СЕТИ КАК ИНСТРУМЕНТ ПОИСКА И НАХОДОК В МУЛЬТИЯЗЫЧНЫХ ПАРАЛЛЕЛЬНЫХ КОРПУСАХ

А. А. Бонч-Осмоловская
Л. В. Нестеренко
(Национальный исследовательский университет
«Высшая школа экономики»)

1. Введение

Цель этой статьи — представить новую методологию анализа семантических полей в области лексической типологии. Исследование лексических систем разных языков предполагает использование определенного набора лингвистических методов сбора и анализа данных. Источником данных на разных этапах служат словари, языковые корпуса, в том числе и параллельные, и, конечно, данные, полученные с помощью анкетирования носителей языка. Собранные из разных языков лексические контексты на следующем этапе требуют обобщения на более абстрактном уровне, для этого используется метод семантического картирования, позволяющий «наглядно представить структуру пространства — в данном случае, пространства определенного семантического поля — и конкретные варианты его организации»[1]. Современный подход к визуализации данных предполагает, что наглядное изображение ценно не только само по себе, но еще и как отображение стоящей за ним математической модели: дискретных соотнесенных параметров и измерителей, которые могут становиться объектами оценки и вычислений. В этом случае открывается множество возможностей применения к исходно неструктурированным языковым данным математических техник и статистистических метрик для получения

[1] Цитата из статьи Рахилина Е. В., Резникова Т. И. Фреймовый подход к лексической типологии // Вопросы языкознания. 2013. Т. 2. С. 3–31.

объективных и надежных оценок и предсказаний и, что особенно важно, новых инсайтов о внутренней структуре модели. Важный шаг в этом направлении был сделан в работе Д. Рыжовой и С. Обьедкова[2], где в качестве альтернативы семантическим картам был предложен метод формальных концептуальных решеток, строгий математический формализм, позволяющий предсказывать возможности и ограничения распределения значений между лексическими единицами.

Техника сетевого анализа параллельных корпусов, предлагаемая ниже, тоже формализует данные как математическую модель, однако она может быть востребована скорее на первом предварительном этапе сбора материала. Ценность сети, как представляется, состоит в том, что она строится автоматически и сразу дает возможность исследователю на раннем этапе выявить наиболее важные фреймы — сделать предварительный замер будущего семантического поля. Таким образом, формируются ожидания относительно структуры семантического поля в целом и конкретных распределений лексических единиц в разных языках, своего рода мостик к формированию анкеты для работы с носителями. Кроме того, конкретные лексемы — вершины сети — получают набор числовых параметров, характеризующих их относительно остальных лексем. Числовые параметры в дальнейшем могут быть использованы для построения семантической карты значений.

Ниже мы опишем суть метода и обозначим перспективы его применения в лексической типологии на материале глаголов ИСКАТЬ и НАЙТИ. Изложение построено следующим образом: в (2) будет рассмотрена специфика ресурсов и инструментов метода. В (3) метод будет опробован на материале глаголов семантического поля ИСКАТЬ-НАЙТИ и предложены некоторые предварительные интерпретации.

2. Описание метода сетевого анализа мультиязычных корпусных данных

Предлагаемый метод состоит из двух этапов. На первом этапе осуществляется извлечение данных из мультиязычного корпуса, на втором этапе данные преобразуются в датасет и визуализируются в виде графовой модели (сети). Рассмотрим последовательно каждый из этапов.

[2] Ryzhova D., Obiedkov S. Formal Concept Lattices as Semantic Maps // Workshop Computational linguistics and language science. CEUR Workshop Proceedings. 2017. С. 78–87.

2.1. Мультиязычный корпус как источник данных для лексической типологии

Параллельные корпуса — корпуса, в которых каждое предложение текста на одном языке соотнесено с его переводом на другом языке, — изначально создавались для машинного обучения автоматических переводчиков. Однако, оказалось, что они предоставляют очень интересный материал для контрастивных межъязыковых исследований. Еще больший потенциал для лингвистических (и прежде всего типологических) исследований имеют мультиязычные корпуса, в которых исходному тексту сопоставлены его множественные переводы на другие языки, при этом все тексты выровнены по предложениям или даже по словам. Поскольку в этом случае мы получаем множество одинаковых контекстов на разных языках, мы можем извлечь данные о распределении языковых единиц (лексических, морфологических, синтаксических) в условиях естественных языковых контекстов, а также получить сведения о том, какие факторы влияют на выбор той или иной единицы в разных языках. Как отмечает Эстен Даль[3], мультиязычные корпуса могут существенно дополнить анкетирование носителей, сняв имеющиеся ограничения этого метода и сохранив его преимущества. Основной проблемой развития в этом направлении является доступ к качественным мультиязычным корпусам. Несмотря на то, что параллельных корпусов создается немало, мультиязычных корпусов пока имеется недостаточно. Среди самых известных оказываются корпуса со стилистически маркированными текстами, например, корпус документов Европейского Парламента, в котором имеются выровненные документы на двадцати европейских языках, или корпус Библии, огромный в смысле мультиязычности, однако весьма специфичный в смысле языкового материала, подчас устаревшего и содержащего множество калек. Тем не менее, нельзя не упомянуть чрезвычайно интересное исследование именно в области лексической типологии, проведенное на корпусе Библейских текстов[4]. Авторы исследуют семантическое поле глаголов движения (go, come, arrive) на материале их употребления в 360 контекстах Евангелия от Марка, переведенного на сто языков. В работе строится вероятностная семантическая карта, в которой так же, как и в традицинных

[3] Dahl Ö. From questionnaires to parallel corpora in typology // STUF-Sprachtypologie und Universalienforschung. 2007. T. 60. № 2. C. 172–181.

[4] Cysouw M., Wälchli B. Parallel texts: using translational equivalents in linguistic typology //STUF-Sprachtypologie und Universalienforschung. 2007. T. 60. №. 2. C. 95–99.

семантических картах, значимой концептуальной единицей является расстояние объектов друг от друга. То, что находится близко, имеет близкие контексты употребления. Распределение одних и тех же контекстов, извлеченных из мультиязычного языкового материала, позволяет вычислить математическую близость лексем между собой и далее отобразить полученные значения на плоскости в виде *дистанции*. Следует отметить, что близкий принцип анализа используется в моделях дистрибутивной семантики, которые также могут использоваться и для типологических исследований лексических систем[5].

Итак, первым этапом предлагаемого метода является извлечение структурированных данных из мультиязычного корпуса. Наличие в корпусе выравнивания предложений дает возможность сформировать списки лексем, использующихся в одинаковых контекстах. Таким образом, мы получаем множество пар лексем с частотными характеристиками для каждой пары. Эти данные используются на следующем этапе для построения графовой модели и ее сетевого анализа.

2.2. Сетевая модель общего семантического пространства

Сети как инструмент хранения и анализа данных достаточно широко применяются именно в гуманитарных науках, в том числе и в лингвистике[6]. Наиболее разработанной областью применения графовых моделей является лексикография — хранение словарной информации. В этом случае вершинами графа являются лексемы, а ребрами графа — отношения между ними. Наиболее известной такой сетью является тезаурус Wordnet, а его мультиязыковым развитием — сеть Babelnet, являющаяся одновременно семантической сетью, толковым словарем, энциклопедией и онтологией на 271 языках[7].

Безусловно, привлекательность словарей и тезаурусов для сетевого анализа состоит в том, что их данные изначально имеют сравнительно

---

[5] Рыжова Д. А. Построение лексико-типологической анкеты с помощью моделей дистрибутивной семантики // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2015. № 3. С. 127–132.

[6] См. например, обзорную монографию Towards a theoretical framework for analyzing complex linguistic networks. Berlin–Heidelberg, 2016.

[7] Navigli R., Ponzetto S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network // Artificial Intelligence. 2012. Т. 193. С. 217– 250.

жесткую структуру — словарные входы естественным образом организуются в связанную систему отношений. Применение сетевого анализа к корпусным данным осложняется прежде всего необходимостью специальной подготовки данных для построения сети. Например, такой подготовкой может стать извлечение множества переводных пар лексем из выровненного мультиязычного корпуса. Тогда каждая лексема будет рассматриваться как вершина графа. Перевод двух лексем в контексте одного и того же выровненного предложения можно будет обозначить как ребро графа, связывающее эти две лексемы. Таким образом множество переводных пар, извлеченных из мультиязычного корпуса, образуют сеть с вершинами и связями, которая может быть визуализирована и проанализирована.

Преимущество сетей состоит в том, что их построение и анализ опирается на разработанный аппарат математической теории графов. Элементы сети (вершины, ребра, кластеры или сообщества) характеризуются набором числовых признаков, определяющих их свойства относительно других элементов. Эти признаки помогают оценить значимость элемента (вершины, узла, какой-то части сети) для сети в целом. Перечислим ниже некоторые из признаков, наиболее релевантных для описываемого здесь метода:

Степень вершины (degree). Самой базовой характеристикой вершины является ее степень — сколько связей имеется у вершины с другими вершинами.

Взвешенная степень вершины. Если связи между объектами имеют числовую оценку (например, как в нашем случае количество контекстов взаимно переведенных лексических пар), то говорят о взвешенной степени вершины — т. е. о параметре, с помощью которого ранжируются связи. В контексте рассматриваемых задач взвешенная степень вершины будет свидетельствовать о широком использовании некоторой лексемы в заданном языке: высокий вес степени означает, что эта лексема часто используется для переводов различных микрофреймов — т. е. она имеет много переводных пар в наиболее частотных контекстах.

Близость, центральность по близости (closeness). Признак центральности по близости характеризует вершину с точки зрения максимума кратчайших путей, связывающих ее с другими вершинами — вершина с высокой степенью центральности по близости связана с максимальным количеством других вершин в сети. В нашем случае, высокий показатель центральности по близости будет свидетельствовать о доминантности лексемы в лексической системе заданного языка. Так, например, если в языке X все рассматриваемое поле покрывается одной лексемой, то это значит, что она участвует в большей части переводов: в сетевом представлении именно к ней будут вести кратчайшие пути от других лексем.

Плотность графа — признак плотности графа показывает отношение между количеством вершин и количеством связей. Если все вершины связаны со всеми, то плотность равна единице. В сообществах с более сложной организацией плотность оказывается ниже.

Кластеры — части сети, характеризующиеся скоплением взаимосвязанных вершин. Кластеризация сети дает возможность выделить части сети (подграфы), элементы которых в большей степени связаны между собой чем с другими вершинами. С точки зрения возможной интерпретации, как правило, вершины одного кластера обладают внутренним сходством. В случае использования сети для лексико-типологического анализа на материале мультиязычного корпуса, кластеры будут показывать распределения лексем между фреймами семантического пространства. Ограниченность и замкнутость связей будет свидетельствовать о том, что именно эти значения общего семантического поля концептуализируются в языках корпуса с помощью специальных лексем и они чаще всего используются для взаимного перевода.

В оставшейся части статьи мы представим пилотное исследование сетевого отображения глаголов семантического поля ИСКАТЬ и НАЙТИ, построенное на материале переводных пар, извлеченных из мультиязычного корпуса. Сначала будет описан метод создания сети, а потом предложены некоторые наблюдения и интерпретации полученных результатов.

3. Кейс *искать/найти*

3.1. Подготовка мультиязычного корпуса и переводных данных.

Для создания мультиязычного параллельного корпуса были взяты тексты семи книг Дж. К. Роулинг о Гарри Поттере на восьми языках: английском, русском, немецком, французском, испанском, итальянском, чешском, болгарском. Общий объем текста для каждого языка составляет в среднем 1,25 млн токенов. В среднем в каждом языке представлено около 80 тысяч предложений. Все тексты были автоматически выровнены друг относительно друга по предложениям при помощи алгоритма Гейла-Чёрча[8]. Таким образом, мы фактически имеем 28 параллельных корпусов, в которые входят любые две пары из рассматриваемых восьми языков и которые в итоге

[8] Gale W. A., Church K. W. A program for aligning sentences in bilingual corpora // Computational linguistics. 1993. Т. 19. №. 1. С. 75–102.

собираются в большой мультиязычный выровненный корпус параллельных текстов. Далее тексты были снабжены морфологической и синтаксической разметкой по стандарту Universal Dependencies версии 2.0 в формате CoNLL-U. Аннотирование производилось автоматически, для это был использован инструмент UDPipe и языковые модели Universal Dependencies.

На следующем этапе были подготовлены словарные фильтры для получения релевантных контекстов. Списки глаголов был составлены следующим образом. Для глаголов *искать* и *найти* был отобран ряд синонимов в русском языке, далее в словарях были найдены их переводные эквиваленты на всех языках, представленных в корпусе. Таким образом, каждому языку был сопоставлен синонимический ряд глаголов полей ИСКАТЬ и НАЙТИ — заданные списки глаголов. При помощи поиска в корпусе для каждого языка собирались контексты употребления соответствующих глаголов его списка.

После этого каждому из полученных контекстов сопоставлялись его переводы на оставшиеся семь языков. Наконец, из всех сопоставленных переводов извлекались лексемы, имеющиеся в глагольных списках каждого из языков. Соответственно, при построении графа были задействованы только парные соответствия глаголов, например,

ENG: *We're gonna try an'* ***find*** *the poor thing.*

RUS: *А этот жив еще, и надо нам с вами его* ***найти****, беднягу.*

и не были учтены случаи, где глагол из одного языка имел другом языке перевод, не являющийся глаголом из списка.

ENG: *We must've been through hundreds of books already and we can't* ***find*** *him anywhere — just give us a hint ...*

RUS: *Может быть, ты хотя бы намекнешь, где нам о нем прочитать?*

Такое ограничение необходимо было для автоматизации процесса выделения пар. Всего было использовано 55 глаголов из исходных списков.

Таким образом, были получены попарные соответствия лексем интересующих нас семантических полей для всех восьми языков, представленных в мультиязычном корпусе. Всего было отобрано 18 110 таких пар глаголов в контекстах. При этом уникальных пар получилось 722. Ниже в таблице 1 и 2 представлены все глаголы восьми языков из исходных списков. Для каждого глагола представлено две характеристики: колонка «количество пар» указывает на то, сколько уникальных пар было получено с этим глаголом — т. е. с каким количеством глаголов из списка мы встречаем данный глагол. Колонка «частотность связей» отражает то, сколько раз во всех контекстах

встречались глагольные пары с данным глаголом. Следует понимать, что один и тот же контекст считался несколько раз для каждых двух переводов. Соответственно если глагол имеет по одному соответствию из списков для каждого языка, то значение ячейки «количество пар» будет равно 7, см.например, болгарский глагол преровя(се). Однако чаще всего, как можно видеть из таблиц 1 и 2, каждый глагол имеет существенно большее количество пар.

Таблица 1. Списки глаголов семантического поля НАЙТИ и их параметрические характеристики

Язык	**Глагол**	**Кол-во пар**	**Частотность связей**
Английский	find	45	3808
	discover	22	259
	detect	8	14
Русский	найти	37	1663
	обнаружить	25	334
	отыскать	8	10
Итальянский	trovare	45	3133
Испанский	encontrar	45	2714
	descubrir	28	385
	registrar	22	111
Немецкий	finden	43	2585
	entdecken	28	219
Французский	trouver	39	2293
	retrouver	35	843
	découvrir	34	515
	apercevoir	25	224
Чешский	najít	49	2486
Болгарский	открия-(се)	50	860
	намирам-(се)	21	243

Данные таблиц 1 и 2 позволяют сделать некоторые наблюдения еще до построения графа. Можно заметить, что поле ИСКАТЬ во всех языках лексически более разнообразно, чем поле НАЙТИ. Важным вопросом является то, насколько вообще эти два поля связаны. Интересно, что в работах по русской семантике вопрос о том, является ли глагол *найти* совершенным видом глагола *искать* обсуждался неоднократно[9]. Однако,

[9] Ср. например о видовой непарности конативного искать и результативного найти Зализняк А. А. Микаэлян И. Л. К вопросу об аспектуальном статусе

Таблица 2. Списки глаголов семантического поля ИСКАТЬ и их параметрические характеристики

Язык	Глагол	Кол-во пар	Частотность связей
Английский	search_en	44	492
	look_for _en	41	1953
	seek_en	26	246
Русский	искать_ru	32	791
	обыскивать_ru	22	80
	обшаривать_ru	19	41
	рыскать_ru	5	5
Немецкий	suchen_de	44	1221
	aufsuchen_de	7	10
	forschen_de	3	3
	aussuchen_de	3	3
Итальянский	cercare_ita	43	1865
	frugare_ita	29	156
	perquisire_ita	25	104
	setacciare_ita	21	30
	ricercare_ita	5	5
Французский	fouiller_fr	33	262
	chercher_fr	33	1152
	rechercher_fr	17	49
Чешский	pátrat	35	321
	hledat	31	996
	prohledat	27	121
	vypátrat	16	44
Болгарский	издирвам_bg	31	143
	търся_bg	30	1114
	претърсвам-(се)_bg	21	57
	затърся_bg	20	107
	претърсвам_bg	16	32
	преровя-(се)_bg	7	7
Испанский	buscar_es	43	1714
	examinar_es	19	83
	escudriñar_es	18	51
	husmear_es	12	16
	rastrear_es	3	3

как показала В. Ю. Апресян (Апресян, этот сборник) в других языках существуют контексты, в которых *искать* и *найти* оказываются взаимозаменяемы. Поэтому было решено построить сеть по общему датасету, состоящему из глагольных пар двух полей.

3.2. Построение графа

Итак, для каждого языка были найдены контексты с употреблением глаголов из изначально заданного списка и соответствия им, т. е. их переводы, во всех других языках выборки. На основании информации о количестве соответствий глаголов друг другу в разных языках был построен граф соответствий. Получившийся граф устроен следующим образом: **вершины** графа — это глаголы исходных списков на разных языках, а **ребра,** связывающие вершины — указание на то, что эти глаголы были использованы в качестве переводных эквивалентов, **вес ребра** подсчитывался следующим образом: было взято отношение количества переводов глагола А глаголом В к общему количеству пар глаголов для двух языков. Т. е. частотность каждой связи нормировалась как доля всех рассматриваемых переводных контекстов между двумя языками.

Общий вид графа представлен на рисунке 1. Лейбл каждого глагола сопровождается сокращенным указанием на язык, к которому он относится[10]. Всего вершин в графе было выделено 55 (по спискам глаголов). Плотность графа равна 0,271. Это значит, что достаточно много вершин связаны далеко не со всеми. В то же время средняя степень вершины равна 18,958. Иначе говоря, в среднем каждый глагол может переводиться с помощью двух-трех глаголов на каждом языке корпуса.

Безусловно, граф такого объема весьма затруднителен для интерпретации в общем виде. Наиболее стандартным решением является выделение сообществ (кластеров). Мы воспользовались функцией Modularity, встроенной в софт для сетевого анализа Gephi. В результате граф был поделен на четыре кластера. Примечательно, что все они соотносятся с определенными семантическим параметрами, различающими фреймы ситуаций ИСКАТЬ и НАЙТИ. Можно предположить, что эти параметры

конативных пар в русском языке: почему искать не может означать ‘найти’? // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). Вып. 15 (22). М., 2016. С. 867–876.

[10] bg — болгарский, cz — чешский, de — немецкий, en- английский, es-испанский, ita — итальянский, fr — французский, ru — русский

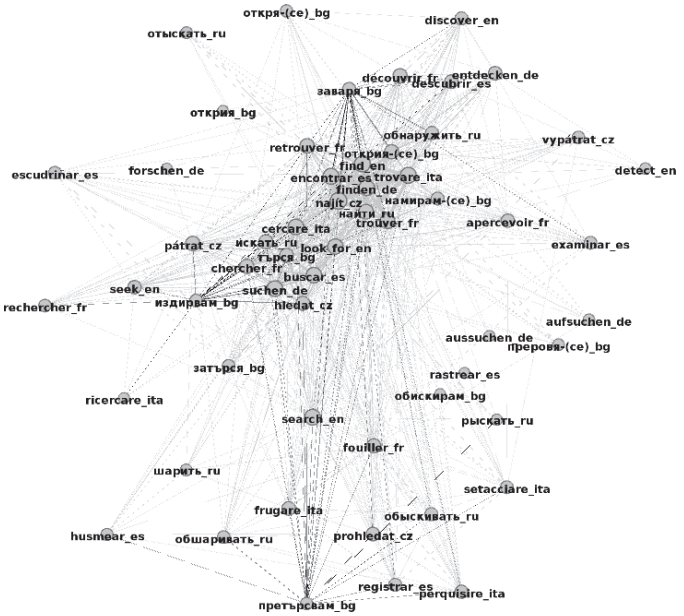

Рис. 1

оказываются принципиально важными для лексической системы в целом, поскольку они последовательно кодируются в разных языках специальными глаголами. Итак, в первый кластер (см. рисунок 2 на с. 316) попали глаголы общего значения, относящиеся к зоне НАЙТИ. Всего к кластеру было отнесено 15 вершин: aufsuchen_de, aussuchen_de, encontrar_es, find_en, finden_de, najít_cz, retrouver_fr, trouver_fr, trovare_ita, vypátrat_cz, заваря-(ce)_bg, намирам-(ce)_bg, находить_ru, отыскать_ru, преровя-(ce)_bg

Кластер имеет высокую плотность — 0,6, это значит, что более половины вершин связаны между собой. Отметим, что значение 1 — т. е. все связаны со всеми — невозможно apriori, поскольку в кластере есть глаголы из одного языка, они не могут быть переводной парой, а значит у них нет связей друг с другом.

В самом кластере отчетливо выделяется ядро — это глаголы высокой частотности с общим значением НАЙТИ. В самом ядре наиболее высокими показателями обладают два глагола — чешский najit, имеющий высокую центральность по близости и немецкий finden, имеющий самую большую взвешенную степень. Это доминантные глаголы в лексических системах чешского и немецкого языков которые имеют наиболее общее значение, поэтому они часто встречаются в самых разных переводных контекстах.

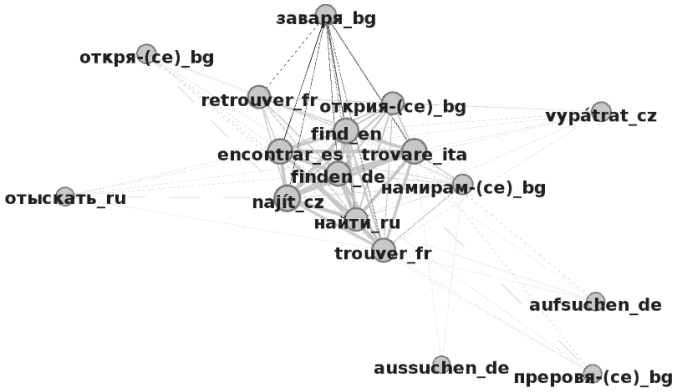

Рис. 2

Во второй кластер (рисунок 3) также попали глаголы, относящиеся к полю *найти*, однако эти глаголы отличает общая особая семантика. Это 8 глаголов (вершин): apercevoir_fr, découvrir_fr, detect_en, descubrir_es, discover_en, entdecken_de, forschen_de, обнаружить_ru. С одной стороны, значения центральных глаголов кластера, связанных между собой, объединяет идея неожиданности находки. Такое значение в целом достаточно часто концептуализируется в языках с помощью специальных грамматических или лексических средств, см., например, статьи в настоящем сборнике: дагестанские языки (Майсак, Даниэль), нанайский (Стойнова), китайский (Холкина, Наний, Цян Сы), сербский (Рыжова, Станкович). Еще одна важная семантическая особенность этой группы глаголов, в которой самой большой взвешенной степенью (0,76) обладает русский глагол *обнаружить,* состоит в специфике отношения к найденному объекту, который не присваивается протагонистом действия[11]. По-видимому, именно благодаря этой семантической характеристике в периферийную зону кластера входят глаголы, определяющие находку как результат интеллектуальной или когнитивной деятельности, — «нашел, потому что думал, замечал, анализировал, исследовал», ср. глаголы forschen_de и detect_en. Речь идет в данном случае о находке абстрактной сущности — свойств объекта, логического вывода, научного инсайта или открытия, т. е. некоторых сущностей, которые принципиально не могут быть присвоены в значении стандартной принадлежности (только X имеет право пользоваться Y, потому

[11] См. Толстая С. М. Глагол найти/находить и его семантические коррелята // Слово и язык. Сб. ст. к восьмидесятилетию академика Ю. Д. Апресяна. М., 2011. С. 338.

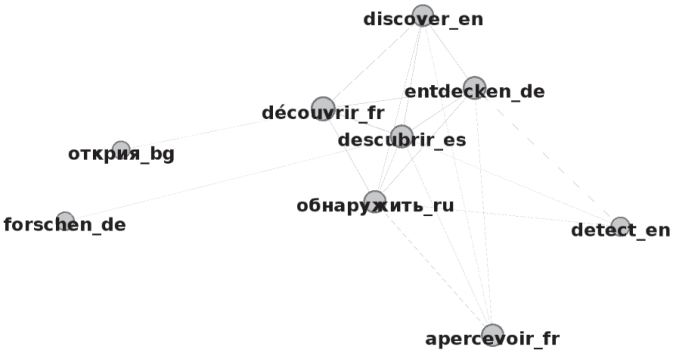

Рис. 3

что это X обнаружил Y). Нельзя не отметить, что именно в этом значении употреблены существительные *поиски* и *находки* в названии этого сборника. Интересно, что все рассмотренные языки, кроме чешского, имеют специальные средства лексикализации этого фрейма. Следует отметить, что второй кластер имеет более низкую плотность, чем первый — 0,522, и очень маленькую среднюю взвешенную степень — 0,108, ср. со средней взвешенной степенью первого кластера, равной 0,901. Маленькая средняя степень узла говорит нам о большом количестве «слабых» вершин — таких, у которых мало связей, и эти связи имеют маленький вес. В нашем случае этому показателю может быть предложена следующая интерпретация: глаголы кластера (фрейма интеллектуального поиска) имеют очень неровную частотность контекстов, часть глаголов встречается часто, а часть редко, кроме того, они реже переводятся регулярным образом (т. е. глаголами исходного списка). Можно предположить, что в результате семантического сдвига исходного значения значительная часть контекстов попадает в другое семантическое поле, например, в поле *смотреть, замечать* (descubrir_es, apercevoir_fr, обнаружить_ru) или в поле *узнавать, расследовать* (forschen_de, detect_en). И здесь мы видим различия лексических систем языков — в некоторых языках, как например, английском, немецком, французском, русском, глаголы из «соседних» полей включаются в семантическое поле НАЙТИ, хотя бы в части контекстов. В других языках вообще нет глаголов этого кластера — не находятся примеры из итальянского, чешского. Мы можем предположить, либо что контексты, связанные с этим кластером, в этих языках переводятся с помощью глаголов общего значения, либо что глаголы переводятся каким-то другим образом, без участия глаголов релевантной семантики, либо, что заданные списки глаголов в этих языках требуют уточнения. Методологически эти

вопросы, возникающие в результате анализа кластера, являются указанием на то, что именно требует более глубокого изучения в дальнейшем. Также для анализа могут быть привлечены контексты, которые легли в основу данных, отображаемых во втором кластере.

Итак, в первые два кластера вошли все глаголы поля НАЙТИ и несколько глаголов поля *искать*. Два других кластера покрывают полностью поле ИСКАТЬ, однако делят его примерно поровну — 16 и 15 глаголов. Рассмотрим, что связывает значения глаголов в каждом кластере.

Так же, как и в случае поля НАЙТИ, два кластера делят глаголы на общие и специфические. В третий кластер попало 16 глаголов, выражающих поиск в самом широком значении: buscar_es, cercare_ita, chercher_fr, escudriñar_es, examinar_es, hledat_cz, look_for_en, pátrat_cz, ricercare_ita, rechercher_fr, seek_en, suchen_de, затърся_bg, издирвам-(се)_bg, искать_ru, търся_bg (см. Рисунок 4).

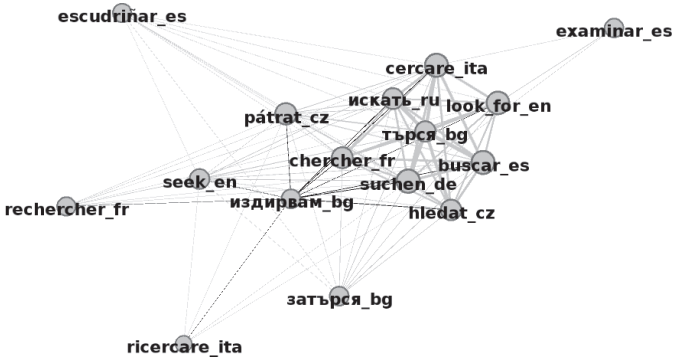

Рис. 4

Интересно, что и по параметрическим показателям, и внешне графы третьего и первого кластера похожи. В центре графа мощные глаголы всех восьми языков, связанные между собой. У них много взаимных связей и очень много общих переводных контекстов. Именно поэтому третий граф имеет высокую плотность, составляющую 0,733 (показатель выше, чем у первого графа) и высокую среднюю взвешенную степень — 0,838 (ср. 0,901 у первого графа). Однако в периферии кластера мы снова сталкиваемся с глаголами со значением интеллектуального поиска — examinar_es, escudriñar_es, rechercher_fr. Таким образом, анализ двух кластеров вместе — второго и третьего — показывает, что в фрейме интеллектуального поиска может быть два фокуса: на процессе поиска и на результате.

В зависимости от выбранного фокуса, глаголы тяготеют к переводным эквивалентам поля НАЙТИ или к полю ИСКАТЬ. Этот пример подтверждает целесообразность рассмотрения двух полей вместе.

Наконец, последний, четвертый кластер (см. рисунок 5) тоже, как уже было сказано, объединяет глаголы поля ИСКАТЬ: search_en, обыскивать_ru, рыскать_ru, шарить_ru, обшаривать_ru, frugare_ita, perquisire_ita, setacciare_ita, registrar_es, rastrear_es, husmear_es, fouiller_fr, prohledat_cz, претърсвам_bg, обискирам_bg. Общим семантическим свойством всех этих глаголов является дополнительное значение активных повторяющихся агентивных действий, необходимых для совершения поисков — ср. обыскивать (серия повторяющихся осмотров конкретных мест/контейнеров с целью поиска), рыскать (серия передвижений с повторением направлений с целью поиска), шарить (серия повторяющихся тактильных движений с целью поиска), husmear_es (вынюхивать — повторяющаяся проверка запахов с целью поиска), rastrear_es (прочесывать — повторяющиеся или параллельные движения, возможно с инструментом, с целью поиска), fouiller_fr (рыть, разрывать — повторяющееся копание с целью поиска). Очень интересен тот факт, что именно это поле собирает метафорические употребления глаголов в значении ИСКАТЬ (рыть, прочесывать, вынюхивать и т.д.). Очевидно, что метафора каким-то образом концептуализирует специфику повторяющихся действий. Действительно некоторые поиски по типу предпринимаемой активности и типам объектов поиска больше напоминают перебор разных запахов, а некоторые — рытье канавы. Дальнейшее описание метафорического ряда требует углубленного анализа свойств противопоставляемых микрособытий поиска и совершенно другого метода исследования.

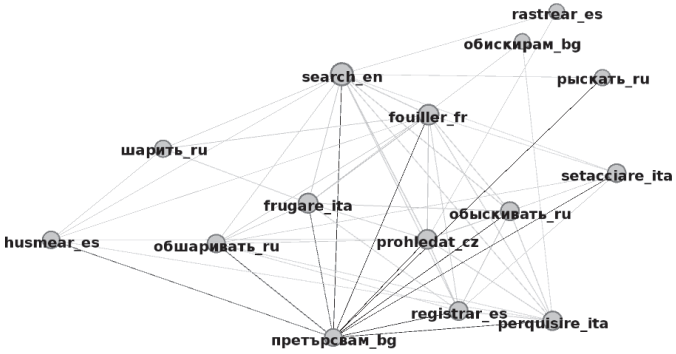

Рис. 5

Возвращаясь к рассмотрению свойств четвертого графа, отметим, что он имеет одинаковую плотность со вторым графом — 0,524. Значение средней взвешенной степени в четвертом графе еще ниже, чем во втором — 0,07. Иными словами, в этом кластере мы имеем дело с разнообразием редких глаголов, каждый из которых различается спецификацией активных действий, предпринимаемых агенсом в процессе поиска. Наибольшим по степени и центральности в этом кластере является английский глагол search. Как показано в (Апресян, настоящий сборник), этот глагол имеет и общее значение поиска, и дополнительное значение «осмотра», хорошо подходящего к фрейму четвертого кластера. Замечательно, что именно этот глагол используется для интернет-поиска, который, видимо, относится не к интеллектуальному поиску второго кластера, но к фрейму повторяющейся серии движений. В той же статье говорится, что в литовском языке для обозначения интернет поиска используется употребленный метафорически глагол *naršyti* ‘рыться, копаться’, что полностью соответствует выделяемому фрейму. Иначе говоря, поиск информации в интернете или в книге (см. Кюсева, Капитонов в настоящем сборнике про язык кунбарланг) является очень хорошим примером итеративно повторяющихся действий, а вовсе не интеллектуального поиска (ср. глаголы второго кластера).

4. Заключение

Выше мы представили результаты сетевого анализа семантических полей ИСКАТЬ и НАЙТИ, осуществленного на материале мультиязычного корпуса. Несмотря на то, что сетевой анализ становится все более широко используемым не только в социальных науках, но и в науках, связанных с анализом языка и текста — лингвистике и филологии, использование этого метода для задач в области лексической типологии является новым. Ценность этого метода состоит, во-первых, в автоматическом анализе значительного числа контекстов — более 18000 в нашем случае, а во-вторых, в построении математической модели, которая помогает ранжировать, сравнивать и противопоставлять результаты. Разумеется, ни один автоматический метод анализа данных не может являться самоцелью, его роль состоит в том, чтобы дать лингвисту новые и функциональные инструменты для поисков и находок.