Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

###### Факультет гуманитарных наук

###### Образовательная программа «Фундаментальная и компьютерная лингвистика»

Пичужкина Ольга Валерьевна

**ЛЕКСИКО-ТИПОЛОГИЧЕСКИЙ ПРОФИЛЬ ФИНСКОГО ЯЗЫКА: АВТОМАТИЗАЦИЯ ПРОЦЕССА СБОРА ДАННЫХ ДЛЯ ЛЕКСИКО-ТИПОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ**

Курсовая работа студента 2 курса бакалавриата группы 162

Академический руководитель

образовательной программы

канд. филологических наук, доц.

Ю. А. Ландер \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Научный руководитель

Д. А. Рыжова

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

« » \_\_\_\_\_\_\_\_\_\_ 2016 г.

**Москва 2017**

Оглавление

[1. Введение](#_Toc481848932)

[2. Обзор литературы](#_Toc481848933)

[3. Этапы работы](#_Toc481848934)

[4. Результаты](#_Toc481848935)

[5. Заключение](#_Toc481848936)

[Литература](#_Toc481848937)

[Приложение 6](#_Toc481848938)

1. **Введение**

Лексическая типология (лексико-семантическая типология, семантическая типология) – раздел лингвистики, изучающий разнообразие семантики лексических единиц (слов или устойчивых конструкций), или же распределение семантических категорий в языках мира. Мария Копчевская-Тамм и её соавторы определяет лексическую типологию как «кросс-лингвистическое исследование категоризации семантических полей» (Koptjevskaja-Tamm et al, 2016). Семантическое поле – группа понятий, которым соответствуют предметы и явления, связанные между собой семантически. При построении предложений человек использует не отдельно взятые слова, а семантические поля, выбирая из поля наиболее уместное слово (Шенна, 2010). Описание семантического поля – то, чем занимается лексическая типология – содержит в себе то, как входящие в него слова могут быть связаны, в каких контекстах могут употребляться и какие оттенки значений имеют. Первыми работами в области лексической типологии были исследование (Morgan, 1871) терминологии родства, (Darwin, 1872) жестов-эмблем и исследования (Magnus, 1877, 1880) обозначений и различения цветов.

Результаты исследования в области лексической типологии считаются достаточно надёжными, если они получены на материале репрезентативной языковой выборки, включающей в себя не менее 100 языков, максимально разнородных ареально и генетически (Bell, 1978). Однако на практике сбор таких данных сильно затруднён. Так как лексикон большинства языков мира достаточно бедно описан (в грамматиках не обсуждаются значения слов, а стандарты описания слов в толковых и переводных словарях недостаточно формализованны), использование словарей как основного материала лексико-типологических исследований практически невозможно (хотя существуют исследования, применяющие такой подход – e. g., Andersen, 1978; Brown, 2005a, 2005b), и требуется работа с носителями. Для сбора данных о сочетаемости лексических единиц типологами используются невербальные стимулы, предлагающиеся носителям (изображения, фотографии, видео, реальные объекты, предметы определённого цвета, вкуса и запаха и т. д.), а также анкеты (от обычных анкет, в которых от носителя требуется перевести словосочетание, до более сложных фреймовых (семантический фрейм – набор прототипических ситуаций и контекстов, описывающих эти ситуации в тексте) анкет, которые и будут рассматриваться в данной работе), которые затем предлагаются носителями для заполнения, часто дополняющиеся корпусными исследованиями моноязычных корпусов и параллельных текстов (переводов одного и того же текста). Так как опрос носителей -- достаточно трудоёмкий процесс, исследователям приходится ограничивать либо размер выборки языков, либо подробность их анализа.

Одним из возможных способов решения этой методологической проблемы может стать частичная автоматизация процесса заполнения уже разработанных анкет (на материале словарей и моноязычных языковых корпусов), не предполагающая работы с носителями. Такой подход объединяет метод анкетирования и корпусные методы сбора материала для лексико-типологических исследований.

Задача данного исследования – разработать алгоритм для автоматического заполнения (перевода слов + поиска встречаемости слов в контекстах, предложенных в анкете) уже существующих лексико-типологических анкет на основе моноязычных корпусов.

1. **Обзор литературы.**

В [Younetal. 2016] рассматривается возможность сопоставления лексики разных языков на основе автоматического анализа данных переводных словарей. В статье исследуется степень семантической близости 22 концептов из списка Сводеша (обозначающих материальные сущности, небесные тела, природные явления и географические объекты) на материале 81 филогенетически и географически отдалённых языков. Данные концепты представляются в виде взвешенного графа, где вершины – концепты, а вес ребра между двумя концептами (он же семантическая близость между ними) – количество языков, в которых существует слово, обозначающее оба концепта (эти слова находятся по переводным словарям – в процессе автоматической обработки словарей каждое слово делится на несколько подзначений, относящихся к тому или иному концепту). В качестве метаязыков для описания подзначений были выбраны английский, испанский, французский, немецкий и русский. Таким образом, мы видим полностью автоматизированный лексико-типологический анализ значительного количества языков.

В [Wälchli, Cysouw, 2013] предлагается способ типологического анализа основных глаголов движения («идти», «входить», «бежать», «приходить» и т. д.) путём автоматической обработки параллельного корпуса текстов (перевод Евангелия от Марка) для 100 языков. Выделяются 360 контекстов, в которых могут быть употреблены глаголы движения, и данная анкета автоматически заполняется, полученные результаты визуализируются (для визуализации использовался метод многомерного шкалирования – способ уменьшить размерность исходного пространства, потеряв как можно меньше информации, подробнее см. [Cox and Cox, 2011]), помогая определить семантическую близость. Таким образом, мы видим автоматизированное заполнение вероятностной семантической карты (близость между двумя контекстами указывает на вероятность того, что в том или ином языке в них будет употреблено одно и то же слово).

В [Рыжова, Ершов, Мельник, 2017] указывается на недостатки существующих методов (словарные данные не очень надёжны в силу плохой сопоставимости словарей, параллельные корпуса переводов одного и того же текста могут быть достаточными для исследования базовой лексики, но использовать их для исследования любой другой семантической зоны крайне затруднительно) и предлагается метод лексико-типологических исследований, включающий в себя разработку анкет, их заполнение и составление семантических карт на основе материала анкет. Утверждается, что каждая из этих задач может быть как минимум частично автоматизирована. Авторы действуют в рамках фреймового подхода [Рахилина, Резникова 2013], разработанного Московской лексико-типологической группой (Moscow Lexical Typology Group, MLexT) и восходящего к работам Московской Семантической Школы ([Апресян 1974]). Этот подход подразумевает сравнительный анализ лексики через сопоставление её сочетаемостных свойств, для чего вводится понятие фрейма – минимальной группы контекстов, на которые может реагировать язык, выделяя для них специальную лексему.

На основе этого метода были разработаны лексико-типологические анкеты для семантических полей “острый”, “размер” и “гладкость” и автоматический способ их заполнения для итальянского языка. Самыми удобными для автоматического заполнения анкет являются прилагательные, так как для определения фрейма прилагательного достаточно учитывать только семантические особенности соответствующего существительного [Рахилина et al, 2010]. В качестве «золотого стандарта» для оценки качества работы алгоритмов использовались уже существующие составленные вручную анкеты для семантических полей «острый», «толстый», «тонкий» и «гладкий» [Кюсева 2012, Козлов et al 2016, Кашкин 2013]. Русскоязычные и англоязычные анкеты использовались в качестве входных данных, италоязычные заполнялись автоматически. По итогам работы авторы сообщают, что их анкеты могут быть автоматически заполнены и для других языков.

1. **Этапы работы.**

На основе анкет Рыжовой, Ершова и Мельник была разработана программа для заполнения лексико-типологических анкет для финского языка.

Использовались лексико-типологические анкеты по профилям "острый", "размер" (прилагательные "тонкий", "толстый", "широкий", "узкий", "тесный" и "просторный") и "гладкость" (прилагательные "скользкий", "ровный", "плоский" "гладкий" и "прямой"), разработанные Рыжовой Д. А., Ершовым И. А. и Мельник А. А (поля «толстый» и «тонкий» были объединены в одно поле «размер»). Роль микрофреймов в этих анкетах выполняют биграммы “прилагательное + существительное”. Анкеты представляют собой таблицы, в строках которой располагаются существительные, а в столбцах изучаемые прилагательные, на пересечении существительных и прилагательных ставится +, если ячейка соответствует зафиксированному сочетанию. Также те же данные представлены в другом виде: в столбец записаны все засвидетельствованные сочетания с контекстом и примером употребления для каждого.

Работа проводилась в два этапа: автоматический перевод существительных через онлайн-словарь, и поиск встречаемости соответствующих словосочетаний через веб-корпус финского языка.

Для автоматического перевода существительных и прилагательных в анкете существуют два метода: машиночитаемые словари и онлайн-переводчики. Первый удобен прозрачностью (мы гарантированно получим нужное нам слово в начальной форме и нужной части речи), второй тем, что позволяет охватить большое количество языков (но непрозрачен и менее точен, чем перевод через словари). Так как анкеты заполняются только для одного языка, был выбран метод машиночитаемых словарей.

Для перевода существительных в анкете был написан краулер по словарю финского языка fi.glosbe.com. Glosbe.com – мультиязычный онлайн-словарь, предоставляющий как переводы, так и примеры употребления каждого слова, объём финско-английской части словаря составляет 360797 примеров и 31495658 предложений. Программа получает на вход файл с существительными в формате txt и на выходе даёт файл с существительными и их переводами в формате txt.

Прилагательные переводились вручную (для профиля «острый» -- все слова для перевода прилагательного из словаря sanakirja.org, для профиля «размер» -- все слова для перевода прилагательных «big» и «small» из словаря sanakirja.org, для профиля «гладкость» -- все слова для перевода прилагательного «smooth» из словаря sanakirja.org). Для перевода выбирались только прилагательные. Sanakirja.org – финско-английский онлайн-словарь. Объём словаря составляет 7800124 слов и 7172581 перевод.

Для заполнения анкет был выбран одноязычный корпус финского языка Araneum Finnicum Minus (http://aranea.juls.savba.sk/guest/run.cgi/corp\_info?corpname=AranFinn\_x&struct\_attr\_stats=1&subcorpora=1). Araneum Finnicum Minus -- корпус, принадлежащий к семейству корпусов Aranea Web Corpora, созданных на движке SketchEngine путём обкачивания (через программу SpiderLing в приблизительно одно и то же время) интернет-ресурсов на разных языках. Корпус снабжен морфологической разметкой. Документы-дубликаты в корпусе удалены, а предложения и абзацы-дубликаты – маркированы. Объём корпуса составляет 228000 документов и 91819745 слов. Запрос в корпус задавался на CQL (Corpus Query Language – язык, позволяющий задавать сложные запросы в корпус), учитывалась лемма, а не словоформа слова. Запрос на CQL выглядит как “[lemma = “прилагательное”][lemma = “существительное”]”. Словосочетание "прилагательное + существительное" вносится в анкету, если набрано минимум 10 вхождений ("золотой стандарт").

Для каждого словосочетания «прилагательное + существительное» в анкете по профилю «острый» в анкету вносился соответствующий контекст.

Все анкеты заполнялись дважды: в стандартном виде (частота совместной встречаемости в корпусе (items per million) на пересечении прилагательного и существительного, если словосочетание «прилагательное + существительное» встречается больше 10 раз) и в другой форме (для каждого словосочетания приводился найденный контекст употребления).

Код программ написан на языке Python 3. Этот язык программирования очень удобен при работе с естественным языком. При работе использовались библиотеки urllib, html (библиотеки для работы со ссылками и html-страницами), re (библиотека для работы с регулярными выражениями – регулярные выражения использовались для поиска переводов слов, количества вхождений, частоты встречаемости и контекста в корпусе, а также для очистки этих данных от html-тэгов) и xlsxwriter (библиотека для работы с таблицами Microsoft Excel). Программа получает на вход файлы с переводами существительных (переводы прилагательных уже включены в код программы) и даёт на выходе файл формата xlsx – анкеты с двумя листами. Первый лист – анкета “1 строка – одно словосочетание” с полями “лексема”, “язык”, “микрофрейм” (в поле “микрофрейм” вносился перевод словосочетания), “контекст на языке” (в это поле вносился перевод существительного), “фрейм”, “таксономический класс”, “поле”, “тип значения”, “пример” (в это поле вносился первый контекст, в котором слово встретилось в корпусе) и “комментарий”. Поля “фрейм”, “таксономический класс” и “тип значения” оставлялись пустыми. Второй лист – анкета в стандартном виде.

1. **Результаты.**

Для оценки точности перевода слов каждое слово сравнивалось с его переводом из Google Translate. Для существительных по семантическому профилю «острый» доля правильных переводов: 0.9, доля неправильных: 0.028 (неправильно переведён «кол» -- как seiväs («полюс»), «каблук» -- как korko («интерес»), «взгляд» -- как kanta («акции»), «нота» -- как nuotti («примечание»)), доля слов, для которых не нашлось перевода: 0.071 (слова «носок ботинка», «бесцеремонность», «цинизм», «лечо», «недосып», «конфликт», «синева», «аппендицит», «впечатлительность», «глаза»). Для существительных по семантическому полю «размер» доля правильных переводов: 0.941, доля неправильных: 0.029 (неправильно переведено «полоска» -- как juova («рубец»)), доля слов, для которых не нашлось перевода: 0.029 (слово «туфли»). Для существительных по семантическому полю «гладкость» доля правильных переводов: 0.863, доля неправильных: 0.015 (неправильно переведено слово «руки» -- как käpälöidä («лапа»)), доля слов, для которых не нашлось перевода: 0.121 (слова «щёки», «ступеньки», «поверхность стола», «поверхность озера», «стулья», «деревни», «промежутки», «течение реки»). Также некоторые переводы оказались дубликатами, например, “путь” и “дорога” (tie).

Данные анкет показывают, что:

1. Для поля «острый» очень малая часть переводов прилагательных используется в финском при описании соответствующих семантических полей -- 6 прилагательных из 17 (terävä, jyrkkä, tarkka, kirpeä, kärjekas и ankara). В большей части контекстов, связанных с острыми объектами, употребляется прилагательное terävä, с острыми поверхностями – прилагательное jyrkkä, с человеком, чувствами и мыслями – прилагательное tarkka. Количество словосочетаний с остальными прилагательными слишком мало, чтобы делать утверждения об их значении. Необычайно частотны выражения «острая картинка» (tarkka kuva) – 2.12 словоупотреблений на миллион в корпусе, «острое наказание» (ankara rangaistus) – 1.43 словоупотреблений на миллион и «острое исследование» (tarkka tutkimus) – 1.22 словоупотреблений на миллион.
2. Поле “размер” заполнено плотнее всего, так как словосочетания, обозначающие размер, самые частотные. Самые частотные словосочетания – со словом “человек” (“iso ihminen – 0.41 словоупотреблений на миллион в корпусе, “aikuinen ihminen” – 3.7, “suuri ihminen” – 0.49, “pieni ihminen” – 4.5 и “nuori ihminen” – 3.9), что связано с многозначностью словосочетаний “большой человек” и “маленький человек”. Наибольшее количество прилагательных, с которым может сочетаться существительное, опять же у слова “человек” (5 из 8). Наиболее широкий охват значений у прилагательных iso (“большой”, 11 словосочетаний), suuri (“большой”, 10 словосочетаний) и pieni (“маленький”, 14 словосочетаний).
3. Для поля «гладкость» найденных словосочетаний меньше всего (12). Количество словосочетаний с прилагательными слишком мало, чтобы делать утверждения об их значении.

Общее число найденных словосочетаний очень мало (77 из 2996 возможных). Возможно, это связано с тем, что был выбран слишком высокий золотой стандарт встречаемости, и нужно было записывать словосочетание в анкету, если оно встречалось минимум 5 раз, а не 10. Возможно, следовало взять более объёмный корпус (Araneum Finnicum Majus с платным доступом содержит в себе 1200000 486 токенов и 817453523 слова).

1. **Заключение.**

Итак, мы провели эксперимент по автоматизации заполнения лексико-типологических анкет с семантическими полями «острый», «размер» и «гладкость», разработанных Рыжовой, Ершовым и Мельник. Задача автоматического заполнения анкеты состояла из двух этапов: перевод слов и поиск встречаемости (и частотности) слов в контекстах, указанных в анкете. Перевод был проведён через словарь sanakirja.org с достаточной точностью. Для второго этапа работы использовался корпус финского языка Araneum Finnicum Minus. Результаты показывают, что автоматическое заполнение лексико-типологических анкет возможно, хотя, безусловно, предложенный метод не автоматизирует полностью лексико-типологические исследования. Однако же, учитывая недостатки существующих источников данных для лексико-типологических исследований, описанные во введении, стоит считать это достоинством метода, а не недостатком. В дальнейшем данный способ сбора данных можно расширить на другие языки (семейство Aranea Web Corpora содержит в себе корпуса для английского, чешского, болгарского, французского, немецкого, испанского, венгерского, итальянского, нидерландского, польского, русского, китайского и словацкого языков). При дальнейших разработках можно усовершенствовать данный метод для работы с более сложными анкетами, например, анкетами для лексико-типологического анализа существительных и глаголов, которые требуют более большой и сложный набор контекста для однозначного определения, к какому фрейму относится слово.

1. **Литература.**

Andersen, E., 1978. Lexical universals of body-part terminology. In: Joseph H. Greenberg, ed., 1978. *Universals of human language.* Stanford: Stanford University Press, pp. 335–368.

Апресян Ю.Д. Лексическая семантика: синонимические средства языка. М.: Наука, 1974.

Bell A., 1978. Language samples. Universals of Human Language, Method and Theory. Greenberg J.H., Ferguson C.A., Moravcsik E.A. (eds.). Stanford Univ Press, 2. Palo Alto, CA. Vol 1: 123–156.

Benko, Vladimír: Aranea: Yet Another Family of (Comparable) Web Corpora. In *Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655.*Springer International Publishing Switzerland, 2014. pp. 257-264. ISBN: 978-3-319-10815-5 (Print), 978-3-319-10816-2 (Online).

Benko, Vladimír: Compatible Sketch Grammars for Comparable Corpora. In *Andrea Abel, Chiara Vettori, Natascia Ralli (Eds.): Proceedings of the XVI EURALEX International Congress: The User In Focus. 15–19 July 2014.* Bolzano/Bozen: Eurac Research, 2014. pp. 417-430. ISBN 978-88-88906-97-3.

Brown, C.H., 2005a. Hand and arm. In: M. Haspelmath et al., eds., 2001. pp. 522-525.  
Brown, C.H., 2005b. Finger and hand. In: M. Haspelmath et al., eds., 2001. pp. 526-529.

Cox T.F., Cox M.A.A, 2011. Multidimensional Scaling. Chapman and Hall.

Darwin, C. 1965 [1872]. The expression of the emotions in man and animals. Chicago, London: University of Chicago.

Кашкин Е.В. Языковая категоризация фактуры поверхностей (типологическое исследование наименований качественных признаков в уральских языках). Дисс.канд. филол. наук. М.: МГУ, 2013.

Козлов А.А., Кухто А.В., Привизенцева М.Ю. О реальности семантического поля: лексико-типологический подход // В кн.: ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН / Отв. редактор Н. Н. Казанский. Т. XII. Ч. 2. Материалы Десятой конференции по типологии и грамматике для молодых исследователей (2013 г.) / Отв. ред. тома Д.В. Герасимов. – СПб.: Наука, 2013. С.: 522-533.

Koptjevskaja-Tamm M., Rakhilina E., and Vanhove M. "The semantics of lexical typology", Rimer, Nick (ed.), The Routledge Handbook of Semantics.

Кюсева М.В. Лексическая типология семантических сдвигов названий качественных признаков ‘острый’ и ‘тупой’. Дипломная работа. М.: МГУ, 2012.

Magnus, H. 1877. Die geschichtliche entwicklung des farbensinnes [The historic development of the color sense]. Leipzig: Viet.

Magnus, H. 1880. Untersuchungen ueber den farbensinn der naturvoelker [Investigations on the color sense of the primitive peoples]. Jena: Fraher.

Moore R., Donelson K., Eggleston A. and Bohnemeyer J., 2015 "Semantic typology: New approaches to crosslinguistic variation in language and cognition", Linguistics Vanguard 1(1): 189–200

Morgan, L. H. 1871. Systems of consanguinity and affinity of the human family. Washington DC: Smithsonian Contributions to Knowledge.

Рахилина Е.В., Резникова Т.И. Фреймовый подход к лексической типологии. Вопросы языкознания. – 2013 – No 2: 3-31.

Рахилина Е.В., Резникова Т.И., Карпова О.С. Семантические переходы в атрибутивных конструкциях: метафора, метонимия и ребрендинг // Рахилина Е.В. (ред.). Лингвистика конструкций. М.: Азбуковник, 2010.

Rychlý, Pavel: Manatee/Bonito – A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing.* Brno: Masaryk University, 2007, pp. 65-70. ISBN 978-80-210-4471-5.

Рыжова Д.А., Ершов И.А., Мельник А.А., 2017 "Автоматический сбор данных для исследований по лексической типологии", Проблемы компьютерной лингвистики и типологии.

Шеина, И. М. (2010). Лексико-семантическое поле как универсальный способ организации языкового опыта. *Вестник московского государственного областного университета*, 69-72.

Walchi B., Cysouw M., 2012 "Lexical typology through similarity semantics: Toward a semantic map of motion verbs

Youna H., Suttond L., Smithc E., Moorec C., Wilkinsc J. F., Maddiesong I., Croftg W., and Bhattacharyac T., 2016. "On the universal structure of human lexical semantics", PNAS

1. **Приложение.**

Код программ расположен по адресу: <https://github.com/vyhuholl/kr/blob/master/araneum%20crowler.py> и <https://github.com/vyhuholl/kr/blob/master/translate%20nouns.py> .

Списки слов расположены по адресу: <https://github.com/vyhuholl/kr/blob/master/nouns%20sharp.txt> , <https://github.com/vyhuholl/kr/blob/master/nouns%20size.txt> и <https://github.com/vyhuholl/kr/blob/master/nouns%20smooth.txt>.

Таблицы с результатами расположены по адресу: <https://github.com/vyhuholl/kr/blob/master/questionnaire_sharp.xlsx>, <https://github.com/vyhuholl/kr/blob/master/questionnaire_size.xlsx>, <https://github.com/vyhuholl/kr/blob/master/questionnaire_smooth.xlsx> .