Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2023»

Москва, 14–17 июня 2023 г.

Development of a Morphological Analyser for Siberian Ingrian Finnish

Ivan Ubaleht
Omsk State Technical University /
644050, Russia, Omsk, Mira 11
ubaleht@gmail.com

Abstract

This paper presents our work on the development of a morphological analyzer for Siberian Ingrian Finnish. Siberian Ingrian Finnish is a low-resource language. In this paper, we present an algorithm for analyzing nouns of Siberian Ingrian Finnish and show an example of analysis.

Keywords: morphological analyzer; low-resource language; natural language processing; Siberian Ingrian Finnish
DOI: 10.28995/2075-7182-2023-22-1127-1132

Разработка морфологического анализатора для сибирского ингерманландского идиома

Иван Убалехт
Омский государственный технический университет /
644050, Россия, г. Омск, пр-т Мира, д. 11
ubaleht@gmail.com

Аннотация

Статья посвящена разработке морфологического анализатора для языка ингерманландских переселенцев в Сибири. Данный язык является малоресурсным, это вносит специфику в разработку программного обеспечения для него. В статье представлен алгоритм анализа слов, относящихся к имени существительному, и рассмотрен пример анализа слова.

Ключевые слова: морфологический анализатор; малоресурсные языки; обработка естественного языка; сибирский ингерманландский идиом

1 Введение

Малоресурсные языки (low-resource languages) характеризуются наличием одного или нескольких из ниже перечисленных свойств: отсутствием стабильной письменности и орфографии; отсутствием или небольшим количеством ресурсов в электронном виде; отсутствием или небольшим количеством программного обеспечения для работы с языком [1].

В главе 2 описывается проблема, присущая малоресурсным языкам, обладающим только следующими типами ресурсов: неаннотированные аудио (видео) данные; лингвистическое описание, выполненное в форме, непригодной для компьютерной обработки. Рассматриваемый в статье сибирский ингерманландский идиом относится к этой группе малоресурсных языков. В главе 3 дано краткое описание сибирского ингерманландского идиома, для которого разрабатывается морфологический анализатор и дана характеристика его ресурсов. Глава 4 посвящена рассмотрению морфологического анализатора. Рассмотрен алгоритм и пример анализа слова, принадлежащего к имени существительному.

2 Обзор работ и постановка задачи

Языки, относящиеся к группе малоресурсных языков, обеспечены ресурсами в разной степени. Например, такие языки как карельский [2] или якутский [3] обладают относительно большим количеством текстов различных типов (периодика, массивы сообщений из социальных сетей и т.д.), на основе которых можно строить языковую модель. Наличие языковой и акустической моделей позволит разработать для этих языков системы автоматического распознавания речи (далее САРР) и другое сложное программное обеспечение.

Существует другая довольно большая группа малоресурсных языков, для которых существует только два типа ресурсов: неаннотированные аудио (видео) данные, например, данные из экспедиций; лингвистическое описание, выполненное в форме, непригодной для компьютерной обработки, например, в «бумажной» форме или в виде текста в форматах PDF, Word и т.д. Как видно, обладая ресурсами данных типов невозможно работать с такими языками в электронной среде. Таких идиомов существует довольно много, например, идиомы, описанные в рамках экспедиций, полевой материал из которых хранится в архивах. Важной научной задачей является более широкое введение данных таких языков в научный оборот.

Единственный способ получить возможность работы с такими языками в электронной среде – это построение корпуса текстов через аннотирование аудио (видео) данных. Как известно [4], создание аннотаций для аудиоданных – это чрезвычайно трудоёмкий процесс: на аннотирование одного часа аудиоданных может уйти от 40 до 100 часов (это так называемая проблема «annotation bottleneck»).

Так как размеченные тексты для рассматриваемого типа малоресурсных языков можно получить только из аудиоданных, то хорошей идеей было бы разработка САРР для этих языков, которая бы автоматизировала, хотя бы частично, процесс получения новых текстовых данных из аудио. Так как САРР обычно строятся на основе принципов машинного обучения, то для малоресурсных языков такой подход почти неосуществим, но, тем не менее, есть работы [1,5], которые рассматривают применение САРР для малоресурсных языков. В работе [6] описывается способ получения первоначального обучающего набора для малоресурсных языков. Суть подхода заключается в том, что вместо того чтобы аннотировать аудиоданные из экспедиций, в которых речь может быть спонтанной и не систематизированной, можно сначала взять качественные тексты, потом их озвучить носителями языка и синхронизировать речь с текстом. При таком подходе можно гораздо быстрее получить первый обучающий набор данных.

Чтобы использовать для сибирского ингерманландского идиома САРР и описанный в [6] способ получения первого обучающего набора данных, нужен исходный набор текстов, который можно получить для данного языка только из аннотаций. Поэтому, схема подхода обеспечения ресурсами сибирского ингерманландского идиома и языков с похожим исходным набором ресурсов может быть следующая:

- построение морфологического анализатора и генератора на основе лингвистического описания языка, это позволит работать с языком компьютерными методами;
- использование этого анализатора и генератора для получения словоформ, аугментации данных, расставления PoS тегов, автоматизации части рутинных операций при аннотировании аудио;
- интеграция с системами Apertium [7] и HFST [8], что поможет работать с фразами и автоматизировать перевод текста целевого малоресурсного языка на русский или английский языки;
- получив набор текстов, можно применить метод получения первого обучающего набора данных для САРР, описанный в источнике [6];
- разработка САРР, которая поможет быстрее получить большое количество новых текстов из оставшихся аудиоданных.

3 Сибирский ингерманландский идиом

3.1 Краткая характеристика

Язык ингерманландских переселенцев в Сибири или сибирский ингерманландский идиом – это смешанный язык, основанный на нижнелужских вариантах ижорского и финского языков. Исследователями также отмечается наличие водского субстрата. В современном состоянии отмечается влияние на данный язык эстонского и русского языков. Предки носителей данного языка мигрировали в Сибирь из финских и ижорских деревень, находящихся в районе нижнего течения реки Луги, в 1803-1804 гг.

В настоящее время языковой коллектив, использующий данных идиом в сфере бытового общения, сохраняется только в селе Рыжково в Омской области. Отдельные носители данного языка проживают также в населённых пунктах рядом с селом Рыжково, в городе Омске, в нескольких других населённых пунктах Омской области, а также в Эстонии.

Термин «сибирский ингерманландский идиом» (Siberian Ingrian Finnish) был введён Д.В. Сидоркевич, которая исследовала и документировала данный идиом в 2008-2014 гг. [9,10]. В настоящее время этот идиом продолжает исследовать Н.В. Кузнецова [11]. Язык ингерманландских переселенцев в Сибири обладает довольно значительным количеством фонологических, морфонологических, морфологических особенностей и предоставляет важный материал, например, для сравнительного анализа с другими языками прибалтийско-финской группы. Разработка программных инструментов для работы с этим языком будет полезна для работы с большими массивами аудиоданных таких близких идиомов, как финские и ижорские диалекты района нижнего течения реки Луги.

3.2 Обзор ресурсов данного языка

Обзор ресурсов, доступных для сибирского ингерманландского идиома, представлен в Таблице 1. Можно добавить, что для данного языка существует озвученный словарь в виде веб-приложения, разработанного автором статьи. Веб-приложение доступно в сети Интернет[1], исходный код этого веб-приложения находится в открытом доступе[2] на GitHub. Аннотации аудиоданных в формате ELAN[3], составленные автором статьи, а также другие материалы, включая исходный код морфологического анализатора, опубликованы в рабочем репозитории проекта[4] на GitHub под лицензией Creative Commons 4.0.

Тип ресурса	**Объём ресурса**
Аудиоданные, собранные Д.В. Сидоркевич в 2008-2014 гг.	80 часов
Аудиоданные, собранные автором статьи в 2019-2023 гг.	20 часов
Аудиоданные, опубликованные в открытом доступе под свободной лицензией Creative Commons 4.0	5 часов
Видеоданные	2 часа
Тексты, в основном транскрипция аудиоданных, сделанная вручную	42 тысячи токенов
Аннотированные в ELAN слова	200 слов
Число носителей языка, речь которых записана	31 человек

Таблица 1: Обеспечение ресурсами сибирского ингерманландского идиома

Как видно из Таблицы 1, данный идиом не обладает значительным набором текстов. Для дальнейшей работы с этим идиомом компьютерными методами необходима языковая модель, которую построить без наличия текстов невозможно. Единственный способ получить языковую модель этого языка – это составить набор аннотаций аудиоданных, но как отмечалось во второй главе, процесс аннотирования аудиоданных занимает большое время. Сократить время

[1] http://lexeme.net/sif
[2] https://github.com/ubaleht/Lexeme
[3] https://archive.mpi.nl/tla/elan
[4] https://github.com/ubaleht/SiberianIngrianFinnish

аннотирования, автоматизировать часть ручных операций позволит морфологический анализатор для данного языка.

4 Разработка морфологического анализатора

К настоящему моменту разработан модуль для морфологического анализатора, работающий со словами, принадлежащими к именным частям речи (существительные, прилагательные, местоимения, числительные).

На рисунке 1 показан алгоритм анализа слов, принадлежащих к имени существительному. Вход алгоритма – пять основ анализируемого слова: NOM.SG – анализируемое слово в номинативе, в единственном числе (слово в данной форме является леммой); PRT.SG – основа слова для получения слова в партитве в единственном числе; ILL.SG – основа слова для получения слова в иллативе в единственном числе; OBL.SG – косвенная основа слова в единственном числе, на основе её строятся все формы слова в единственном числе во всех косвенных падежах, кроме иллатива и партитива, а также для получения формы слова в номинативе во множественном числе; OBL.PL – косвенная основа слова во множественном числе. Альтернативный вход алгоритма – это только форма слова в NOM.SG. Выход алгоритма – размеченная строка в формате XML, содержащая все словоформы анализируемого слова, информацию о морфологии, метаданные о данном слове (код информанта, перевод слова, контекст записи и т.д.). В дальнейшем структурированная информация из формируемых XML файлов будет отправлена в базу данных на основе СУБД MS SQL Server.

```
INPUT:  - Либо слово W в форме NOM.SG.
        - Либо W в форме NOM.SG с основами OBL.SG, PRT.SG,
          ILL.SG, OBL.PL
OUTPUT: Все словоформы для W с морфологическим описанием
        в формате XML.
1. IF W только в форме NOM.SG
   THEN BEGIN
3.   Побуквенный анализ W, выбор одного морфонологического
     типа FROM CS1 TO CS8 OR FROM VS1 TO VS8;
4.   Для формирования каждой из основ OBL.SG, PRT.SG, ILL.SG,
     OBL.PL выбор морфонологического преобразования
     FROM CA1 TO CA5 OR FROM VA1 TO VA5 AND применение
     эвристических правил из множества E;
   END
6. Добавление падежных показателей к основам OBL.SG, PRT.SG,
   ILL.SG, OBL.PL и формирование для W форм слова в 10 падежах
   и 2 числах;
5. Запись всех словоформ W и морфологического описания в XML
   файл.
```

Рисунок 1: Алгоритм формирования словоформ для существительных

В настоящий момент данная часть морфологического анализатора с библиотекой анализа слов, относящихся к именным частям речи, реализована с помощью платформы .NET и языка C# как приложение для Windows. Приложение имеет пользовательский интерфейс для удобного, быстрого ввода имеющихся слов, принадлежащих к имени существительному и имени прилагательному. В дальнейшем планируется интеграция с Apertium и HFST для работы с фразами и переводами. В алгоритме на Рисунке 1 представлен анализ слов, принадлежащих к имени существительному.

Работа со словами, принадлежащими к имени прилагательному, осуществляется с помощью этого же алгоритма. Морфонологические типы у имён прилагательных такие же, как и у имён существительных – добавляется лишь несколько особенностей, например, анализ степеней

сравнения прилагательных. Для местоимений и числительных, которые используют те же парадигмы, анализ с помощью алгоритма не нужен, так как все их словоформы известны.

Падеж	**Единственное число**			**Множественное число**		
	Основа	**Суффикс**	**Пример**	**Основа**	**Суффикс**	**Пример**
Номинатив	NOM.SG	∅	*kukk*	OBL.SG	*-t*	*kuka-t*
Генетив	OBL.SG	*-n*	*kuka-n*	OBL.PL	*-n*	*kukki-n*
Партитив	PRT.SG	∅	*kukka*	OBL.PL	*-j*	*kukki-j*
Иллатив	ILL.SG	∅	*kukka*	OBL.PL	*-s*	*kukki-s*
Инессив	OBL.SG	*-s*	*kuka-s*	OBL.PL	*-s*	*kukki-s*
Элатив	OBL.SG	*-st*	*kuka-st*	OBL.PL	*-st*	*kukki-st*
Адессив – Аллатив	OBL.SG	*-l*	*kuka-l*	OBL.PL	*-l*	*kukki-l*
Аблатив	OBL.SG	*-lt*	*kuka-lt*	OBL.PL	*-lt*	*kukki-lt*
Транслатив	OBL.SG	*-ks*	*kuka-ks*	OBL.PL	*-ks*	*kukki-ks*
Комитатив	OBL.SG	*-nka*	*kuka-nka*	OBL.PL	*-nka*	*kukki-nka*

Таблица 2: Парадигма склонения для морфонологического типа CS1 с примером слова *kukk*, относящимся к этому морфонологическому типу

В качестве примера анализируемого слова возьмём слово *kukk* – «цветок». Чтобы получить все словоформы данного слова в соответствии с алгоритмом на Рисунке 1 необходимо: установить морфонологический тип и тип морфонологических преобразований. В настоящий момент для сибирского ингерманландского идиома определено 16 морфонологических типов [10 стр. 163, 170-171]. Восемь типов для слов имеющих консонантную основу (consonant stem) CS1 – CS8 и восемь типов для слов, имеющих гласную основу (vowel stem) VS1 – VS8. В области морфонологических преобразований существуют консонантные чередования (consonant alternation) CA1 – CA5 и вокалические чередования (vowel alternation) VA1 – VA5.

Если для анализируемого слова, например, для слова *kukk* известны пять основ – NOM.SG, PRT.SG, ILL.SG, OBL.SG, OBL.PL, то на основе их можно получить все словоформы добавлением падежных показателей к соответствующим основам, см. Таблицу 2.

Если для слова в форме NOM.SG неизвестны другие основы, то их можно получить с помощью алгоритма на Рисунке 1. Например, слово *kukk* относится к морфонологическому типу CS1 [10, стр. 169]. Алгоритм на рисунке 1 определяет это после побуквенного анализа слова. Тип CS1 определяется следующим образом: «…согласная основа с первичной геминантой, простым консонантным кластером, кластером с геминантой или одиночным согласным, имеющим слабую ступень чередования, в ауслауте…» [10, стр. 164]. К этому типу относятся также: *jalk* «нога», *harakk* «сорока», *penkk'* «скамейка», *rankk* «тяжёлый», *huntt'* «волк», *luut* «веник», *piipp°* «трубка», *kant* «ствол» и т.д. Для типа CS1 являются возможными следующие морфонологические преобразования:

- для формирования основы OBL.SG это правило CA1 или правило VA1;
- для формирования основ PRT.SG и ILL.SG это правило VA1;
- для формирования основы OBL.PL это правило VA2.

Основа OBL.SG была получена с помощью преобразования CA1: «…чередование ступеней последнего консонанта основы…» [10, стр. 161], *kukk* → *kuka-*, см. Таблицу 2. Основы PRT.SG и ILL.SG формируются с помощью преобразования VA1: «появление тематического гласного» (только для согласных основ) [10, стр. 161], в данном случае появление тематического *a* у основ PRT.SG и ILL.SG, *kukk* → *kukka*, см. Таблицу 2. Основа OBL.PL формируется согласно правилу VA2: «мутация тематического гласного» [10, стр. 162], в данном случае *a* мутирует в направлении *i*, см. Таблицу 2.

Почти для каждого из морфонологических типов существуют альтернативные правила формирования, так в текущем примере для типа CS1 это CA1 и VA1 для формирования основы OBL.SG. Для разрешения этой неоднозначности в алгоритме применяются эвристики, и сформированные словоформы помечаются как реконструированные.

5 Заключение и будущие работы

В статье показаны текущие результаты работы над морфологическим анализатором для сибирского ингерманландского идиома. В рамках проекта морфологического анализатора разработана библиотека, позволяющая формировать все словоформы для именных частей речи.

В дальнейшем планируется:

- разработать библиотеку для работы с глаголами этого языка;
- сформировать словарь на несколько тысяч лексем со всеми словоформами для данного языка;
- интегрировать этот проект, с проектами Apertium и HFST;
- унифицировать обозначения элементов морфологической разметки, используемых в морфологическом анализаторе для будущей интеграции с проектом Universal Dependencies[5];
- использовать этот морфологический анализатор для автоматизации аннотирования аудиоданных, что должно обеспечить выигрыш во времени по сравнению с ручным аннотированием, например, с помощью ELAN.

References

[1] Besacier, Laurent, Barnard Etienne, Karpov Alexey, Schultz Tanja. Automatic speech recognition for under-resourced languages: A survey. — Speech communication, 2014. Vol. — 56.

[2] Boyko Tatyana, Zaitseva Nina, Krizhanovskaya Natalia, Krizhanovsky Andrew,Novak Irina, Pellinen Nataliya, Rodionova Aleksandra // The Open corpus of the Veps and Karelian languages: overview and applications. Computing Research Repository. — 2022. — Vol. arXiv:2206.03870. — version 1. Access mode: https://arxiv.org/ftp/arxiv/papers/2206/2206.03870.pdf

[3] Ivanova, Sardana, Jonathan Washington, and Francis M. Tyers. A free/open-source morphological analyser and generator for Sakha // Proceedings of LREC 2022, Thirteen International Conference on Language Resources and Evaluation. European Languages Resources Association (ELRA). 2022. — P. 5137–5142.

[4] Seifart Frank, et al. Language documentation twenty-five years on. — Language, 2018. Vol. — 94(4).

[5] Prud'hommeaux, Emily, et al. Automatic speech recognition for supporting endangered language documentation. — Language documentation and conservation, 2021. Vol. — 15.

[6] Ćavar Malgorzata, Ćavar Damir, Cruz Hilaria. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR // Proceedings of the tenth international conference on language resources and evaluation (LREC'16). 2016. — P. 4004–4011.

[7] Forcada Mikel L., et al. Apertium: a free/open-source platform for rule-based machine translation. — Machine translation, 2011.

[8] Lindén Krister, Silfverberg Miikka, Pirinen, Tommi. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers // Proceedings of State of the Art in Computational Morphology: Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009). — Zurich, Switzerland, 2009. — P.28–47.

[9] Sidorkevich Daria 2011. On domains of adessive-allative in Siberian Ingrian Finnish. — Acta Linguistica Petropolitana, 2011. Vol. — 7(3).

[10] Сидоркевич Д. В. Язык ингерманландских переселенцев в Сибири: структура, диалектные особенности, контактные явления. Кандидатская диссертация. — СПб: Институт лингвистических исследований РАН, 2014. Режим доступа: https://iling.spb.ru/theses/1999

[11] Kuznetsova Natalia. Evolution of the non-initial vocalic length contrast across the Finnic varieties of Ingria and adjacent areas. — Linguistica Uralica, 2016. Vol. — 52(1).

[5] https://universaldependencies.org/