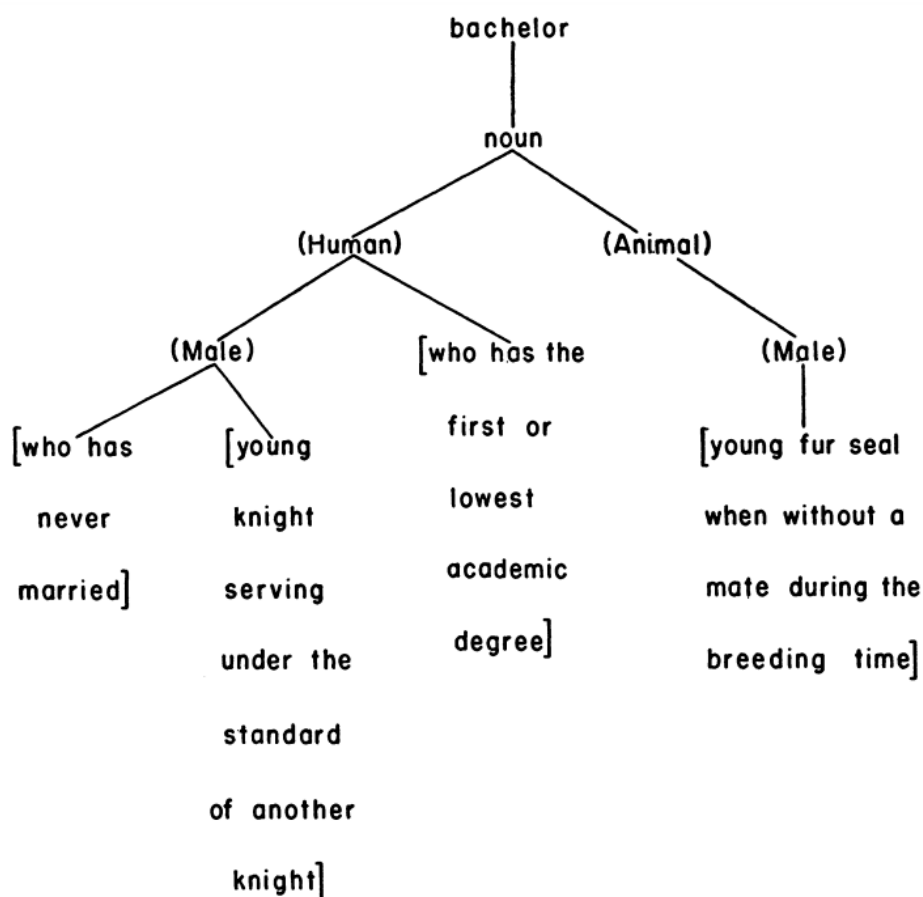
Идея об анализе слова с помощью разложения его на семы подробно описана в работе Каца (Katz) и Фодора (Fodor) «Структура семантической теории»<sup>3</sup>. Авторы предлагают следующий алгоритм компонентного анализа: учитывая часть речи анализируемого слова (его грамматические маркеры) и возможные значения (семантические маркеры), необходимо поэтапно разложить на атомарные элементы смысла каждый более крупный вышестоящий элемент. Получив таким образом семантическую структуру лексической единицы становится возможно сравнивать ее значение со значением других единиц. Две лексические единицы можно считать более или менее синонимичными в

---

<sup>3</sup> Katz Jerrold J., Fodor Jerry A. The Structure of a Semantic Theory. Language, Vol. 39, No. 2. 1963, P. 170-210

зависимости от количества общих элементов структуры (общих ветвей декомпозиции); два понятия расцениваются как идентичные тогда и только тогда, когда их семантические структуры идентичны (Katz, Fodor 1963:185-186). Семантические маркеры призваны как отображать связи между словами, так и подчеркивать идиосинкратические аспекты их значений, то есть фиксировать сходства и различия между ними.<sup>4</sup>

Несмотря на то, что теория Каца и Фодора впоследствии критиковалась (в частности, Ноамом Хомским), она остается фундаментальной теорией в основе



*Рис. 3 Компонентный анализ, представленный в работе Каца и Фодора*

<sup>4</sup> Правила декомпозиции, разработанные авторами, являются частью созданной ими семантической теории, являющейся ярким примером лингвистического формализма и в отличие от многих других формальных теорий не упускающей из виду семантические аспекты. О проблемах подхода декомпозиции см Widyastuti S. Componential analysis of meaning: theory and application или Wunderlich D. Lexical decomposition in grammar.

генеративной семантики. Разговор о ней, однако, уходит слишком далеко от области теории поля.<sup>5</sup>

Принцип композициональности существенно повлиял на современный взгляд на семантическое значение (см Апресян). Примеры компонентного анализа, описанные еще в 1977 году Джеффри Личем (Leech), сегодня возможно воспроизвести с помощью компьютерных методов, основанных на векторах. Сравним пример из Лича и операции, произведенные с векторами путем сложения и вычитания семантических признаков:

components	<i>man</i>	<i>woman</i>	<i>boy</i>	<i>girl</i>
[human]	+	+	+	+
[adult]	+	+	-	-
[male]	+	-	+	-

Рис. 4 Композиционный анализ Джеффри Лича

```
model.most_similar(positive=["king", "woman"], negative=["man"])
[('queen', 0.8523603677749634),
```

Рис. 5 Результат сложения векторов «king» и «woman» и вычитания из результата вектора «man» с помощью векторной библиотеки FastText

Несмотря на некоторые несовершенства, выявленные в данной теории, она находит применение на практике в современных методах машинного анализа. В 2021 году была опубликована работа М. Вакуленко, в которой автор предлагает своего рода компонентный анализ с применением методов обработки естественного языка, суть которого заключается в разложении каждой семы на более мелкие единицы и последующем разложении этих единиц на более мелкие, выделяя таким образом семантические слои анализируемой лексемы (Vakulenko 2021).

Одним из наиболее значимых для последующих исследований в области семантики положений, высказанных Соссюром, является тезис о том, что

<sup>5</sup> См. “лингвистические войны” между Ноамом Хомским и его последователями и сторонниками теории композициональности (George Lakoff, John R. Ross, Paul Postal and McCawley)

значение слова рождается из присутствия других слов в системе и определяется теми отличиями, которые присутствуют между значениями, а также оппозициями между понятиями (Saussure 1963: 114). Из данного заявления берет свое начало понятие «différance» Деррида, определяемое им как «положение о том, что слова и знаки <...> могут получить осмысление только посредством обращения к дополнительным словам, от которых отличаются» («The notion that words and signs can never fully summon forth what they mean, but can only be defined through appeal to additional words, from which they differ.») (Palmer 1976).

Особое внимание автор уделяет именно отличиям между значениями слов, заявляя, что «differences carry signification» – «значение заключается в отличиях». Таким образом, чтобы определить значение слова, необходимо располагать его «окружением» («environment»). Окружение слова не только позволяет определить смысл слова, но напрямую влияет на него: при изменении ближайших «соседей» некоего термина изменяется и смысл этого термина, при том что его звуковая репрезентация и непосредственно связанное с ним понятие остаются неизменными (Saussure 1963:120). В 1957 году Джон Ферс сформулировал данное положение следующим образом: «You shall know a word by the company it keeps» – «Понять слово можно по его компании». Хотя утверждение Соссюра о том, что в языке нет ничего, кроме отличий, можно назвать как минимум экстраполированным, ближайшие элементы лексемы все же критически важны для ее осмысления и по сути являются единственным рациональным методом ее толкования.

В случае, когда происходит изменение ближайших элементов, можно говорить о так называемом семантическом сдвиге или скачке.<sup>6</sup> Природа

---

<sup>6</sup> У В.А. Звегинцева встречается схожее явление, трактуемое как “выключение отдельных слов из системы”, которое может происходить по лексико-семантическим или грамматическим признакам. «В результате резкого или частого изменения значения, например, возникновения нового лексико-семантического варианта и отмирания старого, слово может совершенно уклониться от того значения, которым оно было связано с системой» (Васильев 1990).

семантического сдвига заключается в пересечении, наложении полей друг на друга, вызванном тем, что некоторые элементы, в особенности многозначные и широкозначные понятия, могут входить одновременно в несколько полей, тем самым выступая связующими звеньями между теми полями, членами которого являются. Данное положение вносит коррективы в теорию поля Йоста Триера, о чем пойдет речь далее (Шерстюк, Резникова 2021).

Принято выделять несколько разновидностей системных отношений между лексическими единицами: синонимию, антонимию, омонимию, на уровне плана содержания и гиперонию и гипонию на уровне родо–видовых отношений. Синонимию можно по праву назвать особенной связью, существующей внутри лексического строя языка, поскольку синонимичные связи характеризуются максимальной тождественностью элементов. Так как синонимы обладают схожим, а часто и идентичным планом содержания, Соссюр предлагает определять нюансы значения каждого слова внутри группы синонимов как отличиями, так и общими элементами, которые существуют внутри их значений (Saussure 1963:116).

Компонентный анализ – не единственный подход к структурированию лексики. Некоторые исследователи выделяют также полевой, прототипический и типологический подходы к семантическому анализу, однако лежащие в основе выделения данных подходов методы анализа часто пересекаются и на деле применяются в совокупности. Компонентный анализ представляет особую ценность как доминирующий долгое время в науке подход, по сути лежащий в основе теории поля. Целесообразным представляется изучать лексическую семантику во всем ее разнообразии, не пренебрегая множеством связей, существующих в ее структуре. Полевой подход предлагает именно такой вариант изучения, так как предполагает построение полей на основе некоего общего признака.

### 1.3 Теория поля

Понятие «семантического поля» («semantic field / lexical field») принадлежит Йосту Триеру (Jost Trier). В своей работе он определяет семантическое поле как «группу семантически связанных слов, значения которых определяют друг друга». Опубликованные ранее работы в области семантики не подразумевали непосредственно репрезентацию лексических связей с помощью полей; встречаются термины «функционально-семантическое поле», «лексико-семантическая группа». Апресян, например, выделяет

В своей работе «Теории поля в лингвистике» 1974 года Г.С. Щур изучает историю развития данной теории и сравнивает взгляды разных исследователей на проблему определения семантического поля.

Сам термин «поле» используется во множестве наук и пришел в лингвистику, как отмечает автор, из точных наук, в частности из физики, и значение термина в лингвистике трактуется разными учеными по-разному. Кроме того, термин применяется к разным областям лингвистики и разным уровням языка: к фонетике, морфологии, лексике, грамматике. В зависимости от уровня исследователями были выделены функционально-семантические, морфемные, фонемные, словообразовательные поля, поля множественности, одушевленности, поля модальности и другие (Щур 1974). Немало исследований, граничащих с социолингвистикой и культурологией, посвящены изучению одного конкретного концепта в культуре. Основной проблемой понятия «поле» Г.С. Щур считает его синонимичное употребление с понятиями «структура» и «система» без должных на то оснований, в то время как понятие поля имеет право на существование в том только случае, когда не может быть заменено другим уже существующим понятием и не стремится само заменить его (Щур 1974).

Автор приходит к выводу о том, что в основе объединения любых элементов в поле всегда лежит наличие общего дифференциального признака. Однако, поскольку этот признак может быть экстралингвистическим, а также в роли признака часто выступает скорее функция, данное основание можно



считать как минимум недостаточным. Именно в обращении к экстралингвистическим признакам автор видит также несостоятельность компонентного анализа.

Таким образом, в результате большинства классификаций мы получаем скорее функционально–инвариантные группы (к которым автор также относит группы синонимов). Частным случаем функционально–инвариантной группы Щур считает функционально–семантическую группу (Щур 1974: 96).

Наиболее оправданно, по мнению исследователя, трактовать как «морфосемантические поля» группы элементов, выделенные одновременно по инвариантным и функциональным признакам, причем именно инвариантность служит основным критерием такой классификации; когда инвариантность отсутствует, и единицы языка выделяются согласно их общей функции, полученная классификация есть скорее система, а не поле. Немаловажно, что морфосемантические поля в лексике и в морфологии представляются в некотором роде отличными: в лексике они представляют собой инвариантные группы и в отличие от морфологических групп не связаны общностью морфемного состава и структурных функций. Существует предположение, что в рамках морфологии таким образом возможно выделить макрополя (например, поле модальности), однако в таком случае между элементами поля не будет отношений аттракции и взаимодействия.

Итак, наиболее значимое утверждение Щура звучит так: «В языке категории могут репрезентовать поля, но отнюдь не всякая категория отражает поле, и отнюдь не всякое поле может быть отражено категорией». Для выделения в некоторой категории полей необходимы следующие условия:

- наличие инвариантного свойства;
- репрезентация категорией действительного способа существования элементов в языке;
- суперпозиция (наложение и пересечение классов);
- аттракция (притяжение элементов);
- отсутствие синтагматики.

В особенности автор подчеркивает полевой характер групп синонимов, совмещающих в себе функциональный и инвариантный принципы, в результате чего их элементы обладают общими свойствами и общими функциями. Трактовка синонимов как полей, тем не менее, не совсем правомерна, как отмечает автор, так как для них характерно отсутствие общей коммуникативной функции, что свойственно всем группам элементов с общим семантическим признаком. На основе данных рассуждений предлагается считать фонемные, морфемные, семантические и морфосемантические поля частным случаем инвариантного поля (термин можно считать авторским).

Однако же нет оснований не определять группы синонимов как поля, о чем говорит и сам Щур, а заявление об отсутствии общей коммуникативной функции может быть оспорено хотя бы тем, что данная функция никак не определена и может иметь размытые границы.

#### **1.4 Структура поля**

В теории поля, предложенной Триером, находят свое применение положения структурализма, описанные ранее. Таким образом, теория подразумевает, что значение слова устанавливается из его отношений с другими элементами поля, к которому оно принадлежит. Принадлежащие к разным семантическим полям единицы автор считал не связанными друг с другом. Это положение неоднократно подвергалось критике, так как предполагало существование четких границ между полями, в то время как на уровне лексики, ввиду различных явлений, таких как, например, полисемия и синонимия, говорить об однозначных границах понятия - решение спорное. Границы, однако, должны иметь место, что также отмечалось Щуром, что доказывается хотя бы наличием оппозиций в системе. Предположим, что наиболее оправданно будет осмыслить границы понятия как весьма размытые переходы между соседствующими полями, располагающиеся на периферии поля.

Согласно теории поля Триера, в семантическом поле выделяются ядро и периферия, включающие в себя соответственно наиболее важные для данного

концепта или понятия лексические единицы, формирующие суть понятия, и наименее важные единицы, значением которых, однако, не стоит пренебрегать, так как периферийные элементы не только обеспечивают полноту всех аспектов значения, но также определяют структурные связи, в том числе границы, между разными полями.

По утверждению Васильева, ядерные компоненты образуют инвариантное значение некоего семантического класса слов – ядро семантического поля. Периферийные же компоненты могут неограниченно варьироваться в его составе. Ядерные и периферийные компоненты выделяются не только в составе семемы, но и в составе отдельных ее макрокомпонентов: лексического, грамматического, словообразовательного, денотативного, сигнификативного, коннотативного (Васильев 1990). Утверждение о неограниченном количестве варьирований на периферии поля, по-видимому, не подразумевает бесконечное количество периферийных элементов (ведь лексический инструментарий языка все-таки конечен), но отсылает к нечеткости границ между полями и возможности обнаружения у некоторого понятия самых разнообразных составляющих ее семем. Структура семантического поля таким образом представляет собой конечное количество разветвлений.

Основным недостатком теории Триера принято считать отсутствие четких правил, регулирующих распределение элементов по семантическим полям. Не предложив подобных критериев, автор, однако, положил начало их осмыслению, оставив данный вопрос открытым, что и привело к появлению самых разнообразных взглядов на явление поля и непосредственно самих полей, выделенных на основе множества критериев (с разбором критериев можно подробно ознакомиться в работе Щура).

Решение проблем, отмеченных в теории Триера, предлагается в работе Альфио Глиоззо (Alfio Gliozzo) «Semantic Domains and Linguistic Theory». Автор отмечает сходство положений теории поля с положениями Людвиг Витгенштейна о языковой игре: подобно тому, как, согласно теории поля, слово получает свое значение исходя из элементов поля, в котором находится, в теории

Витгенштейна слова обретают свое значение в том случае, когда могут быть сформулированы с помощью языковой игры, и их значение приравнивается таким образом к их употреблению в той иной ситуации коммуникации. Таким образом, в обоих случаях значение рождается из более широкого контекста. Соответственно, если слова употребляются в не пересекающихся контекстах, несовместимы в определенном ряду похожих ситуаций внеязыковой действительности, то они не могут быть употреблены в одной языковой игре и их контексты, соответственно, отличны, а, следовательно, они не принадлежат к одному полю. Именно так Глиоззо предлагает определять границы между полями - посредством объединения языковой игры и теории поля (Gliozzo 2006).

Стоит отметить, однако, что в понимании автора предлагаемая им «semantic domain» – семантическая область – есть «область обсуждения, например, экономика, политика <...> объединенная лексически. Семантическая область в сотрудничестве с конкретным полем определяется набором специфических для данной области терминов и характеризуется определенными словами, характерными для данной области (domain words), основное свойство которых – употребляться вместе в текстах» (Gliozzo 2006). Таким образом, semantic domain не может быть приравнен к семантическому полю в его привычном понимании, а скорее расширяет идею поля до «области», в которую входят все единицы языка, так или иначе тематически объединенные.

Уход от анализа отдельных лексических единиц к созданию более обширных структур вообще свойственен современной лингвистике. Описанный выше подход находит свое применение в кластеризации текстов и продвигает методы тематического анализа, проявляя себя в этом весьма успешно.<sup>7</sup> В задаче кластеризации и тематического анализа также оказываются важны семантические отношения: например, при объединении разных элементов важную роль играет анализ их гиперонимов. Информация о гиперониме позволяет компьютерным методам лучше понимать связи между терминами и

---

<sup>7</sup> См Hotho A., Staab S., Stumme G Wordnet improves Text Document Clustering, 2003.

соответственно показывать более достоверный результат. На данном подходе строится упомянутый выше WordNet. Также посредством соотнесения определенного набора слов с тематической областью становится возможно присваивать тексту некий тематический лейбл и таким образом решать проблему полисемии (дезамбигуации) с помощью контекста, который несет информацию о языковой ситуации и, следовательно, конкретизирует значение (Gliozzo 2006).

Таким образом, описанный метод, имея в основе теорию поля Триера, позволяет решать актуальные вопросы компьютерной лингвистики. Нерешенным по-прежнему остается вопрос достоверности данных, полученных с помощью такого типа методик – есть ли у нас достаточные основания с теоретической точки зрения подходить именно таким образом к лексической структуре языка? Думается, вопрос останется открытым до прихода ученых к консенсусу в области теории поля и до ее более детальной проработки.

## ГЛАВА 2

### 2.1 Обработка естественного языка

На данный момент сфера обработки естественного языка развивается неравномерно. Ввиду того, что даже родственные языки обладают между собой множеством системных отличий на уровне грамматики, лексики, фонетики и орфографии, не представляется возможным разработать единые инструменты анализа, которые показали бы в равной степени достойные результаты для множества разных языков, вследствие чего методы и модели зачастую создаются или модифицируются индивидуально в соответствии с особенностями конкретного языка.

Английский язык в силу своего распространения не только в мире, но в особенности в области IT остаётся основным в области машинного анализа и располагает наибольшим количеством применимых к нему методов и, следовательно, формируемых с их помощью баз данных. Так, например, WordNet является самым крупным на сегодняшний день ресурсом лексико-семантических данных английского языка и широко используется в лингвистических исследованиях, в том числе в области семантики. WordNet позволяет извлекать из базы так называемые синсеты – списки слов, иллюстрирующих некое понятие и отношения между ним и другими словами (в том числе отношения гиперонимии и гипонимии) – своего рода семантические поля. Всего в базе 117 000 синсетов, преимущественно они иллюстрируют понятия, связанные синонимией.<sup>8</sup>

---

<sup>8</sup> Подобный интерфейс, позволяющий с удобством работать с корпусом, был разработан для крайне ограниченного количества языков, однако, исследователи со всего мира постоянно вносят свой вклад в развитие WordNet и расширение его языковой базы.

WordNet URL: <https://wordnet.princeton.edu/>

Clics (URL: <https://clics.clld.org/>) является довольно значительной базой лексикографических данных. В базе также есть русский язык.

Данная база лексических данных и прочие подобные ей во многом базируются на структуралистском подходе, описанном ранее, так как синсеты опираются на идею о семантической структуре – «network» (Gliozzo 2006).

За последнее десятилетие было разработано немалое число моделей, применимых к самым разнообразным языкам. Так, например, одна из ведущих моделей – FastText – адаптирована для 157 языков, в числе которых диалекты, исчезающие и мертвые языки: западнофризский, нижнесаксонский, латынь, эсперанто.

Построение семантических полей в данной работе производится в несколько этапов. Для полного их понимания необходимо углубиться в процесс получения векторной репрезентации слова.

Обработка естественного языка подразумевает представление анализируемых языковых единиц в подходящем для компьютерного анализа формате. В случае с обработкой лексем наиболее достоверным и результативным представлением служит формат вектора, позволяющий оперировать словом в формате кортежа, состоящего из чисел, и таким образом осуществлять сравнение его с другими словами, определять тематическую направленность текста (что впоследствии применяется для задач, связанных с кластеризацией текстов), выявлять семантические связи, существующие между словами внутри текстового документа. Вектор, используемый в данной работе, был получен с помощью модели Fasttext cc.fr.300.bin. Ввиду удобства использования и наличия различных библиотек методов практическая часть работы написана с помощью языка программирования Python версии 3.8.

Модель была предварительно обучена на текстовых данных, полученных из Википедии и Common Crawl – открытого репозитория текстовых интернет-данных. Исходный вектор считается с помощью составления матрицы с частотой встречаемости каждого отдельного слова в текстовом корпусе. Далее производится расчет условной вероятности встречаемости в тексте сочетания некоторой последовательности слов.

Для обучения модели использовался разработанный в 2013 году исследователями Google алгоритм continuous bag of words (CBOW) – «непрерывный мешок слов» – улучшенная версия представленного ранее алгоритма bag of words (BOW) и в отличие от него учитывающая контекст.

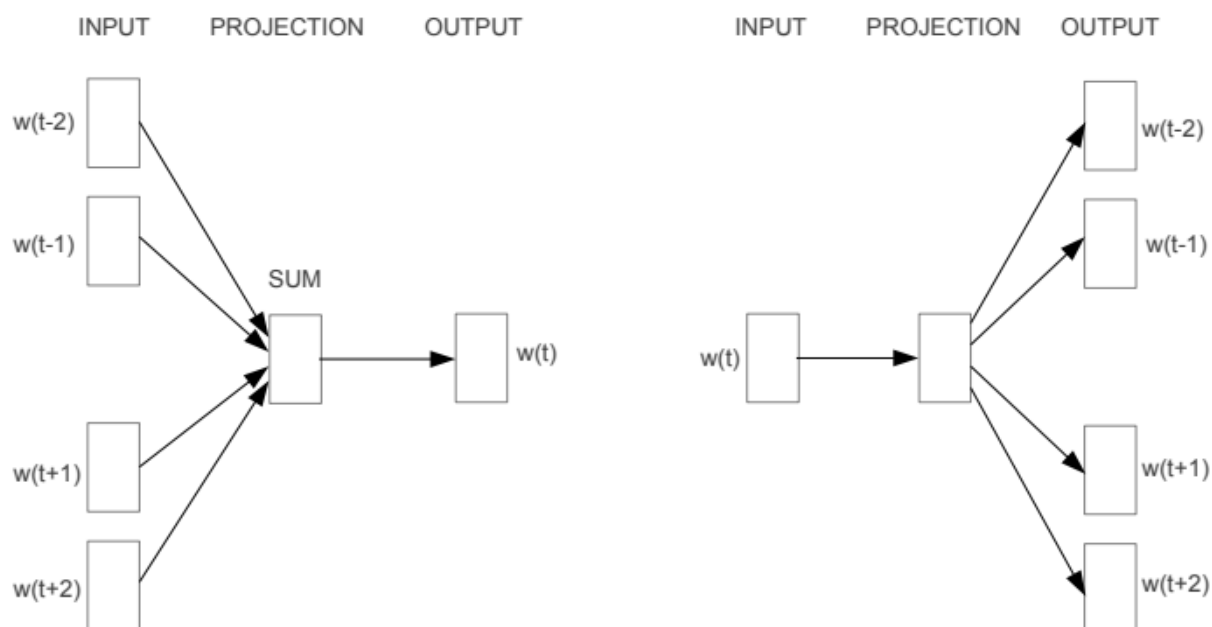


Рис. 5 Схема устройства алгоритмов skip-gram (слева) и continuous bag of words (справа)

Суть алгоритма заключается в приведении слов, стоящих вокруг анализируемого слова (данный параметр поддается изменению в случае, если по какой-то причине нам необходим более широкий контекст; изначально в модели Mikolov анализировались 2 предыдущих и 2 последующих слова (см. Рис. 4)) к общему вектору, который впоследствии способен предсказывать недостающее слово. Иными словами, модель способна предсказывать слово по его контексту посредством вычисления векторов и статистического анализа лексической сочетаемости.<sup>9</sup>

При обучении используемой в данной работе библиотеки метод CBOW был доработан: учитывалось расстояние между центральным словом и

<sup>9</sup> Подробное описание модели в Mikolov T., Wenau Yih, Geoffrey Z. Linguistic Regularities in Continuous Space Word Representations. // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013, P. 746–751.



окружающим контекстом. Таким образом, чем ближе находится некоторое слово к интересующей нас единице, тем больше его «вес» и тем сильнее оно влияет на итоговый вектор. Это изменение сделало модель более точной.<sup>10</sup>

Помимо описанного выше алгоритма CBOW при обучении создателями векторной библиотеки был также использован метод Skip-gram. По сути он представляет собой обратный CBOW алгоритм, предсказывающий сочетание некоторого слова с другими словами на основе подсчета условной вероятности.<sup>11</sup>

Итак, после установки модели необходимо извлечь из нее интересующие нас вектора. В качестве материала была взята группа самых близких синонимов прилагательного «intéressant» согласно исследовательской базе данных Centre national de ressources textuelles et lexicales. Итоговый список прилагательных можно наблюдать на следующем изображении, иллюстрирующем процесс получения векторов с помощью FastText:

```
synonymes = ["intéressant", "beau", "séduisant", "attrayant", "attirant", "brillant",  
             "charmant", "avantageux", "bon", "joli", "plaisant", "captivant", "lucratif",  
             "remarquable", "ravissant", "prenant", "important", "fructueux", "facinant",  
             "curieux", "piquant", "passionnant", "attachant", "étonnant"]  
vectors = {adj: ft.get_word_vector(adj) for adj in synonymes}
```

*Рис. 6 Список анализируемых прилагательных*

Полученные векторы имеют длину 300 и представляют собой список чисел. Каждому слову присваивается собственный уникальный вектор, что впоследствии позволяет производить со словами различные математические операции.

---

<sup>10</sup> Подробное описание процесса обучения модели описано в: Grave E. et al. Learning word vectors for 157 languages. 2018

<sup>11</sup> Описание принципов модели см.: Mikolov T., Kai C., Corrado, G.s, Dean J. Efficient Estimation of Word Representations in Vector Space, 2013 или Bojanowski P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics, 5. 2017

## 2.1 Измерение семантической близости

Подсчёт семантической связи между словами осуществляется с помощью формул подсчета расстояния между векторами-репрезентациями данных слова. Сравнение эффективности данных формул применительно к семантическому анализу на сегодняшний день изучено недостаточно и в основном применительно к кластеризации текстов, кроме того, большинство исследований, посвященных выявлению наиболее достоверного способа измерения семантического расстояния, проводились на основе данных, полученных на материале английского языка<sup>12</sup>. Этим своего рода упущением вышеописанных методов обуславливается необходимость анализа существующих метрик семантической близости с целью определения наиболее достоверной формулы для последующего построения графа.

Итак, основными метриками для пары векторов  $u$  и  $v$  можно назвать следующие две формулы:

Эвклидово расстояние (Euclidean distance):

$$\left( \sum (|u_i - v_i|^2) \right)^{1/2}$$

Косинусное расстояние (Cosine distance):

$$1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$$

---

<sup>12</sup> см Steinbach, Michael & Karypis, George & Kumar, Vipin, A. Comparison of Document Clustering Techniques. Proceedings of the International KDD Workshop on Text Mining. 2000

Huang, Anna. Similarity measures for text document clustering. Proceedings of the 6th New Zealand Computer Science Research Student Conference. 2008

Perreault-Jenkins M. A study of similarity measures for natural language processing as applied to candidate-project matching. 2020

В результате вычислений были получены коэффициенты семантической близости. Для оценки результатов в список векторов также были добавлены вектора для слов «table», «affreux» и «venir». По тому, насколько успешно алгоритм отделит эти семантически не близкие к полю «intéressant» вектора, можно судить о достоверности вычислений. Более того, результаты не должны противоречить здравому смыслу: эмпирически мы понимаем, что слова «beau» и «attrayant» менее близки к «curieux» или, например, «lucratif».

Полученные данные изображены на таблицах в следующем параграфе.

## 2.2 Полученные результаты

### 2.2.1 Таблицы семантической близости

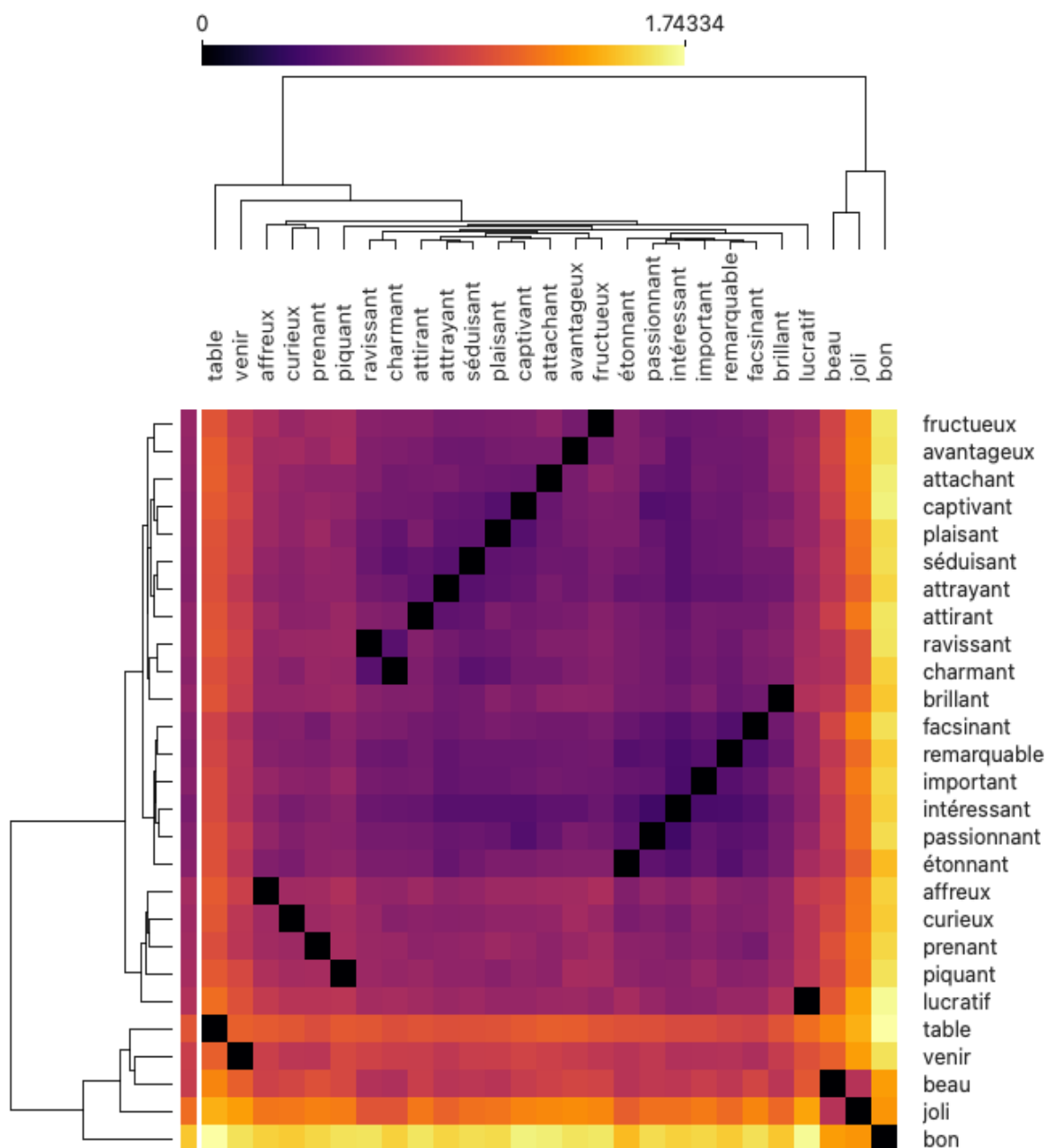


Таблица 1 Семантическое расстояние между векторами, полученное с помощью формулы евклидова расстояния

Цвет иллюстрирует расстояние между понятиями; шкала в заголовке таблицы отображает зависимость цвета от расстояния (расстояние от понятия до самого себя соответственно равно нулю и выделено самым темным цветом). Расстояние иллюстрируется шкалой вверху таблицы: максимальное расстояние, иллюстрирующееся желтым цветом, соответственно, равно 1.74334. Ветви у

границ таблицы отражают объединения векторов по кластерам на основе статистических метрик.

Наблюдения:

- семантическое расстояние между «table» и прилагательными показывает, что с в данном случае модель частично справилась с фильтрацией не близких слов. Близость же слов «venir» и «affreux» к большинству прилагательных оставляет желать лучшего;
- можно наблюдать, что наибольшее количество близких связей выделяется именно у слова «intéressant», что позволяет считать его центральным элементом поля и ядерным понятием;
- статистические алгоритмы классификации показали себя довольно продуктивным методом;
- по всей видимости, особенностью вектора «bon», семантическое расстояние которого со всеми прилагательными в списке довольно большое, является широкосмысленность и многозначность понятия;
- минимальное расстояние позволяет выделить некоторые области значения: например, объединить между собой прилагательные «fascinant», «remarquable», «important», «intéressant», «passionnant», «étonnant» как наиболее близкие друг другу и формирующие определенную область поля;
- еще одна область наблюдается у прилагательных «charmant», «attrayant», «séduisant», «plaisant», «attirant»;
- наиболее близким к прилагательному «intéressant» оказывается прилагательное «passionnant»;

В целом результаты не противоречат здравому смыслу; алгоритмы классификации также показывают себя довольно эффективными в объединении областей значений.

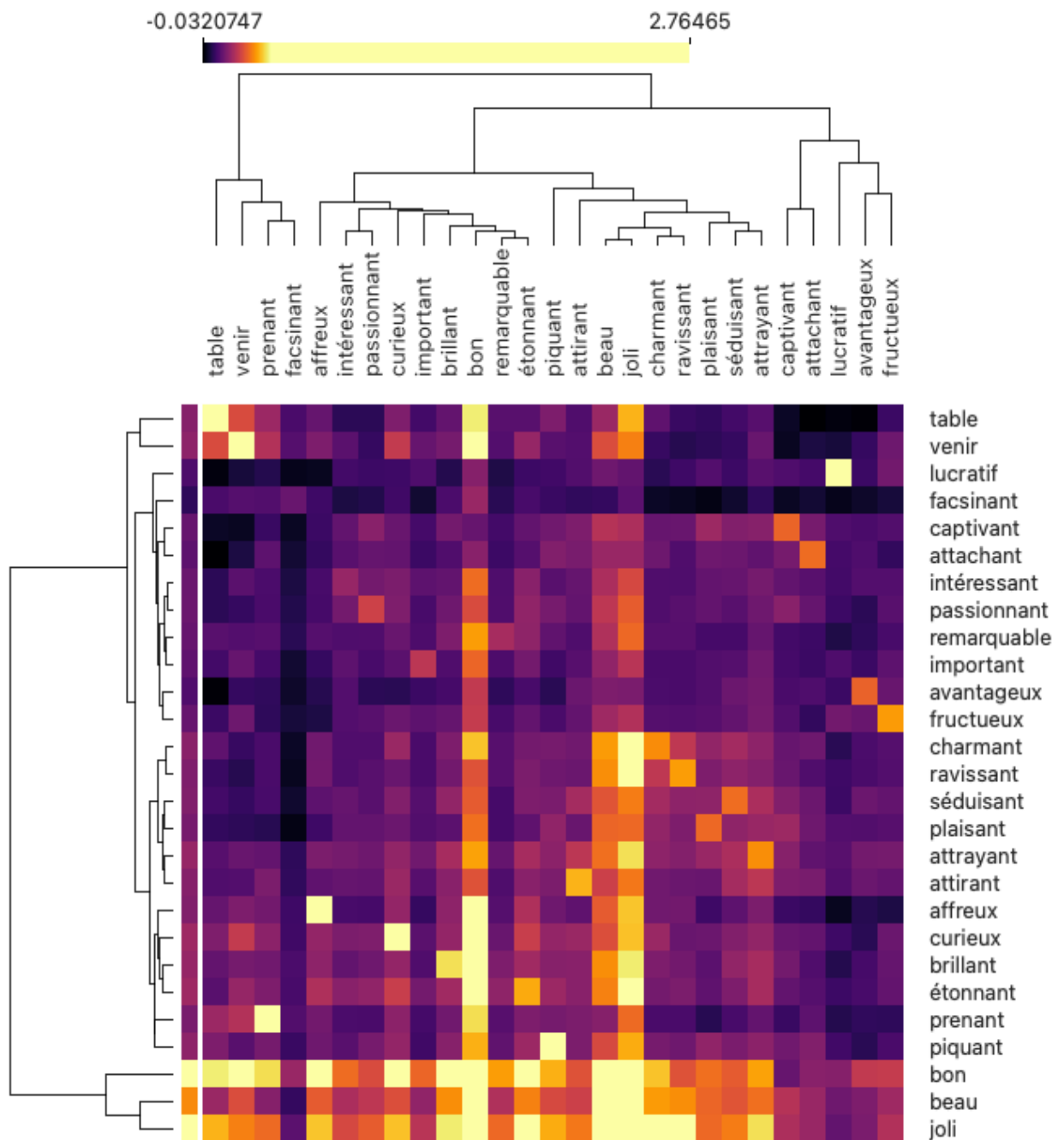


Таблица 2 Семантическое расстояние между векторами, полученное с помощью формулы косинусного расстояния

В случае с косинусным расстоянием (см. Таблица 2) в отличие от предыдущей таблицы анализировать результаты разумнее не столько с помощью расстояния, сколько с помощью изучения пересечений более похожего цвета, поскольку данная метрика учитывает угол между векторами и ее результаты несколько отличны от результатов, полученных в Таблице 1. Также, в данном случае расстояние от вектора до самого себя не всегда является наименьшим

расстоянием в сравнении: некоторые слова оказываются одинаково близки как к самим себе, так и к некоторым другим единицам, что прослеживается в особенности на примере прилагательного «joli».

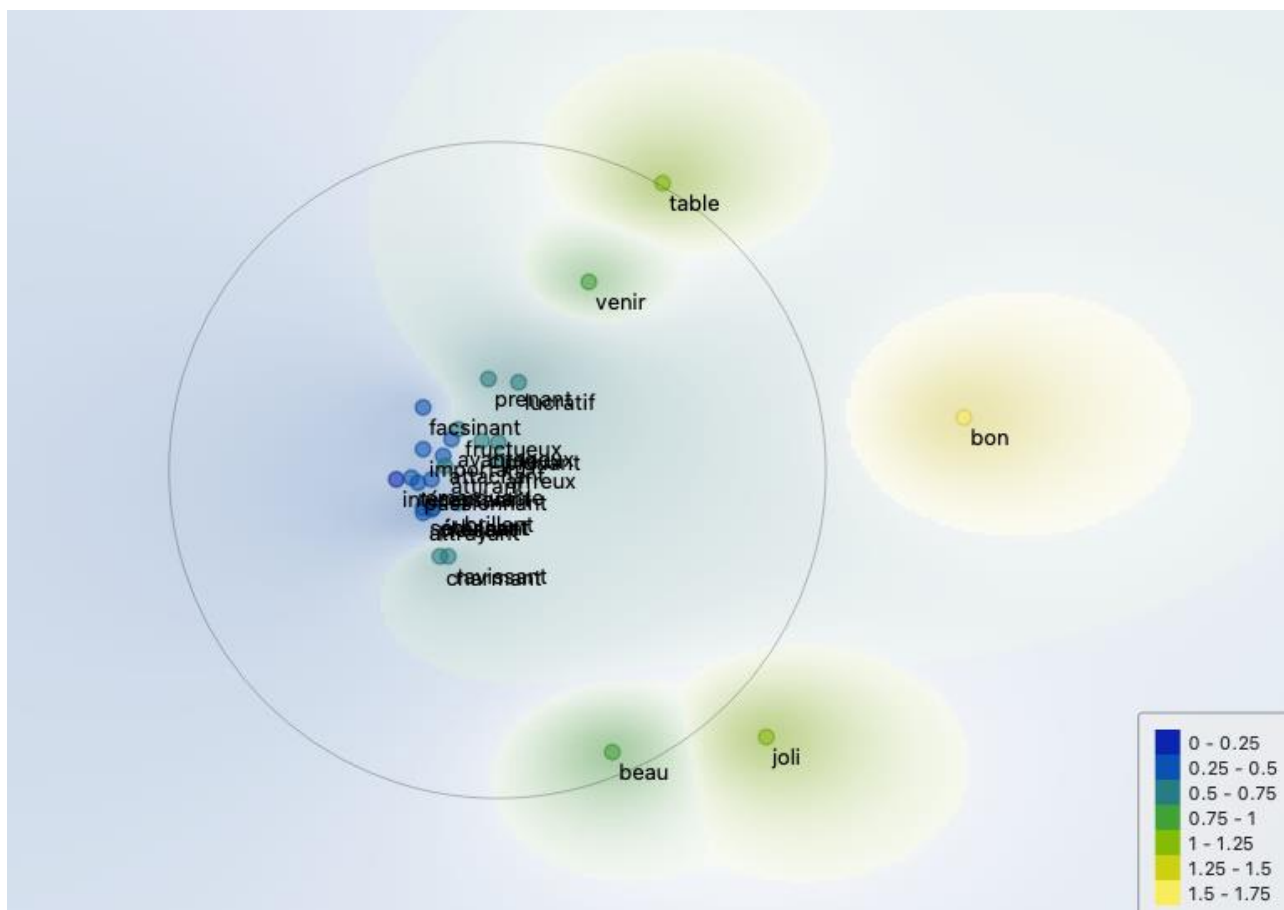
Наблюдения:

- при данном расчете прилагательное «bon» становится значительно ближе к остальным словам в группе;
- «fascinant», в свою очередь, становится значительно дальше от большинства близких ему слов;
- метрика значительно лучше справилась с фильтрацией тестовых слов, нежели эвклидово расстояние, что можно наблюдать особенно отчетливо на примере прилагательного «affreux» и существительного «table»;
- «intéressant», однако, не выделяется в данной классификации как основополагающий элемент;
- несмотря на большую хаотичность результатов, чем в Таблице 1, коэффициенты все же выделяют области значения, например, в случае с «joli», «beau», «bon», «charmant» и «attirant».

В целом, различия между двумя метриками не столь существенные. Стоит учитывать, что изначально была взята группа слов, связанных наиболее тесными отношениями из всех возможных, а именно – синонимией. Это затрудняет задачу сравнения и классификации, поскольку требует применения довольно мощных методов, способных учитывать минимальные отличия и оттенки значения.

### 2.2.2 Построение графов

После анализа полученных в предыдущем параграфе результатов таблицы семантической близости были загружены в программу Orange, инструмент для анализа данных, для получения впоследствии графов зависимости, иллюстрирующих рассчитанное семантическое поле. Результаты приведены ниже.

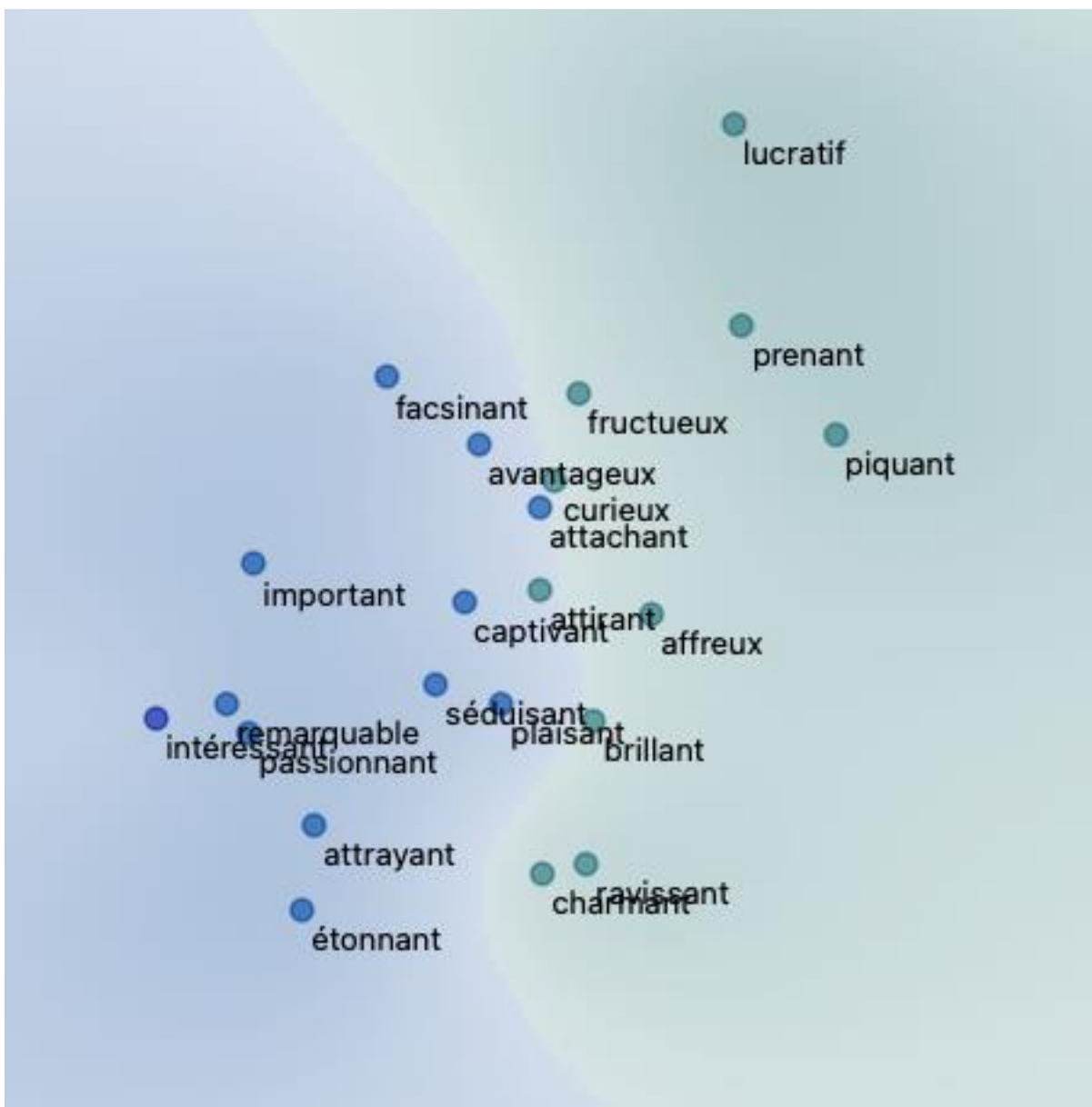


*Граф 1 Семантическое поле, построенное с помощью метрики евклидова расстояния*

Полученный граф показывает, что метрика евклидова расстояния выделила несколько областей значений в исходном списке. Так, отдельными областями изображены «bon», «table», «venir», «beau» и «joli», в то время как основная группа прилагательных находится в центре поля. Граф построен вокруг центрального понятия, которым было выбрано прилагательное «intéressant»; цвет, соответственно, иллюстрирует расстояние остальных слов до центрального понятия. Таким образом, наиболее далекими на графе оказываются области «beau», «bon», «venir», «table», что, в принципе, не противоречит действительности и логике.

Интересным представляется ядро графа, включающее в себя все основные синонимы. Подробное изображение ядра представлено на Графе 2.



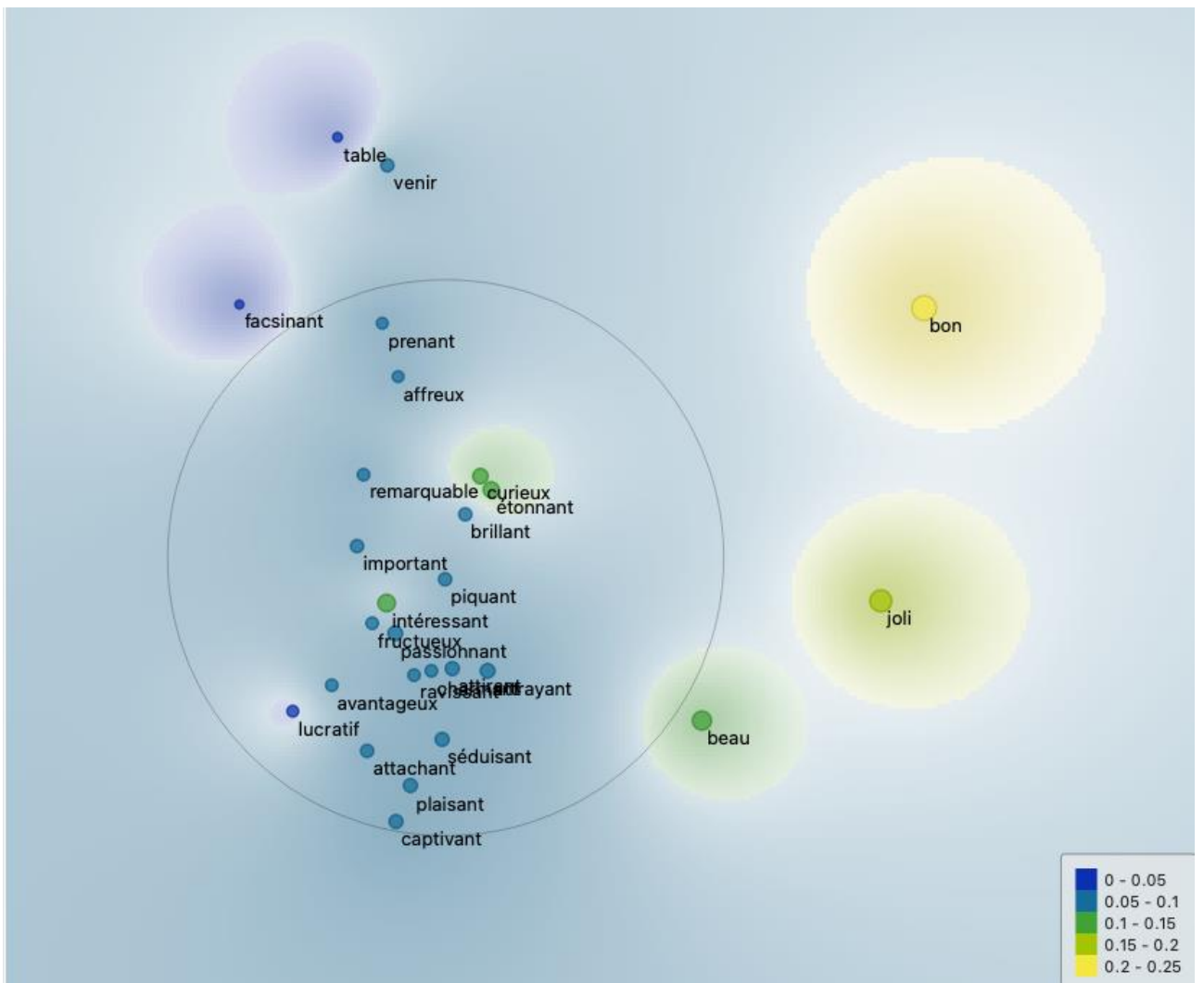


*Граф 2 Ядро поля, изображенного на Графе 1*

«Intéressant» оказывается объединен наиболее тесно с «remarquable», «passionnant», «attrayant», «important», «étonnant», «fascinant», «séduisant», «plaisant», «captivant». Правая половина графа отражает чуть более далекие, но все еще довольно близкие прилагательные.

Характерно, что «lucratif» оказывается крайним ядерным элементом, что можно считать достоверным с логической точки зрения.

Построенное с помощью косинусного расстояния поле изображено на Графе 2.



Граф 3 Семантическое поле, построенное с помощью метрики косинусного расстояния.  
Примечание: чем темнее цвет, тем дальше некое понятие находится от «intéressant».

Как и на Графе 1, можно наблюдать обособленность областей значения «bon», «joli», «beau» и размещение основной части прилагательных внутри ядра поля. Обе метрики также не справились с «фильтрацией» прилагательного «affreux», которое можно наблюдать внутри ядра на обоих графах. «Venir» и «table», в свою очередь, оказываются отделены от основной области поля в обоих случаях, однако, на Графе 2 оказываются определены как максимально далекие по расстоянию

Полученные с помощью данной метрики результаты имеют несколько отличий в сравнении с результатами, полученными с помощью эвклидова расстояния:

- «bon», хоть и расположен дальше всех остальных понятий, имеет наименьшее расстояние до «intéressant»
- ядро поля оказывается менее целостным, его элементы не так тесно связаны между собой, что, однако, не влечет за собой негативных последствий с точки зрения классификации;
- оказываются обособлены области «curieux», «lucratif» и «fascinant»;
- классификация определяет «curieux», «étonnant» и «beau» как находящиеся на одном расстоянии с «intéressant», при этом позиционно «intéressant» находится по центру ядра в окружении ближайших синонимов.

По-видимому, косинусное расстояние менее эффективно справляется с задачами измерения семантической близости, нежели эвклидово расстояние, так как имеет большее количество противоречий, в том числе по отношению к окружающей действительности, что, скорее всего, является результатом полученных ранее в таблице неоднозначных показателей расстояния.

Таким образом, эвклидово расстояние показало себя как наиболее эффективная метрика применительно к семантическим отношениям векторных репрезентаций лексем. На основе полученных с ее помощью результатов был построен граф, иллюстрирующий семантическое поле прилагательного «intéressant», в котором прослеживаются ядерные и периферийные компоненты, а также области значений, соответствующие человеческому представлению о языковой и внеязыковой действительности.

## Заключение

Как говорилось ранее, теория поля находится на довольно незрелом этапе развития. С момента ее первого упоминания в научных работах Триера исследователи уделяли внимание преимущественно недостаткам данного подхода, изредка предлагая доработки в данной области, которые могли бы привести научное сообщество к единой полноценной теории. Работы последних десятилетий, в которых фигурирует теория поля, по-прежнему ссылаются на Триера, так как со времен их научной деятельности теория не претерпела существенных изменений и, соответственно, не имеет более актуальной формы.

Ошибочно было бы полагать, будто семантика находится на недоступном формальному анализу уровне. Многообразие существующих точек зрения на значение как таковое, на его структуру, атомарные элементы и способы формализации, в свою очередь, приводит к тому, что понятийный аппарат лингвистики оказывается не сформирован на желаемом уровне, и исследователи из разных областей языкознания зачастую оперируют одними терминами применительно к различным явлениям, либо при номинации одного и того же явления прибегают к самым разным и иногда необоснованно применяемым понятиям.

Отсутствие сколько-нибудь однородного общепринятого описательного и формального подхода приводит к забвению теории поля в ее первоначальном виде и замыкает структурную семантику в области культурологических и социальных исследований, специализирующихся преимущественно на анализе концептов и явлений окружающей действительности, но не структуры, лежащей внутри лексического состава языка.

Лексический уровень, однако, все же поддается формализации и структуризации, что доказывается успешным применением компонентного анализа на практике, а именно в области машинного перевода, кластеризации текстов и лексикографии. Лексикографические базы данных на сегодняшний день широко применяются в области сопоставительного языкознания, в синхронических и диахронических лингвистических исследованиях, а также

оказываются незаменимыми при переводе, в особенности, когда речь идет о наименее распространенных языках.

Немаловажно, что английский язык, будучи основным языком в сфере компьютерных технологий, существенно превосходит прочие языки по количеству научно обоснованных применяемых методов. Разработка методов для остальных языков в области обработки естественного языка происходит в основном за счет усилий исследователей и программистов со всего мира и зачастую не имеет под собой достойного теоретического фундамента, что также нуждается в корректировке.

Идеи формалистов 20 века, несмотря на отсутствие консенсуса в отношении многих аспектов, все же находят свое отражение, пусть не всегда в связи с намерением разработчиков, в устройстве большинства самых актуальных разработок в области обработки и генерации естественного языка, таких как векторная семантика, синтез речи, анализ тональности текста, классификация текстов и создание виртуальных помощников. Этот факт позволяет уверенно судить о чрезвычайной эффективности синтеза лингвистики и компьютерных технологий, который способен внести существенный вклад в языкознание и значительно продвинуть исследования во многих областях.

Таким образом, теория поля требует доработки. Будучи основной теорией в области структуризации лексического строя языка, она нуждается в полноценном понятийном аппарате и в целостности основных положений и критериев выделения поля, на данный момент весьма абстрактных.

В дополнительном изучении нуждаются также метрики семантического расстояния. Косинусное расстояние широко применяется во многих библиотеках при измерении семантической близости, однако в настоящей работе демонстрирует не самые лучшие результаты. По-видимому, данная метрика более эффективно справляется с бинарными векторами или с задачами, в которых не фигурирует векторное представление каждого отдельного слова. Эту особенность стоит учитывать при последующей работе с готовыми

библиотеками, предлагающими косинусное расстояние в качестве основной метрики.

Эвклидово расстояние, напротив, демонстрирует вполне приемлемые результаты, на основе которых, как показано в данной работе, возможно осуществлять построение семантических графов. Графы позволяют весьма успешно моделировать семантические поля, что могло бы внести вклад в изучение внутреннего устройства лексической системы и существующих в ней связей.

## Список использованной литературы

- 1) Апресян Ю.Д. Исследования по семантике и лексикографии. Том. 1: Парадигматика – М.: Языки славянских культур, 2009. 568 с.
- 2) Васильев Л.М. Современная лингвистическая семантика: Учеб, пособие для вузов. - М: Высш. шк., 1990. - 176 с.
- 3) Потемня А.А. Из записок по русской грамматике // Хрестоматия по истории русского языкознания. – М.: Высш. Шк., 1973.
- 4) Шерстюк А.Ю., Резникова Т.И. О семантической непрерывности: поле ‘толкать’ в славянских языках // Известия Российской академии наук. Серия литературы и языка. Т. 80. No 5, 2021. С. 21–33.
- 5) Шипицына Г.М. Учение А.А. Потемни и современные семантические теории БелГУ, 2005.
- 6) Щур Г.С. Теории поля в лингвистике М.: Наука, 1974. 256 с.
- 7) Wojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5. 2017
- 8) Derrida, J. Structure, sign and play in the discourse of the human sciences // The Languages of Criticism and the Sciences of Man: The Structuralist Controversy, 1970, P. 247-272
- 9) Gliozzo, Alfio. Semantic Domains and Linguistic Theory, 2006.
- 10) Grave E. et al. Learning word vectors for 157 languages, 2018.
- 11) Hotho A., Staab S., Stumme G. Wordnet improves Text Document Clustering, 2003.
- 12) Huang, Anna. Similarity measures for text document clustering. Proceedings of the 6th New Zealand Computer Science Research Student Conference, 2008.
- 13) Katz Jerrold J., Fodor Jerry A. The Structure of a Semantic Theory Language, Vol. 39, No. 2. 1963, P. 170-210
- 14) Lyons J. Semantics. Cambridge University Press, 1977. 371 pp.
- 15) Mikolov Tomas, Wentau Yih, Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. // Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013, P. 746–751.

- 16) Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR, 2013.
- 17) Nida A Componential Analysis of Meaning: An Introduction to Semantic Structures. Mouton, 1975.
- 18) Palmer F. R. Semantics. Cambridge University Press, 1981, 221 pp.
- 19) Perreault-Jenkins M. A study of similarity measures for natural language processing as applied to candidate-project matching, 2020.
- 20) Saussure F. Course in general linguistics, 1963.
- 21) Steinbach, Michael & Karypis, George & Kumar, Vipin. A Comparison of Document Clustering Techniques. Proceedings of the International KDD Workshop on Text Mining, 2000.
- 22) Vakulenko M., From Semantic Metrics to Semantic Fields 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 2021, P. 44-47.
- 23) Widyastuti, S. Componential analysis of meaning: theory and application. // JEE: Journal of English and Education, 4(1), 2016. P. 116–128.
- 24) Wunderlich, D. Lexical decomposition in grammar // The Oxford Handbook of Compositionality, 2012, P. 307-327.

#### Электронные ресурсы:

- 1) Материалы лекции “Формальная семантика как порождение лингвистики и философии” Барбара Парти ,2012 URL: <https://polit.ru/article/2012/05/18/Partee/> (дата обращения: 26.03.2022)
- 2) Semiotics and structuralism (Saussure, Levi-Strauss, Derrida) URL: <https://partiallyexaminedlife.com/2012/02/01/now-taking-questions-on-semiotics-and-structuralism-saussure-levi-strauss-derrida/> (дата обращения: 8.04.2022)



- 3) Key theories of Louis Hjelmslev // Электрон. журнал Literary theory and criticism URL: <https://literariness.org/2018/03/19/key-theories-of-louis-hjelmslev/> (дата обращения 20.04.2022)
- 4) Centre national de ressources textuelles et lexicales // URL: <https://www.cnrtl.fr/> (дата обращения: 16.01.2022)
- 5) Clics Database of cross-linguistic colexifications URL: <https://clics.clld.org>
- 6) WordNet Lexical database for English URL: <https://wordnet.princeton.edu/>
- 7) Документация библиотеки FastText: <https://fasttext.cc/docs/en/support.html>
- 8) Документация библиотеки SciPy: <https://docs.scipy.org/doc/>