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 Коптьевская-Тамм определяет лексическую типологию как «кросс-лингвистическое исследование категоризации семантических доменов» (Koptjevskaja-Tamm et al, 2016). 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 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.

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 для 100 языков.

В [Рыжова, Ершов, Мельник, 2017] предлагается метод лексико-типологических исследований, включающий в себя разработку анкет, их заполнение и составление семантических карт на основе материала анкет.

На основе этого метода был разработан автоматический способ заполнения анкет на материале корпуса финского языка.

1. **Этапы работы.**

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. Анкеты представляют собой набор

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.

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 путём обкачивания интернет-ресурсов на разных языках. 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. 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 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. При работе использовались библиотеки urllib, html (библиотеки для работы со ссылками и html-страницами), re (библиотека для работы с регулярными выражениями) и xlsxwriter (библиотека для работы с таблицами Microsoft Excel).

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»).

Получившиеся анкеты:

Данные анкет показывают, что

1. **Заключение.**
2. **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.

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.

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.

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

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

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.

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

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> .