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 – раздел лингвистики, изучающий разнообразие семантики лексических единиц (слов или устойчивых конструкций) в языках мира.

Результаты исследования в области лексической типологии считаются достаточно надёжными, если они получены на материале репрезентативной языковой выборки, включающей в себя не менее 100 языков. Однако на практике сбор таких данных сильно затруднён. Для сбора данных о сочетаемости лексических единиц типологами используются анкеты, которые затем предлагаются носителями для заполнения.

Одним из возможных способов решения этой методологической проблемы может стать частичная автоматизация процесса заполнения анкет, не предполагающая работы с носителями.

1. **Обзор литературы.**

В [Younetal. 2016 ] рассматривается возможность сопоставления лексики разных языков на основе автоматического анализа данных переводных словарей,

В [Wälchli, Cysouw 2013] предлагается способ типологического анализа основных глаголов движения путём автоматической обработки параллельного корпуса текстов.

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

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.

Для перевода существительных в анкете был написан краулер по словарю финского языка fi.glosbe.com.

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 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. 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 (+ на пересечении прилагательного и существительного, если словосочетание «прилагательное + существительное» встречается больше 10 раз) и в другой форме (для каждого словосочетания приводился найденный контекст употребления).

1. **Результаты.**
2. **Заключение.**

**Литература.**

Randi Moore, Katharine Donelson, Alyson Eggleston and Juergen Bohnemeyer, "Semantic typology: New approaches to crosslinguistic variation in language and cognition", Linguistics Vanguard 2015; 1(1): 189–200

Maria Koptjevskaja-Tamm, Ekaterina Rakhilina, and Martine Vanhove To appear in Rimer, Nick (ed.), The Routledge Handbook of Semantics, "The semantics of lexical typology"

Hyejin Youna, Logan Suttond, Eric Smithc, Cristopher Moorec, Jon F. Wilkinsc, Ian Maddiesong, William Croftg, and Tanmoy Bhattacharyac, "On the universal structure of human lexical semantics"

Bernhard Walchi, Michael Cysouw, "Lexical typology through similarity semantics: Toward a semantic map of motion verbs"

Рыжова Д.А., Ершов И.А., Мельник А.А., "Автоматический сбор данных для исследований по лексической типологии"

**Приложение.**

Код программ расположен по адресу: <https://github.com/vyhuholl/kr/blob/master/краулер%20по%20финскому%20корпусу.py> и <https://github.com/vyhuholl/kr/blob/master/прога%20курсач.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: