Санкт-Петербургский государственный университет

КОРТЕГОСО ВИССИО Николас

Выпускная квалификационная работа

Частеречная разметка для современного суринамского языка (сранан-тонго)

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5805.

«Компьютерная и прикладная и лингвистика»

Профиль «Компьютерная лингвистика»

Научный руководитель:

кандидат филологических наук,

доцент кафедры математической лингвистики,

Захаров Виктор Павлович

Рецензент:

Prof. Dr.,

von Waldenfels Ruprecht

Санкт-Петербург

2021

Аннотация

В данной работе исследуются общие морфологические и синтаксические черты современного суринамского языка (сранан-тонго). Целью исследования является разработка автоматической частеречной разметки (POS-теггера) для этого языка с использованием небольшого лексикона и минимальных обучающих данных. В работе рассмотрены теоретические вопросы, связанные с областью частеречной разметки, особенности существующих подходов, проблемы, возникающие при их использовании для малоресурсных языков, а также предложен способ преодоления этих ограничений для конкретного случая сранан-тонго. В практической части исследования даны описания разработанного теггера и эксперимента по проведению разметки. Методика проанализирована и оценена, высказаны предложения по дальнейшему развитию теггера.

Ключевые слова: суринамский язык, частеречная разметка, POS-теггер, стохастический POS-теггер, АОТ

This graduation qualification work examines the general morphological and syntactic features of modern Sranan Tongo. The aim of the research is to develop an automatic part-of-speech (POS) tagger for this language using a small lexicon and minimal training data. The work discusses theoretical issues related to the area of part-of-speech tagging, the characteristics of the existing approaches, the problems that arise when using them to tag low-resource languages, and a way to overcome these limitations for the specific case of Sranan Tongo. The practical part describes the proposed POS tagger and introduces an experiment to evaluate its performance. The experiment results are analyzed and suggestions are made for future development of the tagger.

Keywords: Sranan Tongo, part-of-speech, POS tagging, stochastic, rule-based, NLP

Оглавление

Введение

Процесс автоматической классификации слов текста по частям речи и соответствующей разметки известен как частеречная разметка, POS-теггирование или просто теггирование. POS-теггер обрабатывает последовательность слов и приписывает каждому слову тег части речи. Автоматическое присвоение метки (тега) части речи играет большую роль в синтаксическом анализе, в алгоритмах устранения неоднозначности слов и в поверхностном анализе текстов для быстрого поиска имен, времени, дат или других именованных сущностей в задачах извлечения информации. Корпуса с частеречной разметкой - полезный ресурс для теоретических и прикладных лингвистических исследований. Например, такие корпуса могут помочь найти экземпляры или частоту определенных конструкций.

POS-теггеры доступны не более чем для нескольких сотен языков. Целью настоящей работы является представление общих принципов проекта по созданию POS-теггера для современного суринамского языка и построение работающего прототипа, который может быть использован для разметки текстов на этом языке.

Актуальность данной работы обусловлена выбором мало изученного языка, для которого отсутствуют ресурсы автоматической обработки текстов. Суринамский язык упоминается по-разному: «сранан-тонго» (букв. «язык Суринама»), «сранан», «суринамский креол», «ненгре», «таки-таки», хотя последние два варианта считаются унижающими [Arends J., 1989:2]. В данной работе употребляется термин сранан-тонго, именуемый в дальнейшем «СТ».

СТ - креольский язык, основанный на английском языке, на котором в мире говорят около 647600[1] человек (ethnologue.com). В настоящее время это

1 www.ethnologue.com/language/srn

родной язык для примерно 126000 суринамцев, а также это второй (или третий) язык для большей части остального населения Суринама, которое составляет около 400000 человек. Также подсчитано, что более чем 200000 человек суринамского происхождения, которые сейчас живут в Нидерландах, также говорят на этом языке. Как и большинство креольских языков, истоки СТ восходят к эпохе европейской колонизации Америки, Азии и Африки в XV—XX веках. Первоначально он служил языком общения между рабами и хозяевами, а также между рабами из разных африканских регионов. В течение короткого времени он стал родным языком суринамских рабов. По сравнению с другими креольскими языками в Атлантике, развитие языка хорошо документировано: существуют письменные записи с 1888 года. После эмансипации рабов в 1863 году СТ оставался родным языком креолов низшего класса (людей рабского происхождения), но он также служил как *lingua franca* между креолами и азиатскими иммигрантами (индейцами, яванцами и китайцами). Постепенно СТ стал презираемым языком, явным признаком низкого социального статуса и отсутствия надлежащего обучения. После 1946 года, в канун независимости, за очень короткое время СТ стал более уважаемым (и респектабельным) благодаря достижениям креольских поэтов [Creole drum, 1975:vii]. Несмотря на развитие литературной культуры, СТ остается преимущественно устным языком: в Суринаме официальным языком является нидерландский, и поэтому пресса, законы и административные документы публикуются на этом языке. По этой причине объем письменного материала СТ совсем небольшой.

Что касается вычислительных методов присвоения словоформам тегов частей речи, большинство алгоритмов теггирования делятся на два класса: теггеры на основе правил и вероятностные или стохастические теггеры. Теггеры на основе правил обычно включают большой свод правил вычисления частеречной принадлежности и устранения неоднозначности,

которые определяют, например, что неоднозначное слово является существительным, а не глаголом, если ему предшествует определитель. Стохастические теггеры разрешают неоднозначность тегов, используя обучающий корпус для вычисления вероятности того, что данное слово имеет данный тег в данном контексте. Модели на основе правил требуют больших лингвистических знаний и большого количества правил, созданных вручную. Специализированная языковая инженерия стоит дорого и требует лингвистически подготовленных носителей языка. С другой стороны, статистические модели требуют меньше инженерии, они более надежны и работают лучше на больших реальных данных. Однако для обучения моделей высокого качества необходимы большие объемы обучающих данных, которые недоступны для большинства языков, и поэтому статистические модели плохо работают в условиях ограниченных ресурсов.

С точки зрении автоматической обработки текстов, СТ считается малоресурсным языком. Для малоресурсных языков отсутствуют большие одноязычные или параллельные корпуса и/или вручную созданные лингвистические ресурсы, достаточные для применения моделей, которые требуют больших объемов обучающих данных. Несмотря на отсутствие аннотированных данных, обычно существуют некоторые ресурсы, которые могут быть полезны для языков с низким уровнем ресурсов, включая двуязычные лексические ресурсы или подсказки из связанных языков. Способы эффективного использования этих ресурсов для повышения производительности модели при обработке малоресурсных языков – вопрос исследования и цель данной работы.

Гипотезу исследования можно сформулировать следующим образом: в случае СТ (и языков с подобной структурой), при решении задачи построения POS-теггера, дефицит обучающих данных можно частично компенсировать гибридным подходом, сочетающим подход, основанный на

правилах и стохастические методы. Предлагаемый подход использует пару простых правил и небольшой лексикон для присвоения набора возможных тегов одной словоформе и модель 3-грамм для устранения неоднозначности возможных вариантов. Поскольку ручная разметка текстов - дорогостоящая и трудоемкая задача, в представленном здесь эксперименте предлагается обучать модели 3-грамм на небольшом объеме данных при попытке преодолеть разреженность данных.

Цель создания POS-теггера языка СТ требует выполнения следующих задач:

- изучение структуры слов и синтаксиса СТ;
- составление небольшого лексикона;
- формирование набора частей речи на основе распределения слов в рамках предложения;
- выбор и построение модели POS 3-грамм для теггера;
- составление небольшого набора вручную размеченных предложений для обучения и тестирования модели;
- обучение модели на основе аннотированных предложений;
- тестирование и оценка модели.

Работа состоит из шесть глав, заключения, списка литературы и приложений. Первая глава знакомит с областью лингвистических исследований в СТ и электронными ресурсами, доступными в сети для СТ. Во второй главе представлен краткий анализ письменного варианта СТ, и обсуждаются такие его особенности, которые необходимо учитывать, чтобы выбрать и создать подходящий метод токенизации (выделения словоформ) и дальнейшей обработки. Во второй главе также исследуется морфология СТ в связи с проблемой определения частей речи. Третья глава относится к общим синтаксическим и грамматическим категориям современного СТ с целью определения набора тегов для обучения модели на основе их распределения в

синтаксических конструкциях. В четвертой главе представлены основные подходы к частеречной разметке в области автоматической обработки текста (АОТ), и предлагаются некоторые способы включения вспомогательных знаний в архитектуру модели для уменьшения разреженности данных при обучении модели с небольшими наборами аннотированных данных. В последней главе представлены методология, план эксперимента и его результаты.

Эксперимент предназначен для оценки производительности POS-теггера, обученного на небольших обучающих наборах, с использованием трех различных метрик. Обучающие наборы состоят из примеров предложений, извлеченных из набора данных APiCS (327 предложений, 2851 токен) [Winford D. et al., 2013] и описания языка (219 предложений, 1858 токенов) [Nickel M. et al., 1984]. Дополнительное лингвистическое знание кодируется в виде лексикона, содержащего 346 ключевых словоформ и те частеречные теги, которые они могут принимать. Остальные словоформы (те, которых нет в лексиконе) будут предсказываться моделью.

В тестовой выборке из 70 предложений POS-теггер показал следующие средние значения для набора из 31 тега: 0.81 для точности и 0.81 для полноты. Ожидается, что эти значения будут расти с увеличением объема данных в лексиконе и в обучающей выборке. Это является гипотезой для дальнейших экспериментов.

Новизна работы заключается в том, что впервые разрабатывается POS-теггер для CT, который может быть адаптирован для других креольских языков с аналогичными структурами.

Теоретическая значимость данной работы состоит в разработке модели работающего стохастического POS-теггера для CT с обучением на малом объеме данных и апробации методов включения вспомогательных знаний в архитектуру модели.

Практическая значимость исследования состоит в том, что пользователь POS-теггера получает возможность вводить в модель свои собственные параметры: переопределять и использовать собственные теги, изменять параметры модели и обучать ее на основе других данных.

Код работающего POS-теггера доступен по ссылке: https://github.com/nicolascortegoso/pos-tagger-sranan-tongo

Глава 1. Предыдущие работы

В этой главе рассматривается общая область лингвистических исследований о СТ и основные подходы для работы с языками, у которых мало лингвистических ресурсов для создания статистических приложений для автоматически обработки текстов. Далее представлены несколько электронных ресурсов, которые доступны в сети для СТ. Ближе к концу главы вводится тема разметки частей речи.

1.1 Лингвистическое исследование языка сранан-тонго

СТ исследуется часто в рамках атлантических креольских языков. Атлантические креолы включают большое количество разнообразных языков, на которых говорят в Африке, Карибском бассейне, на материковой части Южной Америки, а также в Северной и Центральной Америке. К ним относятся креолы, которые лексически связаны с нидерландским, английским, французским, испанским и португальским языками. Креолы с английской лексикой варьируют от довольно близких по структуре к их суперстратам (например, баджанский и тринидадский креольский) до других, которые довольно сильно расходятся(например, суринамские креолы). Последних иногда называют «радикальными» креолами, что означает, что по сравнению с другими креольскими языками они оставались достаточно свободным от внешних влияний с момента своего создания. [Winford D., 2008:19]. Объяснение различий и сходства креолов остается одной из центральных задач креольских исследований. Например, проект APiCS (The Atlas of Pidgin and Creole Language Structures)[2] собирает сопоставимые синхронные данные о грамматических и лексических структурах большого количества пиджин и креольских языков.

2 https://apics-online.info/

Лингвисты занимаются также факторами, которые участвуют в формировании креолов, и вопросом: отличается ли развитие креолов по своему характеру и степени от развития других языков [Migge B., 2011]. Специалисты выдвинули три основные гипотезы относительно структурного развития креольских языков – гипотезы субстрата, суперстрата и универсализма. Согласно субстратистам, креолы были образованы языками, на которых ранее говорили африканцы, порабощенные в Америке и Индийском океане, что наложило свои структурные особенности на европейские колониальные языки. Гипотеза суперстрата предполагает, что основными исключительными источниками структурных особенностей креола являются нестандартные колониальные разновидности европейских языков, из которых они развились, и вклад субстрата предположительно незначителен. Влияние субстрата в основном определяло, какие альтернативы в европейских колониальных языках станут частью креольских систем. Универсалисты утверждают, что креолы развивались в соответствии с универсалиями языкового развития. Согласно версии этой гипотезы, называемой гипотезой языковой биопрограммы «language bioprogram hypotesis» [Bickerton D., 1984], дети, которые познакомились с пиджином в раннем возрасте, создали креольский язык, усвоив только словарный запас пиджина. Дети разработали новые грамматики, следуя стандартным спецификациям биологической схемы языка, известной как универсальная грамматика или биопрограмма. При сравнении случаев, когда язык лексификатора (тот, от которого унаследована большая часть словарного запаса) является одним и тем же, считается, что структурные различия креолов возникают из-за разной степени влияния субстрата. До сих пор продолжается дискуссия о роли универсалий овладения языком в креольском образовании. С этим тесно связан вопрос о том, какие аспекты креольской грамматики сформировались под влиянием суперстратных или субстратных

источников или возникли в результате нововведений и внутренних изменений [Winford D., 2008:20]. Список библиографии показывает, насколько эти дебаты стимулировали исследование СТ.

СТ особенно интересен лингвистам, поскольку в отличие от других английских креолов, возникших в подобных обстоятельствах, он отделился от лексификатора на ранней стадии своего развития: Суринам был первоначально колонизирован с Барбадоса английскими поселенцами в 1651 году, но с 1668 года прочно находился в руках голландцев. В следующем году английские плантаторы уехали и, следовательно, влияние английского исчезало [Voorhoeve J., 1977:139]. Кроме того, по сравнению с другими креолами, СТ имеет довольно большое количество исторических письменных документов и даже словари или словарные записи (самый старый из которых датируется 1783 годом), которые позволяют проводить диахронические исследования. Например, работы Arends J. [1989], van den Berg M. [2000], Migge B. et al. [2009] занимаются разными аспектами синтаксиса СТ с диахронической точки зрения.

Что касается современного языка, обширных лингвистических описаний СТ до сих пор нет. Грамматика Donicie A. «De Creolentaal van Suriname Spraakkunst» [1954] часто цитируется и появляется в списке библиографии, но, к сожалению, невозможно было найти эту публикацию в открытом доступе. Данная диссертация во многом опирается на описание языка, представленное в работе Nickel M. et al. «Papers on Sranan Tongo» [1984], где авторы описывают на 46 страницах общую синтаксическую структуру языка. Другим источником общего описания СТ, к которому обращались в рамках диссертации, является Winford D. et al. [2013][3]. Поскольку Winford D. et al. цитируют работу Donicie A., можно ожидать, что часть его вкладов будет отражена в их исследовании. Описание СТ из этих

3 Доступно в онлайн сайте APiCS по ссылке https://apics-online.info/contributions/2)

двух главных источников было по возможности углублено и уточнено на основе работ, перечисленных в списке литературы. Некоторые характеристики языка относительно лучше изучены (маркеры времени и вида), чем другие (например, наречия).

Публикаций в области автоматической обработки текстов, относящихся к СТ, не обнаружено.

1.2 Исследовательские подходы для малоресурных языков

Для СТ отсутствуют ресурсы автоматической обработки текстов (АОТ). Большинство инструментов АОТ обучаются с использованием методов машинного обучения на больших аннотированных корпусах, которые недоступны для СТ. АОТ все больше и больше зависит от статистических методов и машинного обучения, до такой степени, что сейчас редко можно найти статью, опубликованную с использованием каких-либо основанных на правилах или других нестатистических методов.

Лингвисты, работающие с языками с ограниченными ресурсами, прибегли к различным методам решения этой проблемы. В принципе, существуют два основных подхода к АОТ в условиях ограниченных ресурсов, когда объем данных недостаточен для традиционных подходов. Первый подход начинается с этапа сбора данных на интересующем языке и обычно приводит к созданию инструмента АОТ. При этом используются такие ресурсы как списки слов, словари переводов и морфологические описания. Как правило, полученные результаты напрямую не применимы к другим языкам.

Второй подход относится к передаче аннотаций или моделей из богатого ресурсами источника на бедные ресурсами целевые языки. Существуют разные методы, но центральная идея заключается в том, что, при наличии параллельных текстов или определенных общих черт между двумя

языками, их можно использовать для построения языковой модели для одного языка из другой модели.

Хотя для СТ есть параллельный текст и переводы на английский или нидерландский, что делает применимым метод передачи аннотаций для частеречной разметки, в рамках данной работы применяется первый подход. Следовательно, выбранный здесь путь для создания POS-теггера – использование существующих электронных ресурсов для языка.

1.3 Электронные ресурсы

Одна из самых больших проблем с языками с низким уровнем ресурсов заключается в том, что ресурсы трудно получить. Большая часть существующих описаний языков либо не опубликована, либо существует только в бумажном формате. В случае СТ в сети находятся полезные ресурсы в виде онлайн-архивов и баз данных, которые можно использовать для создания POS-теггера. Следующие сайты являются одними из самых полезных.

OLAC Open Language Archives Community[4] (Сообщество Открытых Языковых Архивов) — это инициатива по созданию унифицированных средств поиска в онлайн-базах данных языковых ресурсов для лингвистических исследований. Информация о ресурсах хранится в формате XML для облегчения поиска. OLAC был основан в 2000 году и размещен на веб-сервере консорциума лингвистических данных Университета Пенсильвании.

OLAC дает рекомендации по передовым методам работы с языковыми архивами и работает над улучшением взаимодействия между ними. OLAC преследует:

4 URL: http://www.language-archives.org/language/srn

- достижения консенсуса в отношении передовой текущей практики цифрового архивирования языковых ресурсов;
- разработку сети взаимодействующих репозиториев и служб для размещения и доступа к таким ресурсам.

Архив OLAC содержит ссылки ресурсов на и о языке СТ. Каталог, содержащий относящиеся к этому языку ресурсы, включает лексические ресурсы, описания языка, слуховой и письменный материал. Большая часть этих ресурсов доступна онлайн, и многие из них перенаправляют на архив “SIL International”.

SIL International[5] (также известный как «Летний институт лингвистики») — это некоммерческая организация, принадлежащая евангелическому протестантскому христианству, основной целью которой является изучение, развитие и документирование малораспространённых языков для расширения лингвистических знаний, развития литературы и переводы Библии на эти языки. Это учреждение ведет базу данных “Ethnologue” и специализируется в основном на ненаписанных языках.

Исследователи SIL впервые приехали в Суринам в 1968 году и более 30 лет они занимались изучением языка, грамотностью, а также работали над созданием словарей, чтением книг, описаний грамматики и переводом. С 2001 года SIL больше не имеет официального присутствия в стране, но результаты их исследования доступны для всех, кто заинтересован в изучении этого языка.

В архиве SIL о СТ доступны заметки с описанием языка и набор народных сказок в формате PDF (большинство из них включают свой перевод на нидерландский).

5 URL: https://www.sil.org/resources/search/language/srn

SIL издали два словаря для СТ: “Wortubuku fu Sranan Tongo” нидерландско-сранан и английско-сранан под редакцией John Wilner [Wilner J., 2007]. Оба доступны в формате PDF по лицензии Creative Commons.

SIL создал также веб-сайт “The Languages of Suriname. Publishing the research of SIL in Suriname”[6], который содержит ресурсы для языков, на которых говорят в Суринаме. Материалы на этом сайте включают интерактивные словари “html”, которые позволяют искать слова онлайн, народные рассказы и библиографию в формате PDF всех работ SIL, опубликованных в Суринаме. В данном списке перечисляются упомянутые выше народные сказки, переводы из рассказов или отрывков из Библии и разные лингвистические исследования о СТ.

На сайте доступны также словари с исторической ценностью, которые представляют интерес для сравнения современного СТ с тем, каким он был и, возможно, как на нем говорили в прошлом:

- C.L. Schumann Neger-Englisches Wörterbuch ed. tertia (1783 Sranan - Deutsch)
- H.C. Focke Neger-Engelsch Woordenboek (1855 Sranan - Nederlands)
- H.R. Wullschlägel Deutsch - Negerenglisches Wörterbuch (1856 Deutsch - Sranan)
- Rediman Woordenlijst Nederlands - Surinaams (1971 Nederlands - Sranan

APiCS The Atlas of Pidgin and Creole Language Structures Online[7]: проект «Мировая Карта структур креольских языков и пиджинов» объединяет доклады от 88 специалистов по языку, которые работали над систематическим сравнением 120 ключевых структурных особенностей 76 креолов, пиджинов и смешанных языков в областях синтаксиса, семантики, морфологии, лексики и фонологии. Набор языков содержит не только

6 URL: http://www.suriname-languages.sil.org/
7 URL: https://apics-online.info/

наиболее широко изученные креолы Атлантического и Индийского океанов, но также менее известные пиджины и креолы из Африки, Южной Азии, Юго-Восточной Азии, Меланезии и Австралии, включая некоторые вымершие разновидности и несколько смешанных языков.

Языковые специалисты сделали два вклада в общий проект: создали “structure dataset” для каждого языка, с комментариями и примерами и написали главу книги “Survey of Pidgin and Creole Languages”. Онлайн версия составлена редакторами Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber. Donald Winford и Ingo Plag внесли вклад в изучение CT.

Пользователь может совершать поиск в базе данных по разным критериям: по языкам, свойствам, сравнению с базой данных WALS–APiCS, исследованиям, примерам, источникам и авторам. Набор данных CT включает описание синтаксических характеристик языка и список исследовательской литературы.

DBNL Digitale Bibliotheek voor de Nederlandse Letteren[8] (Цифровая Библиотека Нидерландской Литературы) — это сайт о нидерландских языке и литературе, содержащий около 11000 оцифрованных книг и журналов, которые можно просматривать бесплатно. Оцифрованная библиотека включает раздел суринамской литературы, где находятся творения суринамских авторов и работ суринамистов о литературе, языке и культуре страны.

В архиве библиотеки доступны выпуски журнала о суринамистике «Oso» (дом). С 1980 года в Нидерландах существует Instituut ter Bevordering van de Surinamistiek (Институт содействия суринамским исследованиям), который издает этот журнала два раза в год. В этом журнале обычно

8 URL: https://www.dbnl.org/

появляются статьи о лингвистических аспектах СТ и анонсируются новые публикации и появляются отчёты работ других авторов.

В DBNL также доступны оцифрованные версии журналов «Moetete» [1968] и «Creole Drum» [Lichtveld U. et al. 1975], содержащие художественные тексты на СТ, прозу и поэзию на СТ.

Glosbe[9] – многоязычный бесплатный общественный онлайн-словарь, связанный с поисковой системой в параллельных корпусах. Словарь разработан интернет-пользователями и энтузиастами лингвистики. На сайте есть словари между языковыми парами, которые невозможно найти на других сайтах.

Словарь позволяет осуществлять поиск словосочетаний при условии, что они присутствуют в предложении обоих корпусов. Источником текстов на СТ является сайт «Онлайн библиотека Свидетелей Иеговых».

Энциклопедия Википедия часто используется для построения инструментов АОТ и составления корпусов. Портал Википедии на СТ[10] включает (по состоянию на февраль 2021 года) около тысячи записей. Из них статьями можно считать лишь около десятка, потому что большинство из них – записи о странах, административных делениях, населенных пунктах, состоящие из менее чем сотни токенов. Что касается их содержания, имеется много повторений, так как они составлены при помощи ботов из шаблонов. В статьях также есть несоответствия в написании. По всем этим причинам Википедия не может пока рассматриваться как надежный источник для создания ресурсов для СТ.

9 URL: https://glosbe.com/
10 URL: https://srn.wikipedia.org/wiki/Fesipapira

1.4 Частеречная разметка

Частеречную разметку (POS-теггирование) можно определить как задание автоматически присвоения части речи каждому слову в предложении. Насколько известно, опубликованных работ по этой теме для СТ нет. Несмотря на это, частеречная разметка – это хорошо изученная задача, для которой предложены разные подходы. Большая часть исследований в этой области посвящена тому, как реализовать/адаптировать доступные подходы к реализации POS-теггеров для одного конкретного языка.

POS-теггеры в основном делятся на две группы: основанные на правилах и стохастические. Подходы, построенные на нейросетях, не обсуждается в рамках этой работы, так как они обязательно требуют большого количества оцифрованных данных, которых нет для СТ.

Набор тегов используемый POS-теггерами определяется на основе разных критериев. Обычно он кодирует как целевую функцию классификации, содержащую полезную для пользователя информацию о грамматическом классе слова, так и функции прогнозирования, то есть характеристики, которые будут полезны для прогнозирования поведения других слов в контексте. Эти две задачи пересекаются [Manning C. et al., 2003:144].

Manning C. упоминает, что понятие части речи на самом деле является сложным, поскольку части речи могут быть определены на различных основаниях, таких как семантические, синтаксические или морфологические основания. Часто эти представления о части речи противоречат друг другу. Для целей предсказаний можно было бы использовать определение части речи, которая лучше всего предсказывает поведение соседних слов, и это, предположительно, строго дистрибутивные теги. Но на практике часто используются теги, которые отражают условные или морфологические критерии.

В рамках этой работы набор частей речи определен на основе их распределения в синтаксисе. Для целей предсказания идеально было бы дать отличительные теги словам, которые имеют различную дистрибуцию. Теги не будут очень хорошими предсказателями, если они слишком общие, как в классическом различении: существительное, глагол, прилагательное, наречие, местоимение, предлог, союз, междометие, числительное, артикль или определитель. Более тонкие различия, например, знание того, является ли слово притяжательным местоимением или личным местоимением, может сказать нам, какие слова могут встречаться рядом с ним (притяжательные местоимения, вероятно, будут сопровождаться существительным, личные местоимения - глаголом) [Jurafvsky, 2008:124]. Но если теги слишком узкие, появляется риск того, что они могут привести к разрежению данных.

Глава 2. Графематический анализ сранан-тонго

В этой главе рассматриваются орфография и написание СТ и обсуждаются особенности, которые необходимо учитывать, чтобы создать подходящий токенизатор для выделения слов и дальнейшей обработки языка. Код для сегментирования текста на предложения и значимые единицы находится в приложении, в разделе B1. «Реализация программного кода алгоритма POS теггера».

2.1 Алфавит и структура слог

В СТ, как и в многих других языках, использующих ту или иную форму латинского алфавита, пробел является хорошим приближением к разделителю слов. Однако есть ограничения, например, из-за вариативности, с которой языки (или носители языка) воспринимают словосочетания и составные слова. Несмотря на то, что в СТ стандартизована орфография, есть степень вариации в написании текстов.

Гласные пишутся на СТ используя графемы «a», «e», «i», «o», «u». Есть еще две дополнительные «è» и «ò», встречающиеся в словах, которые недавно были приняты в язык. В СТ шесть дифтонгов «ai», «aw», «ei», «èi», «oi» и «ow». Сочетание гласных «ew» и «ui» не являются дифтонгами. Пропущенная гласная обозначается апострофом.

Для указания согласных используются 15 графем: «b», «d», «f», «g», «h», «k», «l», «m», «n», «p», «r», «s», «t», «w» и «y». Также используются пять диграфов (две буквы, которые вместе представляют один звук): «dy», «ng», «ny», «sy» и «ty». Символы «c», «j», «q», «v», «z», «x» в словах на СТ не встречаются, поэтому их можно использовать для выделения иностранных слов.

Слова состоят из одного или нескольких слогов. Большинство слогов состоит из одного или нескольких согласных, за которыми следует гласная или дифтонг. Слоги иногда заканчиваются буквой “m” или “n” и, реже, другими согласными. В нормальной речи часто встречаются удлиненные гласные или согласные. Обычно это результат исключения гласной или комбинации со следующей гласной. Удлиненные гласные встречаются очень редко, но бывают исключения. Они указываются в написании с акцентом с циркумфлексом, как в pôti «бедный».

2.2 Разница между правилам старого и нового написания

В течение многих лет СТ использовал написание, основанное на нидерландском правописании, записывая приблизительно так, как СТ звучал для нидерландцев.

С появлением движения, борющегося за статус СТ как респектабельного языка, возникла потребность в орфографии, основанной на собственной фонологии. Более подходящее написание возникло как неформальный консенсус из публикаций лингвистов, изучающих СТ и родственных креольских. В то время как некоторые литературные авторы переняли (варианты) лингвистического правописания, написание основанное на нидерландском оставалось распространенным для повседневного обихода. Чтобы решить эту проблему, правительство Суринама поручило комитету лингвистов и писателей определить стандартную орфографию, которая была принята и вступила в силу в 1986 году. Правила стандартной орфографии в основном следовало лингвистическому консенсусу. Однако, поскольку СТ не преподается в школах, в отличии от нидерландского языка, большинство говорящих на СТ четко не осведомлены о принципах стандартной орфографии и продолжают использовать вид написания под влиянием нидерландской орфографии.

Новое написание отличается от предыдущего стандарта из 1960 года по следующим пунктам.

- вместо графемы «oe» используется графема «u».
- вместо графемы «j» теперь пишется «y», также в сочетании с буквами «n», «s», «d» и «t».
- графема старого написания «é» заменяется в новой орфографии диграфом «ei».

Некоторые слова почти никогда не используются в их полной форме или только при особых обстоятельствах. Стандартное написание позволяет использовать апостроф при особых обстоятельствах, чтобы указать, что часть слова была опущена. Апостроф обычно встречается в письменных материалах для обозначения нормального произношения. Например: “m’ma” от “mama” «мать».

2.3 Написание составных слов

Во многих случаях такие категории как “слово” и “часть речи” не сопоставляются один в один. Комитет по орфографии рекомендует записывать слова вместе, когда части в сочетании имеют другую функцию или значение, чем когда они написаны отдельно. Например, “ala dei” означает «каждый день» и состоит из двух слов; “aladei” означает “повседневный, ежедневно” и поэтому записывается одним словом.

Несмотря на эту рекомендацию, образованные суринамцы склонны объединять слова в составные под влиянием нидерландского языка, который обычно следует этой практике. Например: “wansma” «кто-то», “bromkisiri” «семена цветов», “agumeti” «свинина» потому что они пишутся одним словом на нидерландском (iemand, bloemenzaad, varkensvlees). Согласно рекомендации комитета по орфографии эти слова могут быть написаны как два “wan sma”, “bromki siri”, “agu meti”.

В составных словах, где две гласные соединяются последовательно, первая гласная обычно опускается, а вторая часто немного удлиняется. Например: at’oso «больница».

2.4 Переключение кода отдельного слова и заимствование

Суринам – многоязычное общество, и суринамцы обычно заимствуют слова из того или иного языка в повседневной речи. Поэтому, трудно решить, принадлежит ли языку данное слово. Граница между СТ и суринамским вариантом нидерландского обычно очень размытым. СТ широко заимствует из других языков, изменяя способ произношения слов и корректируя их значение. Особенно с креольскими языками возникает вопрос, в какой степени можно говорить о заимствованных словах, поскольку лексика креольского языка в любом случае была создана из многоязычного контекста.

Другая проблема состоит в том, как отличать явление переключения кода от заимствования. Оба термина переключение кода отдельного слова и заимствование описывают вставку единой лексической единицы иностранного происхождения в матричный язык. Разница состоит в том, что переключение кода остается частью встроенного языка, а заимствования являются частью лексикона матричного языка. Они не только синтаксически и морфологически, но и фонетически адаптированы к своему матричному языку. Следующий пример из работы Radke H. [2017:124] демонстрирует это явление. СТ служит матричным языком, в который встроены отдельные лексемы нидерландского языка (подчеркивание указывает на слова на нидерландском языке):

- “Want we tan kree nomo fu den prijs ma un ap wan president nanga regering ete!”
- «Потому что мы продолжаем плакать о ценах, но у нас все еще есть президент и правительство»

В СТ есть эквиваленты для слов «правительство» и «потому что». Оба слова имеют фонетически похожий аналог в СТ (“presidenti” и “wanti”). Вопрос, представляют ли используемые формы изменение языка или они вызваны орфографическими причинами, в которых окончательный “i” был опущен, не является окончательным выясненным [Radke H., 2017:124]. Хотя слово “prijs” все еще показывает типичные орфографические особенности нидерландского языка (дифраф “ij” не появляется в орфографической системе СТ), в своей работе Radke H. указывает, что по мнению Blanker G. et al. [2010] и van der Hilst E. [2013] все равно оно считаются частью словарного запаса СТ.

При обработке текстов на СТ есть большая вероятность встретить много слов из нидерландского, которые принадлежат к разным частям речи и не только к существительным.

Глава 3. Морфология сранан-тонго

Как правило, креольские языки имеют небогатую морфологию и, поэтому, передача значения в основном выражена синтаксически, например через фиксированный порядок слов и использование нескольких отдельных слов, а не через изменение их формы.

Креольский имеет также довольно небольшой словарный запас по сравнению с некреольскими языками. Креолы весьма бедны в словообразовательных морфемах, и СТ не исключение. Несмотря на это, у креольских есть и другие ресурсы для увеличения ссылочной способности своего лексикона. Например, СТ полагается на словосложение, редупликацию и мультифункциональность слов для расширения словарного запаса. Последние две стратегии не требует существования или развития отдельных грамматических морфем.

Часть речи или класс слов, к которому принадлежит конкретный лексический элемент, определяется положением в предложении. В большинстве случаев нет никакого морфологического признака, чтобы показать класс слов отдельно от положения. Поэтому, части речи СТ рассматриваются в рамках синтаксиса в следующем разделе.

3.1 Аффиксация и словосложение

Креолы демонстрируют огромные различия в степени и типах производной аффиксации, которыми они обладают. С одной стороны, есть креолы, которые практически не усвоили какие-либо аффиксы языка-лексификатора, в то время как другие креолы имеют целый ряд аффиксов из их суперстратов. СТ разработал свой собственный довольно ограниченный набор аффиксов, обычно на основе свободных морфем из лексификатора, и лишен каких-либо английских лексических аффиксов [Plag I., 2009:339].

Словарь Wilner J. [2007] перечисляет шесть аффиксов:

- “gran-”: в существительных, где “gran-” является первым элементом, он указывает, что это самый большой или самый важный элемент. Например “granman” верховный вождь, “grankrutu” верховный суд, “grankownu” император. Использование: “grani sma” относится к пожилым людям, тогда как “gransma” относится к важным людям.
- “-man”: в составных словах он используется для обозначения того, кто делает или делает то, что указано в заглавном слове, например “fufuruman” (воровать → вор), “gronman” (земля → земледелец), “bereman” (беременность → беременная женщина), “guduman” (богатство → богатый человек).
- “-oso”: в составных словах, где вторым элементом является “-oso”, это относится к зданию, используемому в связи с существительным первого элемента. Например “datra-oso” (дом врача → больница).
- “-sei”: в словах, где “-sei” - второй элемент, он служит для обозначения общего местоположения или направления. Например “fotosei” в сторону города.
- “-sma”: в сложных словах, где главное слово является прилагательным, оно используется для обозначения человека, обладающего этим качеством, например “dedesma” мертвый человек, “granisma” старый человек. Когда главное слово является существительным, обозначающим место или группу людей, оно относится к человеку из этого места или группы, например “fotosma” городской человек.
- “-tenti”: суффикс для двузначных чисел, оканчивающихся нулем (например, 20, 30 и т. д.), кроме десяти.
- “-wan”: используется с прилагательными для обозначения человека, обладающего этим качеством. Например “breniwan” слепой, “pôtiwan” бедный человек.

Если под аффиксами подразумеваются исключительно морфемы, которые не могут существовать отдельно, предыдущий список сокращается до префикса “gran-” и суффикса “-tenti”, который появляется в ограниченном наборе числительных.

Словосложение очень распространено. Sebba M. [1981:107] разделяет его модели на следующие классы:

- существительное + существительное;
- прилагательное + существительное;
- глагол + существительное.

Существительные “ten” «время», “presi” «место», “man” «человек» и “sani” «вещь» могут сочетаться практически с любым существительным, глаголом или прилагательным, чтобы образовать соединения с ожидаемыми значениями.

В комбинациях с глаголом:

- глагол + “man” → человек, который обычно или профессионально что-то делает: “wroko” «работать» + man «человек» → wrokoman «рабочник»;
- глагол + “ten” → подходящее время для определенного действия: “sribi” «спать» + ten «время» → sribiten «время сна»;
- глагол + “sani” → вещь, используемая для выполнения действия: prei «играть» + sani «вещь» → preisani «игрушка».

Компаунды от прилагательных и от наречий могут образовываться из чисел, прилагательных и глаголов путем добавления “fasi” «образ» [Sebba M., 1981:107].

trafasi «разное» ← другое + образ

gersifasi «подобным образом» ← похож + образ

3.2 Редупликация

Редупликация – это шаблон, при котором языковая форма (полностью или частично) повторяется непосредственно перед или после базовой формы, чтобы выразить изменение ее значения. Чаще всего редупликация выражает интенсивность/затухание и итерацию [Haspelmath M., 2013].

Sebba M. утверждает, что роль редупликации в СТ ограничена, и ее функция определяется категорией простой формы. Если существует несколько гомофонных лексических единиц, каждое из которых относится к разным категориям, обычно только один подвергается редупликацию, хотя есть некоторые исключения [Sebba M., 1981:113]. Редупликация затрагивает в основном самые старые односложные или двусложные слова и редко более свежие заимствования из нидерландского [Sebba M., 1981:103].

В СТ есть некоторые существительные, структура которых указывает на редупликацию, но чья нередуплицированная форма не существует. Это характерно для некоторых существительных, относящихся к животным, растениям и частям тела. Например: “konkoni” «кролик».

Есть группа существительных, которые представляют собой названия инструментов, используемых для выполнения определенного действия. Например [Sebba M., 1981:104-105]:

kankan «расческа» ← kan «расчесать»
nanay «игла» ← nay «шить»

Существительные, образованные от глаголов, прилагательных и других существительных:

fonfon «наказание» ← fon «бить»
rediredi «дизентерия» ← redi «красный»
busbusi «подлесок» ← busi «куст»

Sebba M. [1981:105] замечает, что редуплицированные глаголы всегда производятся от других глаголов. Обычно их значение — многократное действие, обозначаемое глаголом. Например:

wakawaka «прогуливаться» ← waka «ходить»
takitaki «сплетничать» ← taki «говорить»

Редуплицированные прилагательные могут быть образованы от других прилагательных или от существительных. Такие производные от существительных обычно имеют значение атрибутами существительного как характерного качества [Sebba M., 1981:105-106]. Например:

tifitifi «зубчатый» ← tifi «зуб»

Редуплицированное прилагательное обычно усиливает значение. С некоторыми ограничениями это кажется продуктивным процессом, который можно применить к любому прилагательному. В некоторых случаях повторение прилагательного дает переносное значение [Sebba M., 1981:106].

bisi «занят» → bisibisi «очень занят»
krin «чистый» → krinkrin «полностью»

Наречия могут образовываться от числительных [Sebba M., 1981:106]:

wanwan «по одному» ← wan «один»
afaafu «умеренно» ← afu «половина»

Есть также наречия, которые не действуют как прилагательные, но могут подвергаться дублированию [Sebba M., 1981:106]:

wantenwanten «тотчас же» ← wanten «немедленно»

Sebba M. [1981:106-107] приходит к выводу, что редупликация регулярно применяется к глаголам, прилагательным и наречиям, и результаты имеют предсказуемые значения (за исключением прилагательных с переносным» значением. С другой стороны, значение дублированных существительных непредсказуемо, за исключением подкласса существительных, указывающих на инструменты.

3.3 Мультифункциональность

Мультифункциональность - это способность одного лексического элемента выполнять более одной грамматической функции без каких-либо морфологических изменений [Lefebvre C., 2004:155]. Например в CT, слово “sukru” «сахар» является существительным в предложении: “Sukru no de fu feni” «сахара нет», но оно может действовать как атрибутивное прилагательное, в “a sukru te” «сахарный чай» или как предикативное в “a te sukru” «сладкий чай». В предложении “a sukru a te tumsi” «он / она добавил слишком много сахара в чай» - слово “sukru” является глаголом [Defares J., 1982:52]. Многофункциональность – широко распространенное явление в CT, но не полностью продуктивна. Sebba M. [1981:114] отмечает, что существует некоторые ограничения. Например, наиболее предпочтительными глаголами для перехода к существительному являются те, которые требуют только

субъект и прямого объекта, а те глаголы, которые могут иметь третий аргумент (“gi” «давать», “bai” «покупать» и т.д.) и непереходные глаголы движения исключаются из числа существительных. Sebba M. считает мультифункциональность особенно продуктивной в случае прилагательных: практически все в СТ могут также функционировать как существительные, обозначающие их абстрактные качества. Например [Sebba M., 1986:116]:

“ogri” «уродливый, плохой» (прилагательное) → “ogri” «зло, поступок» (сущест.)
“fri” «свободный» (прилагательное) → “fri” «свобода» (сущест.)

Sebba M. [1981:106] также утверждает, что в СТ прилагательные и наречия не различаются и, поэтому, большинство прилагательных также могут функционировать как наречия. Что касается происхождения существительного от глагола, он выделяет три обычных семантических отношения, которые могут иметь место между конкретным глаголом и относящимся к нему существительным:

- существительное - это общее название объекта или продукта действия, названного глаголом. Например: “nyan” → «съесть», «еда»;
- количество связанных пар глагол / существительное инструмент, в которых инструмент не подвергался дублированию. Например: “frey” → «летать», «крыло».

Многофункциональность обычна в СТ, но не полностью продуктивна: она имеет синтаксические ограничения.

Глава 4. Синтаксис сранан-тонго

В этой главе обсуждается общая структура предложения СТ. Также выделяются части речи и лексические элементы, характеризующие эти конструкции, с акцентом на их положение в порядке слов. На основе распределение слов определяется набор тегов для частей речи, который используется при разметке выборок для обучения и тестирования модели. Номенклатура предлагаемых тегов вдохновлена Броуновским Корпусом.

Что касается порядка слов, некоторые конструкции имеют более консервативный вариант, а другие возникли под влиянием нидерландского. Также среди специалистов существуют разные мнения о классификации некоторых элементов. Нижеследующее в значительной степени опирается на общие описания языка, предоставленные Nickel M. et al. [1984] и Winford D. et al. [2013] с некоторыми замечаниями других исследователей.

4.1 Порядок слов

В СТ основной порядок слов можно описать формулой SVO: subject – verb – object (подлежащее – сказуемое – дополнение). Порядок слов довольно строгий и определяет функцию слов. В большинстве случаев в СТ, исходя из словоформы, не видно, к какой части речи принадлежит слово. Следовательно, часть речи слова определяется, изучив возможности комбинации слова в языковой ситуации. Например, следующее предложение:

mi nen Juwan [Voorhoeve J., 1956:191]

Словоформа “nen” может функционировать как существительное «имя» или глагол «называть». Кроме того, “mi” может указывать как на личное «я», так и на притяжательное местоимение «мой» первого лица единственного

числа. Хотя толкование пары “mi nen” как «мое имя» возможно, не получится прочесть целое предложение как «мое имя – Джуван», потому-что отсутствует нужная связка, как в следующем предложении:

en fesnen na Christoforus [Nickel M. et al., 1984:17a]
his first name is Christopher
его имя – Кристофер

Таким образом, в контексте примера Voorhoeve J., словоформа “nen” должна быть глаголом. Если бы был отрицатель “no” перед “nen”, тогда было бы еще одно дополнительное доказательство, что “nen” в рамке этого предложения действует как глагол. Voorhoeve J. предлагает включать в содержимое словаря для CT информацию об окружающих словах. Например, в записи словоформы “nen” [Voorhoeve J., 1956:192]:

1. называется (глагол), когда встречается с маркерами времени и вида “ben”, “sa” и “e”;
2. название (существительное), когда встречается с артиклями “wan”, “a”.

В разделе о мультифункциональности было рассмотрено как классы слов прилагательных и существительных полностью пересекаются. Было также упомянуто, что практически все прилагательные могут участвовать в качестве наречия. Что касается категорий глаголов и существительных, хотя многие лексические элементы могут принадлежать к обеим категориям, степень многофункциональности этого типа не безгранична.

4.2 Главные предложения

В CT все основные типы предложений – декларативные, вопросительные да/нет, императивы – имеют порядок слов SVO. Вопросительные предложения используют возрастающую интонацию, в

отличие от двух других типов, которые имеют убывающую интонацию [Winford D., 2008:21].

Анализ Nickel M. et al. [1984] разделяет предложения на четыре типа: активные, стативные, процессуальные и относительные. В основном, активные предложения выражают действие, стативные и процессуальные указывают на состояния и процессы соответственно и относительные предложения модифицирует некоторые существительные.

4.2.1 Активное предложение

Активное предложение состоит как минимум из глагола (V). Глаголу может предшествовать субъекту (S), состоящему из именной группы. За глаголом может следовать косвенный объект (IO), прямой объект (O), локатив (LOC), обстоятельства (ADV) и определение (MOD).

Таблица 1: Общая структура активного предложения (Nickel M. et al., 1984:2)

(S)	V	(IO)	(O)	(LOC)	(ADV)	(MOD)
подлежащее	глагол	косвенный объект	прямой объект	локатив	обстоятельства	определитель

И косвенный и прямой объекты состоят из именных групп. Локатив может быть местоположением, географическим названием или именной группой. Обстоятельства могут указывать время действия с помощью слова времени или именной группы, в которой главное слово является словом времени. Наречие также может встречаться в этом месте, чтобы уточнить действие.

Определитель в конце обычно состоит из предложной группой, указывающей инструмент, бенефициант, комитатив или определитель содержания объекта. Любой из аргументов, следующих за глаголом,

предположительно может быть приведен перед предметом, чтобы привлечь внимание к этому аргументу.

4.2.2 Стативное предложение

Стативные предложения описывают состояние. В СТ существуют три класса стативного предложения: описательное, экзистенциальное и эквативное. Они различаются по используемым глаголам и типам составляющих, которые могут следовать за глаголом [Nickel M. et al., 1984:7].

4.2.2.1 Описательное предложение

Описательное предложение состоит из именной группы в качестве подлежащего, прилагательного дополнения и необязательного определения.

Таблица 2: Структура описательное предложения (Nickel M. et al., 1984:7)

S	(V)	complement	(MOD)
подлежащее	глагол	дополнение	определение

Прилагательное дополнение придает существительному подлежащего некое качество[11]. В настоящем времени нет явного глагола, хотя может встречаться маркер продолжительного вида “e” [12c]. Описательное предложение может быть модифицировано наречием [14a] или предложной группой, выражающей сравнение [15a].

a doti [Nickel M. et al., 1984:12a]
it dirty
it's dirty
оно грязное

11 Некоторые специалисты утверждают, что это действительно глаголы (см. раздел о предикатных прилагательных конструкций в такой же главе.

mi frow e siki [Nickel M. et al., 1984:12c]
my wife CONT sick
my wife is sick
моя жена болеет

faya tumsi [Nickel M. et al., 1984:14a]
it hot very
it's awfully hot
ужасно жарко

mi oso moro bigi leki dati [Nickel M. et al., 1984:15a]
my house more big like that
my house is bigger than that
мой дом больше чем этот

4.2.2.2 Эквативное предложение

Эквативное предложение состоит из именной группы в качестве подлежащего и номинального дополнения, введенного связкой “na/a”. Прошедшее время выражается словом “ben de”.

Таблица 3: Структура эквативного предложения (Nickel M. et al., 1984:7)

S	COP	complement	(Mod)
подлежащее	связка	дополнение	определение

en na mi mati [Nickel M. et al., 1984:18a]
he is my friend
он мой друг

den tu sma disi na Sranansma [Nickel M. et al., 1984:19a]
these two people are Surinamers
эти два человека суринамцы

4.2.2.3 Экзистенциальное предложение

Эта конструкция используется для утверждения, что что-то либо существует сейчас, либо существовало когда-то. Экзистенциальное предложение отличается от описательного и эквативного предложений тем, что оно не может иметь ни прилагательного, ни именного дополнения соответственно [Nickel M. et al., 1984:9]. Необязательные дополнения, которые могут встретиться в этом предложении, указывают либо время существования (MOD), либо местоположение (LOC), либо могут изменять предложение каким-либо другим способом (ADV).

Таблица 4: Структура экзистенциального предложения (Nickel M. et al., 1984:9)

S	V	(ADV)	(MOD)	(LOC)
подлежащее	связка	дополнение	определение	локатив

dotdoti no de inisei [Nickel M. et al. 1984:16a]
dirty things no be inside
there isn't any dirt inside
внутри нет грязи

nownow wroko no de [Nickel M. et al. 1984:16b]
now work no be
there isn't any work now
сейчас нет работы

4.2.3 Процессуальное предложение

Процессы или понятие «становление» выражаются в СТ с помощью глаголов “kon” и “tron”. Глагол “tron” используется с прилагательными и является эквивалентом описательного предложения.

a kon fatu [Nickel M. et al., 1984:24c]
he's gotten fat
он растолстел

Описательное предложение, использующее маркер продолжительного вида “e”, также может указывать на процесс, как в следующем примере:

a e fatu [Nickel M. et al., 1984:24b]
he's getting fat
он толстеет

Глагол “tron” используется с прилагательными и является эквивалентом эквативном предложения. Дополнение, следующее за глаголом, неизменно является существительным или именной группой.

a strati tron wan liba [Nickel M. et al., 1984:25a]
the street became a river
улица стала рекой

4.2.4 Относительное предложение

Относительное предложение используется для дальнейшего определения существительного. Относительные придаточные предложения вводятся относительным местоимением, указывающим на роль, которую определяемое существительное играет в относительном предложении [Winford D. et al., 2013].

Таблица 5: Относительные местоимения

местоимение	
di, dati	человек
di, san	нечеловеческая сущность
pe	местоположение
fa	способ

Относительные местоимения представляют собой прямые объекты, косвенные объекты и объекты предлогов. Им присвоено тег «REL»

a tori disi san mi e go taki tide na fu srafuten [Nickel M. et al., 26b]
the story this that I CONT go talk today be about slave-time
this story that I'm going to tell today is about slavery.
эта история, которую я расскажу сегодня, о рабстве.

Отношение обладатель-владелец может быть указано при помощи “fu”, где предыдущий элемент станет обладателем, а следующий владельцем. Например:

a buba **fu** a tigri [Winford D. et al., 2013:2-6]
the tiger’s pelt
шкура тигра

Эта функция указывается в наборе тегов тегом «OF».

4.3 Вопросительное предложение

Вопросы в CT могут запрашивать информацию или просто ответы «да или нет». Информационные вопросы формируются путем размещения вопросительного слова, например “pe” «где» или “san” «что» в начале предложения. Порядок слов такой же, как указано в активном предложении,

за исключением того, что запрашиваемая информация не встречается [Winford D., 2008:29]. Например:

Oten yu o gi mi moni baka? [Nickel M. et al., 1984:9c]
when? you FUT give me money back
When will you give me the money back?
Когда ты вернешь мне деньги?

Следующие вопросительные слова используются для запроса информации:

Таблица 6: Вопросительное слово

вопросительное слово	запрос информации
san	о чем
suma	о человеке
pe	о месте
fa	о том, как что-то делать или как быть
fu sanede	о причине действия
sortu	«какой»

Таблица 7: Вопросительные слова, в которых всегда присутствует префикс “o”

вопросительное слово	запрос информации
omeni	о количестве (предметы, деньги, времени, и т.д.)
oten	о времени в целом

Частица “o” используется вместе с прилагательными для введения вопросов, например “o fara” «как далеко», “o bradi” «как широко», “o langa” «как долго», “o hei” «как высоко», “o lati” «как поздно, в какое время» [Wilner J., 2007]. Например:

o bradi a liba de? [Arends J., 1989:94]
“what wide the river is?”
какой ширины река?

Arends J. [1989:49] замечает, что в предыдущем предложении слово “bradi” является не прилагательным «широкая», а существительное «ширина». Таким образом, было бы лучше рассматривать вопрос «насколько широк» вместо «какой широкий». Тот факт, что Wilner J. говорит о комбинации “o” + прилагательное, несомненно, связан с тем, что он передает это в контексте сранан-английского словаря, где используются грамматические категории целевого языка, но по этому поводу специалисты имеют разные точки зрения.

Вопросительные слова обозначаются тегом «WP»

4.4 Императив

Субъект почти всегда удаляется, если он не во множественном числе. В остальном, структура такая же, как в активных предложениях.

Bai tu paki brede gi mi [Nickel M. et al., 1984:11b]
Buy two packs of bread for me!
Купи мне две пачки хлеба!

Гортативы вводятся с помощью слов “meki” или “kon” [Winford D. et al., 2013]:

Meki/kon unu libi a tori dati, yere [Winford D. et al., 2013: 34]
make/come we leave DET.SG story DEM, hear
Let’s forget that story, okay?
Давай забудем эту историю, ладно?

4.5 Расщепленные предложения

В СТ можно выделить все части речи, которые имеют самостоятельное значение: существительное, глагол, прилагательное, наречная конструкция, локатив. Выделенное слово или группы слов помещаются в начало предложения и вводятся связкой “na/a”, которая действует как маркера фокуса.

Winford D. [2008:26-27] выделяет два вида фокуса: презентационный (или информационный) фокус и идентификационный (или контрастный) фокус. Конструкции презентационного фокуса представляют некоторую новую тему и обычно включают начало именной группы.

Want **na** tu leisi mi nanga a man meki afsprak kaba. [Winford D. et al., 2013: 2-274]
Because it’s twice that the guy and I made an appointment already.
Потому что уже дважды я договаривался о встрече с этим парнем.

В идентификационном фокусе фронтальный элемент может быть любая основная составляющая предложения, включая предикаты или глаголы. Как правило, функция контрастного фокуса, состоит в том, чтобы идентифицировать какого-либо участника, сущность, место или время, которые, как предполагается, неизвестны слушателю, как действительного участника описываемой ситуации. В стандартных терминах презентационный фокус можно соотнести с темой, а идентификационный – с ремой.

Na yu fufuru mi moni! [Winford D. et al., 2013:2-272]
It’s you that stole my money
Это ты украл мои деньги

Когда глагол или предикат выделен, он всегда повторяется в предложении

A stik yu o **stik** a pikin. [Winford D. et al., 2013:2-276]
You'll suffocate the child this way.
Таким образом вы задушите ребенка**.**

Маркер фокуса помечаются тегом «FOC».

4.6 Сложносочиненное и сложноподчиненное предложения

В CT предложения часто вводятся сочинительными союзами, указывающими на их место в дискурсе. "Wèl" или "so" могут использоваться для открытия дискурса. Прогресс в повествовании обозначается "dan" «тогда». Чтобы резюмировать аргумент или выразить импликацию, используются союзы "dus" «так» или "na so". "Ma" «но» или "ma dan" «но потом» используются для обозначения того, что последующее не ожидается от предыдущих предложений.

4.6.1 Сложносочиненное предложения

Сложносочиненное предложения содержат два или более предложения, соединенных с сочинительной связью между его частями.

Координатные структуры можно разделить на три основных типа: простая координация с "dan" лил "èn" «и»; противодействующая координация с "ma" «но»; и дизъюнктивная координация с "of" «или». Предлог "nanga" «с» может использоваться для простого согласования словосочетаний с существительными [Winford D. et al., 2013]. Сочинительные союзы указываются тегом «CC».

4.6.2 Сложноподчинённое предложение

В сложноподчинённом предложении одно предложение является главным, а другое(ие) ему придаточным(ыми). Обе части сложноподчинённого предложения связываются союзами и союзными словами, которые обозначаются характер отношений. Придаточная часть предложения может указать цель, условие, время или причину действия в главном предложении. Следуя классификации Nickel M. et al. [1984:14-18] придаточные союзы могут также вводить вызванное действие (результат), речевой акт или что-то известное, или сравнение.

Целевое предложение вводится союзом “fu” и подлежащее обычно опускается. Оно всегда следует за главным предложением. Союз “fu” слабо артикулируется в нормальной речи и иногда удаляется. Однако контекст ясно показывает, что цель преследуется. В рамках этой работы, эта функция считается комплесентатором (см. следующий раздел).

a e bukun ini a peti **fu** teki a watra [Nickel M. et al., 1984:28a]
he stooped in the well to draw the water
он нагнулся к колодцу, чтобы черпать воду

Предложения, обозначающие причину, вводятся подчинительными союзами “want”, “bikasi” / “bika” «потому что» или “fu di”. CT также использует подчиненных, происходящих от нидерландского, таких как “want” и “omdat" «потому что». В отличие от целевых, в этих предложениях предмет не опускается. Обычно они следуют за независимыми предложениями.

someni ben dede **fu di** den no ben kisi wan bun yepi [Nickel M. et al., 1984:30a]
so many people died because they didn't receive good care
так много людей умерло из-за того, что не получили должной помощи

Эти предложения вводят отношения «если-то». Условная часть или часть «если» вводится с помощью “efu”. Есть два типа: реальный и нереальный. Нереальные условия включают гипотетические и контрфактические утверждения, которые передаются с помощью использования прошедшего времени в условной части предложения и сочетанием прошедшего плюс модального или будущего в последующем предложении. Это относится как к настоящим, так и к прошлым ситуациям. Часть результата «тогда» может быть или не быть введена с помощью слова “dan”. Для того чтобы выразить противоположное ожидание условия введены при помощи “aladi” или “nanga aladi” «хотя». Результат представлен “toku” «пока».

Efu a no doti mi no e wasi en furui [Nickel M. et al. 1984:31]
if it isn't dirty I don't wash it much
если оно не грязный, я его не стираю

Временные предложения могут предшествовать или следовать за основным предложением. Они указывают время события, указывая окружающие его обстоятельства. Они вводятся двумя подчиненными союзами, означающими «когда» “di” и “te”. Первый используется в случаях, когда упоминается конкретная (обычно прошлая) ситуация, а “te” используется для будущих, спекулятивных или неспецифических, включая привычные и нереализованные, ситуации. Другие временные союзы включают “fosi” «до» и “baka di / te” «после» и “sodra” «как только». Также встречаются сложные формы, такие как “vanaf di”, которое объединяет заимствованное слово нидерландского “vanaf” «от» и “di”. Главное предложение может быть введено словом “dan”.

te mi firi lesi **dan** mi no e teki wan pan [Nickel M. et al., 1984:33a]
when I feel lazy then I don't take a pan
когда мне лень, я не беру сковородку

Уступительные предложения можно разделить на три класса: условно-уступительные, передающие смысл «даже если», неопределенные уступки в смысле «несмотря ни на что» и определенные уступки в смысле «хотя». Первые два типа вводятся союзом “(a)winsi” «хотя, даже если». Определенные уступки делятся на два подкласса: те, которые передают смысл «хотя», и те, которые передают более сильный смысл «вопреки, независимо от того, насколько». Они вводятся “ala di” или “ala fa” соответственно [Windord D. et al., 2013].

Ala fa mi bari a meisje, toku a teki waka nanga a boi [Windord D. et al., 2013:67]
In spite of the fact that I warned that girl, she still went with that guy
Несмотря на то, что я предупреждал эту девушку, она все равно пошла с этим парнем

Подчинительные союзы представлены тегом «CS».

4.7 Комплементатор

В СТ дополнения можно разделить на два типа: факт-тип и потенциальный тип. Эти дополнения могут отображаться как полные предложения или могут быть каким-то образом сокращены, например, без явных субъектов или маркеров времени и вида и т.д. [Winford D. et al., 2013].

- аргументы предикатов типа “gersi” «кажется» и оценочным предикатам, таким как “tru” «истинно;
- дополнения глаголов утверждения (“gersi” «сказать», “taigi” «рассказать» и т. д.), глаголов психологического состояния (“sabi”

«знаю», “bribi” «верю» и т. д.) и глаголов восприятия (“si” «видеть», “yere” «слышать» и т. д.);

- дополнения причинного “meki” «делать».

Дополнения к “gersi” и “tru", всегда являются дополнительными сентенциональными субъектами [Winford D. et al., 2013].

Ma a **gersi** taki den kuli wani teki a kondre now op [Winford, D. 2000b: 96]
but it seem comp the.pl Indian want take the.sg country now up
But it seems that the Hindustanis want to take over the country now.
Но кажется, что иностранцы хотят сейчас захватить власть в стране.

В предложениях, указывающих на что-то сказанное, воспринимаемое или известное, обычно используется “taki”. Слово “taki” может выступать в качестве основного глагола, и в этом случае сказанные слова могут быть проанализированы как прямое дополнение активного глагола [Nickel M. et al., 1984]. Например:

Anansi **taki**: "mi abi wan sani dya..." [Nickel M. et al., 1984:37]
Anansi talk: I have a thing here
Anansi said, "I've got something here..."
Ананси сказал: «У меня здесь кое-что …»

Когда используется глагол, отличный от “taki”, то “taki” становится комплементатором, соединяющим то, что сказано, воспринимается или известно в основном предложении. Например:

ala sma sabi **taki** Anansi abi tumsi furu triki [Nickel M. et al., 1984: 38c]
all person know that Anansi have very many trick
everyone knows that Anansi has very many tricks
все знают, что у Ананси очень много хитростей

В дополнениях к глаголу речи и аналогичным глаголам, комплементатор “taki” часто заменяется на “dati” (нидерландский “dat”) или на нулевой комплементатор [Winford D. et al., 2013]. Например:

En mi hoop **dati** a kondre o kon bun, yere. [Winford D., 2000b:115]
and I hope comp ART country FUT come good hear.
And I hope that the country will get better, right.
Слушай, и я надеюсь, что страна станет лучшей.

Глаголы восприятия принимают два типа дополнения: конечный тип, представленный “taki”, а также сокращенный тип (маленькое предложение), который описывает события, одновременные со временем матричного глагола. Например:

Den si den pikin waka go na skoro [Winford D., et al., 2013:53].
3PL see the.PL child walk to LOC school
They saw the children walking to school.
Они видели, как дети идут в школу.

СТ использует слово “meki” «делать» или «позволять», чтобы выразить причину. Он может использоваться как глагол или как комплементатор, представляющий результаты предыдущего действия [Nickel M. et al., 1984:16-17].

a man dati ben e meki muiti **meki** furu srafu kon fri [Nickel M. et al., 1984:34b]
the man that PAST CONT make effort cause many slave come free
that man tried hard to set many slaves free
этот человек изо всех сил пытался освободить многих рабов

Дополнения потенциального типа вводятся предлогом “fu” «для» и выражают потенциальные события или состояния. Предикаты, которые их принимают, включают в себя глаголы, которые выражают желания, намерения, запросы и команды, «аспектные» глаголы, такие как «начинать», и модальные предикаты, такие как «должен», «способный», «обязанный» и т. д.

Дополнения потенциального типа могут быть сокращенными или полными. Когда объекты матрицы и комплемента являются корреляционными, “fu” может быть опущено [Winford D. et al., 2013]. Например:

Wan pikin aksi a man **fu** a man rij a laast rij. [Winford D., 2000a:433]
a girl ask the.SG man for the.sg man ride the.sg last ride
A girl asked the guy to take one last ride.
Девушка попросила парня прокатиться в последний раз.

Относительное местоимение может также служить дополнением вместе с глаголом, указывающим на познание или восприятие. Например:

a no sabi **pe** den boi kibri [Winford D. et al., 2013:40a]
he no know where thePL boy hide
he didn't know where the boys were hiding
он не знал, где прятались мальчики

В целом, существует близкое сходство между предлогами и союзами, что привело к предположению, что дополняющие друг друга могут рассматриваться как предлоги, принимающие клаузальное дополнение [Plag I., 1998:333].

Комплементаторы указываются тегом «COMPL».

4.8 Элементы именной группы

В активном предложении именная группа может служить субъектом, объектом, косвенным объектом, временным наречием и дополнением в эквативных предложениях.

Именная группа состоит из детерминатива, квантификатора, атрибутивов и главного слова, за которыми следуют один или несколько пост-определителей. Все, кроме главного слова именной группы, необязательно.

Таблица 8: Элементы именной группы [Nickel M. et al., 1984:20].

(DET)	(QUAN)	(ATTR)	HEAD	(SPEC)
детерминатив	кватификатор	атрибутивы	главное слово	спецификатор

4.8.1 Главное слово

Согласно Nickel M. et al. можно разделить главные слова на два класса в зависимости от того, могли ли определительный и показатели количества модифицировать их. Те, которые могут быть модифицированы, называются существительными. Имена собственные, местоимения и указательные местоимения принадлежат к другой группе.

4.8.1.1 Личное местоимение

Личные местоимения в CT не различают пол. Такая же форма личного местоимения используется для субъектных, объектных и притяжательных функций. Единственным исключением является местоимение третьего лица единственного числа, которое имеет отчетливую субъектную “a” и объектную и притяжательную формы “en”.

Таблица 9: Личные местоимения

	ед. число	мн. число	
1-ое лицо	mi	wi	unu
2-ое лицо	yu, i		
3-ое лицо	a, en	den	

Второе лицо не имеет формы вежливости. Второе лицо единственного числа имеет два варианта, причем “i” является наиболее распространенным в нормальной речи. Косая форма местоимения единственного числа 3-го лица “en” функционирует как субъект эквативной связки, а также как форма фокуса в конструкции расщепленны. В парадигме есть синкретическая форма “unu”, означающая и первое и второе лицо множественного числа.

CT различает между инклюзивной и исключительной личностными формами местоимений “мы”. Форма “unu” включает в себя он/она и я, исключая слушающего. Общая форма “wi” может означать и "мы включаем вас" и "мы исключаем вас" [Haspelmath M. et al., 2013]. По мнению Nickel M. et al. [1984] форма “wi” менее распространена, но используется, когда неоднозначность может возникнуть из-за использования “unu”.

Местоимение третьего лица единственного числа также используется как пустой субъект.

Личные местоимения представлены тегом «PRN», как и неопределенные местоимения, такие как “no wan” «никто» и “noti” «ничего».

4.8.1.2 Возвратное местоимение

В CT местоимения становятся возвратными после добавления суффикса “srefi”. Возвратные местоимения используются для обозначения того, что аргумент, не являющийся субъектом, в транзитивном предикате кореференционален с субъектом или связан с ним [Haspelmath M., 2013]. Возвратные местоимения встречаются в обычных переходных конструкций, в которых пациент кореференционален с агентом. Например:

a koti **ensrefi** nanga wan nefi [Winford D.et al., 2013:2-225]
3SG cut himself with a knife
He cut himself with a knife
он порезался ножом

Возвратное местоимение второго и третьего лица единственного числа формируется на основе форм “yu” и “en” личного местоимения соответственно. Также как основная форма, “unsrefi” может указывать на и первое и второе лицо множественное число.

Таблица 10: Возвратные местоимения / интенсификаторы

	ед. число	мн. число	
1-ое лицо	misrefi	wisrefi	unsrefi
2-ое лицо	yusrefi		
3-ое лицо	ensrefi	densrefi	

Возвратные местоимения указаны тегом «PPL».

4.8.1.3 Указательное местоимение

Обычно они действуют как спецификаторы, следующие за существительным, но могут также встречаться как главное слово именной

группы. Они различают расстояние от говорящего: “disi” «это» и “dati” «то». Указательные местоимения помечаются для множественного числа представлением артикля множественного числа “den”: “den disi” «эти», “den dati” «те».

Таблица 11: Указательные местоимения

	ед. число	мн. число
ближе	disi	den disi
дальше	dati	den dati

Тег «DT» помечает указательные местоимения.

4.8.1.4 Имя собственное

Nickel M. et al. утверждают, что имена собственные не могут быть модифицированы атрибутом, показателем количества или определителем, но могут быть помечены для выделения связкой “na”. Хотя, например, имена собственные обычно встречаются без артиклей, это не является абсолютным. Если принимать во внимание графематический анализ, то слова, начинающиеся с заглавной буквы и находящиеся в позиции, отличной от начала предложения, являются хорошими кандидатами в имена существительные собственные.

Имена собственные указаны тегом «NP».

4.8.1.5 Существительное

Существительное – это неизменяемое слово (не изменяется по родам, ни по числам). Они определяются в CT путем их способности встречаться с определителями и принимать притяжательные формы.

Nickel M. et al. [1984:32-34] разделяют существительные на группы в зависимости от того, как они могут быть изменены или где они могут использоваться в словосочетании или предложении.

- неисчисляемые существительные: существительные, которые не могут быть определен числом. Например: “moni” «деньги», “faya” «свет», “watra” «вода», “aleisi” «рис», “merki” «молоко»;
- исчисляемые существительные: существительные, которые могут быть определены числом. Например: “kaw” «корова», “dosu” «коробка», “sma” «человек»;
- собирательные существительные: относятся к группировкам или меркам исчисляемых или неисчисляемых существительных. Например: “kan” «чашка», “anu” «рука», “ipi” «топка». Они могут служить в качестве квантификаторов других существительных;
- локативные существительные (локативы): могут выполнять роль предложной группы, чтобы указать на местонахождение предикации. Следующие не могут быть использованы с локативным предлогом “na”: “dya” / “dyaso” «здесь», “drape” / “dape” «там». Другие локативные могут использоваться с “na”. Суффикс “sei” «сторона» играет важную роль в формировании локативных, которые служат объектом локативных групп, введенной “na”. Следующие локативные также могут быть автономными: “inisei” «внутри», “watrasei” «берег реки», “dorosei” «снаружи», “fotosei” «центр города», “Fransei” «Французская Гвиана»;
- слова темпоративы: они могут указывать, когда происходит действие, или могут использоваться в именной группе, чтобы указать продолжительность времени. Например день недели и время суток. Например: “esrede” «вчера», “tide” «сегодня», “tra tamara” «послезавтра», “mamanten” «утром», “noti” «ночью».

Обычные существительные размечаются тегом «NN». Поскольку слова темпоративы могут встречаться в самом начале фразы, они выделяются в специальном классе «номинальное наречие» и получают тег «RN».

4.8.2 Детерминатив

Детерминативом может быть артикль, притяжательная именная группа или местоимение. У Voorhoeve J. [1985:205-211] находится подробное описание последовательности элементов, которые определяют существительное в именной группе. Схема Voorhoeve J. (рис. 1) присваивает числа для обозначения элементов, которые появляются до и после существительного и могут определить его. Когда они встречаются вместе, такие элементы с наибольшим числом появляются первыми в порядке слов. В дальнейшем, они будут называться “степенями”, например, определенные артикли “a” и “den” соответствуют седьмой степени до существительного.

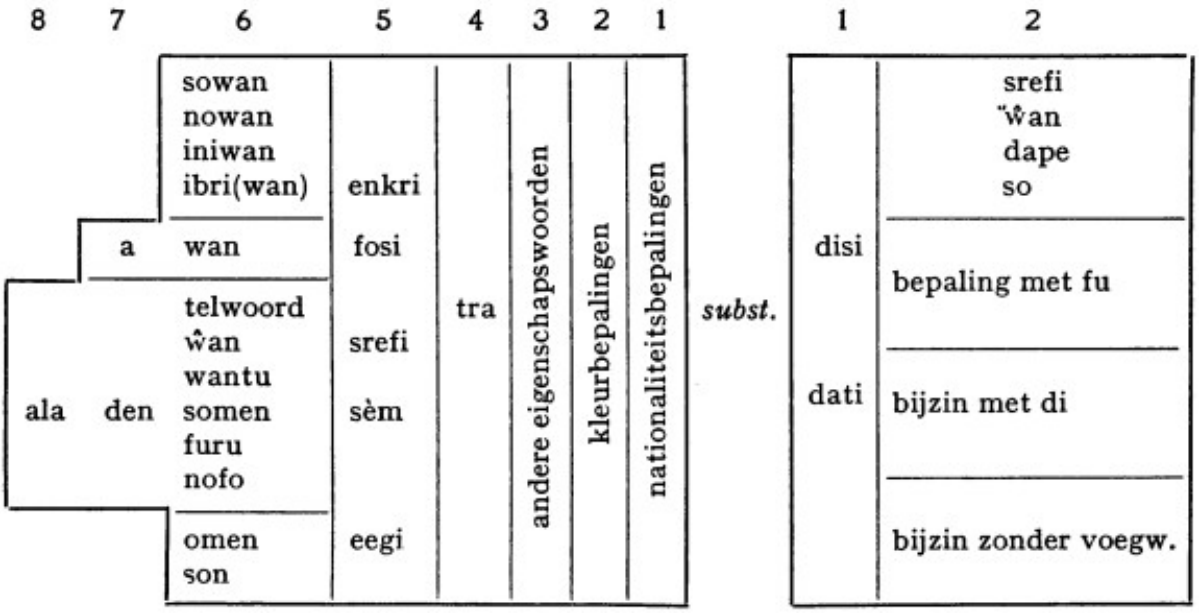


Рисунок 1: Порядок элементов именной группы [Voorhoeve J., 1985:2009]

Таблица 12: Элементы именной группы (Nickel M. et al.) по числам схемы Voorhoeve J.

детерминатив (определенные артикли)		квантификаторы + неопределенные артикли		атрибуты				сущест.	спецификаторы	
8	7	6	5	4	3	2	1		1	2

4.8.2.1 Артикль

Артикли вводят именные группы. Они указывают, является ли существительное в главном слове именной группы определенным или неопределенным. Числовые различия передаются с помощью определенной артиклей “a” для единственного и “den” для множественного чисел. Формы определенных артиклей совпадают с формами личных местоимений третьего числа.

Неопределенные существительные в единственном числе помечаются артиклем “wan” «один» от квантификатора. Неопределенные существительные во множественном числе, существительные общего вида и абстрактные существительные не требуют артикля [Winford D. et al., 2013].

Таблица 13: Артикли

	ед. число	мн. число
определенная	a	den
неопределенная	wan	∅

Определенные артикли отмечают референт, который упоминался ранее или который иным образом может считаться знакомым слушателю. Bruyn A. [2009:317] отмечает, что определенные именные группы иногда встречаются без какого-либо артикля в контекстах, где потребовался бы в других языках, например в английском. Это относится к именной группе с референтами,

индивидуальная идентичность которых идентифицируемая, но тривиальная, как в случае “tafra” «стол» в примере:

Opo a saka aleisi èn poti en tapu **tafra**.
raise def.sg sack rice and put 3sg on table
Pick up the sack of rice and put it on the table
Возьмите мешок с рисом и поставьте его на стол.

но также и к случаям, когда референт является важным участником дискурса, который может быть идентифицирован в конкретном контексте, как “pernasi” «плантация»:

Pernasi feti fu sungu bika sroysi boro.
plantation fight for sink because sluice have holes
The plantation was about to flood because the sluice(s) were cracked.
Плантация собиралась затопить, потому что шлюз(ы) был(и) сломан(ы).

В классификация Voorhoeve J. неопределенный артикль попадает в шестую степень. Определенные артикли единственного и множественного числа, классифицируются вместе в седьмой степени.

Артикли размечаются тегом «AT».

4.8.2.2 Притяжательное местоимение

В CT личные местоимения, выражающие обладателей - это отдельные слова, предшествующие владению. Формы притяжательных местоимений идентичны объектным формам личных местоимений. Это значит, что они не различают пол и для первого и второго лица множественного числа есть общая форма притяжательного [Nickel M. et al., 1984].

Таблица 14: Притяжательные местоимения

	ед. число	мн. число	
1-ое лицо	mi	wi	unu
2-ое лицо	yu		
3-ое лицо	en	den	

По образцу шаблона притяжательных местоимений, когда существительному предшествует другое, первое становится владельцем, а следующее - обладателем. Термин “владение” понимается в широком смысле, включая не только владение в прямой смысле, но также родственные отношения, владение частями тела и субъективное и объектный родительный падеж с номиналами действия, где обладатель будет субъектом или объектом в сентенциальной парафразе [Huber M., 2013].

mi **ma** prasoro [Winford D. et al., 2013: 2-4]
my mother’s umbrella
зонтик моей матери

Тег «PP$» указывает на притяжательные местоимение.

4.8.3 Квантификатор

Существительные количественно определяются кардинальными и порядковыми числами и другими количественными прилагательными.
Квантификатор может быть числительным или квантификатором как “furu” «многие» “someni” «столько», “eri” «целый», “pikinso” «немногие», “ibri” «всякий / каждый», “son” «некоторый», “moro” «больше», “ala” «все».

den **seibi** lowe srafu [Nickel M. et al., 1984: 44a]
the seven runaway slaves
семь беглых рабов

furu gowtu ben de drape [Nickel M. et al., 1984:44b]
much gold was there
там было много золота

Квантификаторы не могут встречаться вместе с числительными. Единственным исключением является “ala” «все», которое может сочетаться с кардинальным числом, и в этом случае “ala” предшествует другому слову.

ala tu go na skoro [Nickel M. et al., 50b]
all two go to school
both (of them) went to school
оба ходили в школу

Квантификатор “ala” также может вводить определенный артикль множественного числа. Поэтому, в схеме Voorhoeve J. “ala” встречается перед всем элементами:

ala den boi doti [Nickel M. et al., 1984:45]
all the boys are dirty
все мальчики грязные

Квантификаторы могут служить в качестве главного слова именной группы как в предыдущем и в следующем примерах. В этом случае существительное понимается, но не указывается.

someni ben dede [Nickel M. et al., 1984:50a]
so many PAST die
so many (people) died
так многие умерли

Квантификатор может также состоять из порядковых групп. Порядковые числа выражаются с помощью “di fu”, за которым следует кардинальное число начиная с двух. Значение “первый” передается словом “fosi”. Эта конструкция заменяется нидерландскими порядковыми номерами, особенно среди людей моложе 25 лет, которые выросли в Парамарибо [Nickel M. et al., 1984:24].

a **di fu fo** boi nen Robin [Nickel M. et al., 1984:46a]
'the fourth boy is named Robin
четвертого мальчика зовут Робин

Существительные могут быть количественно определены с помощью именной группы, главное слово которой является единицей измерения [Nickel M. et al., 1984:24].

gi mi tu **ipi** tomati [Nickel M. et al., 1984:56c]
give me two stacks of tomatoes
дай мне две стопки помидоров

По классификации Voorhoeve J., квантификаторы принадлежат к предопределителям пятой и шестой степени:

- предопределения пятой степени содержат слова типа “enkri” «одиночный», “fosi” «первое», “srefi” «тот же». Например: “nowan enkri tra sani” «ничего единственная другая вещь «ничего другого»»;
- предопределение шестой степени включают числительные. Например: “den furu srefi oso” «много одинаковых домов».

Числительные и квантификаторы размечаются тегом «NUMB» и «AB» соответственно.

4.8.4 Атрибутивное прилагательное

В СТ атрибутивные прилагательные предшествуют главному слову в именной группе, которое они модифицируют. Любые количественные показатели и т. д., которые их изменяют, встречаются перед прилагательными [Sebba M., 1986:114].

Nickel M. et al. [1984:21-22] разделяют четыре класса прилагательных на основе их распределения в цепочке прилагательных, которые могут предшествовать главному слову. В цепочке прилагательных атрибуты, обозначающие неизменные свойства, как национальность, располагаются ближе всего к главному слову. Они соответствуют предопределителям первой степени из классификации Voorhoeve J. До них появляются атрибутивные, указывающиеся цвет (предопределетели второй степени в схеме Voorhoeve J.). Атрибутивные, включающие другие свойства существительного встречаются в самом начале цепочки прилагательных (предопределетели третьей степени Voorhoeve J.). Nickel M. et al. [1984:40] выделяют еще категорию, атрибутивные размеры которой отсутствуют в классификации Voorhoeve J. Атрибутивные размеры находятся до атрибутивных обозначающих цвет.

Таблица 15: Порядок атрибутивных по схеме Voorhoeve J. и описанию Nickel M. et al.

Voorhoeve J.	общая категория [3]		цвет [2]	Национальность [1]	сущест.
Nickel et al.	общая категория	размер	цвет	неизменные свойства	

Атрибутивные могут быть также изменены некоторыми наречиями. В схеме Voorhoeve J. предопределители третьей степени [3] могут быть подробно определены словами “(pikinso) moro” «(немного) больше» или

предопределения второй степени только при условии предопределения третьей степени нет [Voorhoeve J., 1956:208].

Voorhoeve J. классифицирует предопределитель “tra” «другое» в качестве единственного элемента в четвертой степени. Он может быть определен только с помощью “heri” «очень». Например:

wan **heri tra** sani [Voorhoeve J., 1956:208]
een heel ander ding
совсем другая вещь

Когда именная группа используется в качестве атрибута, она в целом служит для модификации главного слова.

tra kondre sma [Nickel M. et al., 1984:48a]
other country people
people from other countries
люди из других стран

mi masra famiri oso [Nickel M. et al., 1984:48b]
my husband's family's house
семейный дом моего мужа

Существует два варианта, включающие порядок слов степени и прилагательных в СТ: в одном слово степени следует за прилагательным, а в другом – перед ним (см. раздел о сравнительной степени в этой главе).

Атрибутивные прилагательные представлены тегом «JJ».

4.8.5 Спецификатор

Под понятием «спецификаторы» [Nickel M. et al., 1984] понимаются слова или конструкции, которые могут пост-модифицировать существительное. В эту категорию включены указательные местоимения (пост-определители первой степени по схеме Voorhoeve J.) и небольшое количество определения качества (пост-определители второй степени). Согласно Nickel M. et al. [1984:40] они три: “wawan” «единственный», “srefi” «сам», “alamala” «все они». Такие спецификаторы размечаются при помощи тега «AP».

Другие спецификаторы включают относительные предложения, предложения, введенные с помощью “fu”, сравнения, введенные с “leki” и локативные группы [Nickel M. et al., 1984:23].

4.9 Элементы глагольной групп

По классификации Nickel M. et al. [1984:36] В СТ можно разделить глагольный класс на те, которые могут служить главным словом глагольной группы в активном предложении, на связочные глаголы и те элементы, которые составляют вспомогательные в глагольной группе. Связочные глаголы уже описаны (см. раздел о связки в этой же главе).

4.9.1 Активные и пассивные глаголы

Активные глаголы [Nickel M. et al., 2013] могут быть сгруппированы синтаксически в зависимости от того, принимают ли они объект или нет, и могут ли они принимать косвенный объект.

- непереходные глаголы не принимают прямого объекта. Например: “opo” «вставать», sribi «спать», sidon «садиться», fadon «упасть»;
- переходные глаголы имеют выраженный или понятый объект. Например: “teki” «брать», “skrifi” «писать», “abi” «иметь»;

- дитранзитивные глаголы принимают косвенный объект. Прямой объект обычно выражается, но иногда просто понимается в контексте. Например: “pai” «платить», “winsi” «желать», “gi” «давать».

Nickel M. et al. выделяют еще две важные группы, которые пересекаются с описанной выше классификацией – это глаголы, выражающие движение и глаголы, указывающие на познание, восприятие или речь:

- глаголы (как переходных, так и непереходных), которые обозначают движение. Например: “waka” «ходить», “lon” «бегать», “tyari” «носить», “puru” «тянуть», “swen” «плавать»;
- глаголы, которые указывают направление движения. Например: “kmopo” / “kmoto” «приходить из/с», “gwe” «уходить», “kon” «приходить», “go” «идти». Когда они встречаются с глаголами движения в сериальной глагольной конструкции, они сразу следуют за ним, если глагол движения переходный, следуют за прямым объектом.
- глаголы познания, восприятия и речи: переходные, и дитранзитивные глаголы могут указывать на познание, восприятие или речь. Эти глаголы обычно берут дополнитель, за которым следует придаточное предложение. Например: “sabi” «знать», “taki” «говорить», “aksi” «спрашивать», “sori” «показывать», “taigi” «рассказывать», “denki” «думать», “prakseri” «думать», “si” «видеть».

Пассивация ограничивается чаще всего глаголами, выражающими действия, которые произвольно контролируются агентом и оказывают прямое воздействие на пациента», например, «мыть», «сушить» и т.д. [Winford D. 1993:123]. Такие, глаголы, выражающие восприятия и психологические состояния, обычно сопротивляются пассивации. Существуют также ограничения, основанные на одушевленности, на то, что может проявляться в качестве пациентов в пассивной конструкции [Winford D., 2008:26].

Все виды глаголов указаны тегом «VB».

4.9.2 Связка

В СТ есть две связки: экквативная связка “na/a”, используемая в экквативных предложенных в настоящее время, и “de”, которая используется с обстоятельственными выраженями, в локативных и экзистенциальных конструкциях и в экквативных предсказаниях с маркерами время и вида.

Связка “de” выражает экзистенциальность. Если предшествует дополнению местности, это означает, что человек или объект находится в рассматриваемом месте. Ей могут предшествовать маркеры времен-вида и отрицатель [Winford D. et al., 2013]. Например:

nownow wroko no **de** [Nickel M. et al., 1984:16b]
there isn't any work now
сейчас работы нет

Основная функция “na” – выделять различные элементы предложения. Связка “na” может относиться к двум номинальным значениям: она выражает, что между двумя именными группами существует идентичность или что первый имеет свойство, описываемое вторым как показано в разделе экквативного предложения. Например:

den tu sma disi **na** Sranansma [Nickel M. et al., 1984:19a]
these two people are Surinamers
эти два человека суринамцы

Связка “na” не может встречается ни с маркерами времени и вида, ни с модальными глаголами (“mu” «должен», “kan” «мочь» и т.д.). Ей не может предшествовать отрицание, но есть специальная форма отрицательной связки “a no” [Arends J., 1989:30]:

en **a no** mi mati [Nickel M. et al., 1984:18b]
he is not my friend
он не мой друг

Когда надо указать время, вид или модальность, “na” заменяется на “de”. Например:

a ben de wan bun datra [Nickel M. et al., 1984:17c]
he was a good doctor
он был хорошим врачом

Связки “de” и “na” получают теги «EX» и «COP» соответственно.

4.9.3 Маркеры времени и вида

В CT глагольная форма не изменяется. Категории отрицания, времени, модальности и вида выражаются инвариантными свободными формами, все из которых появляются перед глаголом, за исключением пост-позиционного маркера совершенного вида “k(a)ba” [Winford D. et al., 2013].

Таблица 16: Основные категории времени / вида в CT

время	относительное прошедшего	ben
	прогнозируемое будущее	o
вид	несовершенный	e
	комплетив	k(a)ba
	совершенный	Ø (неотмеченный глагол)

Согласно Winford D., глагол без маркеров представляет ситуацию как не проанализированную ни по одному из параметров времени или вида. Следовательно, он может поддаваться различным интерпретациям в

дискурсе, в зависимости от контекста и задействованного предиката [Winford D., 2000:394]. В основном, если временной контекст делается явным (это может быть передано контекстом дискурса и наречиями, такими как “noyti” «никогда» и “ete” «пока»), он определяет значение глагола даже без необходимости использования маркеров. Но в тех случаях где временной контекст не указан, по умолчанию глагол имеет значение прошлого для нестатичиских глаголов и значение настоящего времени для статических глаголов и прилагательных, которые являются подклассом статических глаголов.

В CT есть система относительного времени, в которой локус времени может быть либо время выражения, либо другой точкой отсчета. Есть только две временные категории: (относительное) прошлое и будущее. В CT нет категории настоящего времени: ссылка на настоящее время передает вид несовершенный [Winford D., 2000:398].

Частица “ƀen” определяет ситуацию как прошлую по отношению к некоторой другой точке или интервалу времени (локус времени), который может быть либо моментом речи, либо некоторой точкой отсчета в прошлое. Прототипное использование слова “ƀen” состоит в том, чтобы локализовать некоторую ситуацию как происходящую до точки отсчета, находящейся в фокусе в дискурсе.

Di mi doro na oso esde, mi brada **ben** gwe [Winford D., 2000: 0 (E:HL 71)]
When I reached home yesterday, my brother had already left.
Когда я вчера приехал домой, мой брат уже уехал.

Маркер “o” передает ощущение относительно определенного или предсказуемого будущего или намерения в тех случаях, когда говорящий является агентом [Winford D., 2000:416]. В CT также используется

конструкция e + go, чтобы передать ощущение ближайшего или перспективного будущего [Winford D., 2000:418].

Вид выражает способ развития события, то есть, если событие пунктуально и оно полностью реализовано, он или длится, повторяется или скоро сбудется. Поскольку они представляют ситуацию в целом без учета ее внутренней структуры, глаголы без маркеров имеют вид совершенного вида [Winford D., 2000:395].

Комплетив “k(a)ba” передает значение «уже» и выражает смысл “завершенного” результата с не-статическими, и смысл состояния, начинающегося в прошлом и продолжающегося до ориентира в речи с статическими глаголами [Winford D. et al., 2013]. Winford D. утверждает, что, хотя комплетив “k(a)ba” является важной частью аспектуальной системы CT, похоже, он подвергается повторному анализу к наречию [Winford D., 2000:431]. В рамке этой работы, в этой функции “k(a)ba” считается наречием и размечается тегом «RB».

Частица “e” представляет ситуации (как состояния, так и события), рассматриваемые как “неограниченные” и продолжающиеся в время, то есть, повторяющиеся, привычные, текущие или продолжающиеся ситуации [Winford D. et al., 2013]. Winford D. подчеркивает, что “e” представляет чистую категорию вида, довольно нейтральную по отношению к привязке ко времени [Winford D., 2000:428].

Что касается порядок слов:

- слова, которые обозначают время, вид, модальность всегда появляются перед глаголом;
- сочетание нескольких маркеров времени, вида или режима выполняется в строгом порядке;
- Отрицание отмечается частицей “no”, помещаемой перед глаголом и всеми другими частицами.

Маркеры времени и вида получают тег «TAM».

4.9.4 Система модальность

СТ имеет богатую систему модальности, охватывающую ряд значений, связанных с типами возможности, обязательства и потребности. Возможные области включают усвоенные способности, выраженные “sabi” «знать», физические способности “man”, “kan” «может», разрешение “mag” «можно ли» и общие возможности “kan” «может» [Winford D. et al., 2013].

Таблица 17: Модальные вспомогательные глаголы [Winford D., 2000]

sa	неопределенное будущее	Неопределенное или гипотетическое будущее. Будущие ситуации (например, желание или надежда), не зависящие от говорящего или не зависящие от него.
musu	обязательство	Деонтическое значение слова «должно». Маргинальное чувство необходимости с определенными предикатами
kan	возможность	Деонтическая возможность, связана с моральным или социальным законом
man	способность	Деонтические способности, подчиняющиеся физическим или естественным законам
mag	разрешение	Деонтическая возможность, навязанная властью

Для анализа системы модальности СТ, Winford D. различает между эпистемической и деонтической модальностями. СТ грамматизировал множество модальных категорий, выражающих деонтические понятия, такие как обязательство, способность, основная возможность и т.д. и разработал другие стратегии для передачи эпистемических понятий, таких как возможность и вероятность [Winford D., 2000:121].

Понятия, относящиеся к деонтической модальности в CT, часто выражаются вспомогательными средствами, но в некоторых случаях лексическими глаголами или глагольными группами.

Обязательство [Winford D., 2000:70-75]: относится к существованию внешних социальных условий, вынуждающих агента выполнить предикатное действие. Вспомогательный “musu”, охватывает обязательство различной силы, “abi fu” используется только для сильного обязательства.

Необходимость [Winford D., 2000:75-76]: потребность выражается в CT конструкцией “abi NP fanowdu”, где NP относится к тому, что необходимо. В этой конструкции, некоторые говорящие используют вместо “fanowdu” слово “nodig” из нидерландского. В своей модальной функции “abi fanowdu” вводит предложение с “fu”.

Способность и основная возможность [Winford D., 2000:76-86]: в системе модальности CT есть грамматические различия между типами возможности (основная возможность, врожденная [физическая] способность и допустимость). Модальное использование “sabi (fu)” «знать (как)»- это просто расширение его функции в качестве глагола, означающего «знать». Его модальное значение ограничивается умственными способностями, за исключением случаев, когда умственные и физические способности идут рука об руку. Вспомогательный “man” также используется в первую очередь для выражения одного типа способности, включающего физические условия, обеспечивающие внутренние для агента или навязанные агенту силами вне его или его контроля (например, физические законы, несчастье или несчастный случай, травма, бедность и т.д.). Модальный вспомогательный “kan” выражает основную возможность, которая не ограничивается внутренним состоянием способности, но также сообщает об общих внешних условиях, таких как социальные или физические условия. Модальный

вспомогательный “mag” выражает чувство разрешения/допустимость «можно».

Лексический глагол “wani” «хочу» действует как основной глагол с именной группой как объект и как полу-вспомогательный глагол с модальной силой, чтобы выражать желание или намерение. Также глагол “lobi”, может действовать как полу-вспомогательный и выразить чувство симпатий и антипатий говорящему [Winford D., 2000:86-87].

Что касается порядок слов после “musu” и “kan” может следовать “ben”.

Отношение говорящего к содержанию высказывания (возможность, вероятность и предполагаемая уверенность) может быть выражено различными способами, просодические или паралингвистические признаки, наречия и модальные глаголы. Высказывания, связанные с эпистемической модальностью, передают степень приверженности говорящего истинности предложения.

Модальные “kan” и “musu” используются для передачи возможности и предполагаемой уверенности / вероятности соответственно [Winford D., 2000:92]. Они свободно появляются в таких конструкциях, как “kan / musu de taki S” «может / должно быть так, что S», где предложение S, в отношении которого говорящий выражает свою позицию, подчинено безличному главному предложению. Эпистемическая возможность выражается также при помощи наречия “kande” «возможно» [Winford D., 2000:94].

У говорящих есть и другие стратегии, чтобы выразить отсутствие приверженности истинности предложения. Например, они могут использовать такие выражения, как “volgens mi” «по-моему» или “gersi taki” «похоже» [Winford D., 2000:95].

Модальные указываются тегом «MD».

4.9.5 Отрицание

В СТ отрицание выражается оператором отрицания или отрицательной частицей “no”, которая непосредственно предшествует глаголу и его вспомогательным элементам [Migge B. et al., 2009:259].

No, yu **no** musu aksi a man tu. (Migge B., 2009:256).
No, you’re right, you surely shouldn’t ask him.
Нет, ты прав, тебе уж точно не стоит его спрашивать.

Маркеры времени, вида и модальности в отрицательных предложений, такие же как у положительной предложений. Что касается выражения отрицательных предложений с неопределенными местоимениями, последние встречаются с отрицанием сказуемого [Winford D. et al., 2013: feature 102]. В тех случаях, где отрицательный смысл передается неопределенным местоимением, как английский “*I saw nobody”* (никого не видел), СТ не допускает отсутствия отрицания сказуемого.

СТ отличается от таких языков, в которых отрицательно используемые неопределенные местоимения всегда исключают отрицание предикат, как, например нидерландский “*Niemand heeft opgebeld”* (никто не звонил). Например:

Noti **no** pasa nanga mi. [ApiCS, 2:167]
nothing neg happen with me.
Nothing happened to me.
Со мной ничего не случилось.

Для отрицании “no” есть свой тег «NEG».

4.10 Предикатные прилагательные конструкции

Некоторые лингвисты рассматривают предикатные прилагательные в СТ и других креольских языках как стативные глаголы. Один из аргументов, которые поддерживают позицию является отсутствие связки перед такими прилагательными. Следовательно, они подвергаются другому анализу чем атрибутивные прилагательные, непосредственно изменяющие существительное. С другой стороны, анализ Seuren P. (см. Seuren P., [1981]) рассматривает предикатные прилагательные в СТ как истинные прилагательные, которым предшествует нижележащей связкой ∅.

В некоторых случаях предикатным прилагательным может предшествовать связка “de”. Arends J. [1989:57] прилагательные, заимствованные из нидерландского, например, “enthousiast” «энтузиазм» и небольшая группа других прилагательных, при которых есть смысловая разница между “de” + прилагательное и ∅ + прилагательное. Например: “∅ bun” «быть хорошим» (stative) и “de bun” «быть здоровым» (non-stative).

Основываясь на принципах распределения, Sebba M. утверждает, что прилагательное может попасть в сказуемое либо в виде глагола, либо, скорее, в качестве члена подкласса категории глаголов; или как член класса прилагательных, который может функционировать до вербально как атрибутивное прилагательное, но может также появляться в предикате, если ему предшествует квантификатор [Sebba M., 1986:113]. Таким образом, в случае примера [Sebba M., 1986:12], “bradi” является прилагательным, поскольку перед ним стоит квантифиер/наречи. На оборот, в примере [Sebba M., 1986:13], наречие появляется после слово “bradi” и поэтому, в этом контексте, оно считается глаголом:

A liba de [so bradi] [Sebba M., 1986:12]
the river be so broad.
The river is so broad.
Река (очень) широкая.

A liba bradi [so]. [Sebba M., 1986:13]
the river broad so
The river is so broad.
Река (очень) широкая

В итоге, они размечаются как прилагательные «JJ» перед существительным и после экзистенциальной связкой “de”. В других положениях, считаются членами глагольного класса и получают тег «VB».

4.11 Сериальная глагольная конструкция

Под понятием “сериальная глагольная конструкция” (SVC) понимается конструкция с последовательностью из двух или более конечных глаголов V1, V2..., функционирующих как единый предикат и описывающих единое событие, где [Sebba M., 1987:39]:

- и V1, и V2 должны быть лексическими глаголами, т.е. должны быть способны выступать в качестве единственного глагола в простом предложении;
- и V1, и V2 должны интерпретироваться как имеющие одно и то же время и вид. Таким образом, например, V1 может не интерпретироваться как прошлое, если V2 интерпретируется как будущее;
- между V1 и V2 не должно быть установленной границы раздела и они должны находиться в одном разделе;
- никакие союзы не должны разделять глаголы последовательно.

Как говорилось в предыдущей главе, СТ беден словообразовательными морфемами. СТ использует «сериальные глагольные конструкции» (SVC) для увеличения ссылочной способности своего лексикона. Там, где другие языки будут использовать аффикс, добавляемый к простому глаголу или другому лексическому элементу [Sebba M., 1984:37], СТ может расширить свой лексикон, используя второй глагол сериальной конструкции . Например:

bro «дуть» + kiri «убывать» → broko kiri «тушить»
koti «вырезать» + puru «снять» → koti puru «ампутировать»

Sebba M. [1984:37] отмечает, что лексика СТ также весьма бедна предлогами, особенно предлогами, относящимися к пространственным отношениям. За исключением “fu”, все предлоги, обозначающие пространственные отношения, имеют форму “na” + именная группа (NP), где N - одно из небольшого набора существительных, включая “ini” «внутри», “baka” «назад», “ondro” «снизу», “sei” «сторона» и т. д. Предлог “na” является нейтральным по направлению, и в СТ нет других предлогов, как «к» и «от», «внутрь» и «от» указывающих направление движения выражаемое первым глаголом. В SVC второй глагол указывает направление движения, выраженное первым. Некоторые глаголы, которые используются, чтобы указывать направление являются следующими: “kon” «приходить» или “go” «идти» “fadon” «упасть», “komoto/komopo” «из/с», “poti” «положить», “puru” «снять». Примеры:

A tyari a ston komoto na ini a oso [Sebba M., 1981:10]
He brought the stone out of the house
Он принес камень в дом

A e dansi go na ini a oso [Sebba M., 1981:12]
S/he is dancing into the house
Он/а танцует в дом

4.12 Сравнительная степень

Сравнение выражается описательным предложением с использованием “moro” или “p(a)sa”. Что касается слова “moro” есть две модели, включающие порядок слов прилагательных и сравнительной степени. Winford D. et al. объясняют два возможных варианта последовательности “moro bigi” и “bigi moro” на основе различия между одним более консервативным диалектом и другим вариантом, возникающим под влиянием нидерландского. В тех случаях, когда упоминается только сравниваемый элемент, например, “mi brada moro langa” (мой брат выше) или “mi brada na a moro langa wan” (мой брат самый высокий), сравнительный маркер “moro” предшествует прилагательному и может интерпретироваться как обозначение прилагательного. Этот образец общий для обоих вариантов [Winford D. et al., 2013]. Но в случаях, когда присутствуют и сравниваемый элемент, и стандарт, слово “moro” может либо следовать за элементом свойства, либо предшествовать ему. Когда “moro” предшествует элементу свойства, его можно интерпретировать как маркер сравнительной степени. Например:

John moro bigi moro Peter.
John more big exceed Peter
John is taller than Peter.
Джон выше Питера.

Но, когда “moro” следует за элементом свойства, интерпретируется как маркер стандарта:

John bigi moro Peter.
John big exceed Peter.
John’s bigger than Peter
Джон больше Питера

Слово “p(a)sa” при использовании в сравнении всегда следует за прилагательным. Например:

Kenneth bigi psa tvee meter
Kenneth big pass two meter
Kenneth is taller than two meters
Кеннет выше двух метров

Когда они функционируют как сравнительные, слова «moro» и «p(a)sa» помечаются как «COMP».

4.13 Наречие

О категории «наречии» в CT мало написано. Наречия могут изменять целое предложение, прилагательное или квантификатор. Nickel M. et al. [1984:41] перечисляют наречия, которые могут изменять прилагательное или квантификатор: “lekti” «легкий», “bun” «очень», “tumsi” «слишком», “moro” «более». В этом случае, они встречаются перед прилагательным или квантификатором. Наречия выделяются тегом «RB».

Наречия, такие как “ala ten” «всегда», “noiti” «никогда» и т.д., rак правило, помещаются в начальную позицию предложения [Winford D. et al., 2013:11]. Поскольку они имеет другое распределение, чем другие наречия, в рамках этой работы они классифицируются в категории «номинальное наречие» и указываются тегом «RN».

Локусы d(r)ape «там» и dja или djaso «здесь» считаются обычными наречями.

4.14 Локативные конструкции

Помимо вышеупомянутых локусов, есть локусы с объектом определения (существительное, группа существительных или местоимение), которые вводится предлогом [Voorhoeve J., 1956:195]. Пространственные отношения в современном СТ выражаются через широкий спектр конструкций. Некоторые из них довольно четко отражают более недавнее влияние нидерландской суперстраты, в то время как исходная система явно отражает влияние языков-субстратов из западной Африки. СТ позволяет использовать весь спектр локативных структур: и встречающихся в языках-субстратах и в суперстрате [Yakpo K. et al., 2015:171-172].

СТ использует сложные конструкции с указанием места, в которых описания движения совместно реализуются глаголами, предлогами и существительными с указанием места [см. Yakpo K. et al., 2015].

Базовые локативные конструкции, которые отвечают на вопросы «где» обычно выражается с помощью именной группы, введенной локативным предлогом “na”, который служит только для введения других слов и не дает никакой информации о типе локализации. Слово “na” можно рассматривать как основной предлог в местности. Определение местоположения без этого предлога в принципе невозможно (Voorhoeve J.). Например:

a de na wasi-oso [Yapko K. et al., 2015:142]
S/he is in the bathroom
Он/а в ванной

Однако употребление общего местного предлога “na” далеко необязательно. Локативные конструкции, вводимые “na” (или его вариантом “a”), столь же распространены, как и те, в которых местный предлог отсутствует [Yakpo K. et al., 2015]. Местное отношение к объекту может быть дополнительно указано путем добавления локативных элементов “ini” «внутри», “tapu” «вверх», “fesi” «спереди», “baka” «сзади», “mindri” «в середине», «между», “sei” «рядом с». Набор локативных элементов в современном CT в основном состоит из слов английского происхождения, с меньшинством нидерландского происхождения.

Таблица 18: Локативные элементы в CT (Yakpo K. et al., 2015:137)

локативный элемент	значение	значение (русский)	язык источника
ini	inner part, in	внутренняя часть, в	“in” [англ. / нидер.]
na doro	outside	вне	“LOC door” [англ.]
tapu	top, on	сверху, на	“(on) top (of)”
ondro	bottom, under	внизу, под	“onder / under” [англ. / нидер.]
fesi	face; in front	перед, напротив	“face”
baka	back, behind	сзади	“(at the) back (of)”
fu	general location; source oriented	ориентированный на источник	“for” [англ.]
na	general location	общее расположение	“na”

Yakpo K. et al. отмечают, что хотя локативные элементы CT сохранили большую часть фонологической формы и значительную часть лексической информации своих английских этимонов, в то же время они претерпели морфосинтаксическую переклассификацию с предлога на местное существительное (“ini” «внутри») или нарицательное существительное на местное существительное (fesi «перед») [Yapko K. et al., 2015:143].

Несмотря на то, что описано выше, эти локативные элементы могут использоваться как предлоги в результате недавнего развития, вызванного контактом с нидерландским языком. Следовательно, локативные элементы могут быть классифицированы как и существительные и предлоги, в зависимости от конструкции, в которой они встречаются [Plag I., 1998].

Локативные элементы представлены тегом «INL» когда действуют в качестве предлогов и тегом «NN» в тех случаях они являются существительнами. Общему локативному предлогу “na” присвоен тег «LOC».

4.15 Нелокативные предлоги

Помимо предлогов, вводящих дополнения и обозначающих период времени, есть несколько других понятий, которые выражаются с помощью предлогов. Например, предлог “anga” «с» вводит три смысловые роли:

- инструмент, с помощью которого выполняется действие
- способ, которым осуществляется деятельность
- служба поддержки

Предлог “fu” может указывать на притяжательные отношения между двумя номинальными группами. В этом случае, “fu” получает тег «OF».

Yakpo K. утверждает, что контакт с нидерландским отвечает за повторный анализ глагола “gi” «давать» как предлога, эквивалентного английскому «to» в дитранзитивных конструкциях [Yakpo K., 2017:75-76].

Задача предлога “gi” состоит в том, чтобы ввести несколько семамических ролей, таких как:

- человек, который что-то получает;
- бенефициант деятельности;
- тот, кто испытывает эмоции.

Все нелокативные предолгие кроме “fu” в притяжательных конструкций указаны тегом «IN».

4.16 Междометие

Междометие присвоены теги «UH». В эту категорию входят частицы “tog”, указывающие на риторические вопросы.

Глава 5. Построение POS-теггера

В этой главе рассматриваются три основных подхода POS-теггирования и обсуждаются те преимущества и сложности, которые представляет каждый из них при применении для СТ. Затем будет предложен метод построения POS-теггера для СТ, который сочетает некоторые элементы предыдущих представленных подходов. Предлагаемый теггер предназначен для решения некоторых особых характеристик языка и для работы с недостатком данных и ресурсов.

5.1 Теггеры основанные на правилах

Теггеры, основанные на правилах, используют словарь или лексикон, чтобы получить возможные теги для каждого слова. Если слово имеет более одного возможного тега, то теггер использует правила, созданные вручную для определения правильного варианта. Снятие неоднозначности также может выполняться путем анализа лингвистических характеристик слова вместе с предшествующими и последующими словами. Например, в случае СТ, если предыдущее слово – артикль, то слово будет являться, скорее всего, существительным или прилагательным.

Процесс POS-тегирования на основе правил проходит в два этапа:

1. на первом этапе теггер использует словарь для присвоения каждому слову списка потенциальных частей речи;
2. на втором этапе теггер использует большие списки рукописных правил для устранения неоднозначности, чтобы отсортировать список до одной части речи для каждого слова.

Подход на основе правил работает полностью по правилам, созданным вручную лингвистами. Поэтому, этот метод требует знания специалистов, является дорогостоящим и трудоемким. Хотя СТ имеет довольно строгий

порядок слов (с влиянием синтаксиса нидерландского языка), некоторые конструкции в СТ подвержены изменению порядка слов (см. сравнительную степень и предлоги с местным падежом в четвертой главе), что усложняет процесс формулировки правил вручную.

5.2 Брилл-теггер (Transformation-based tagging)

Теггер на основе преобразования Transformation-based tagging, иногда называемый «теггером Brill» - это реализация Transformation-based Learning (обучения на основе преобразования) (TBL), который представляет собой алгоритм, основанный на правилах для автоматического POS-теггирования. Как и классические теггеры, основанные на правилах, TBL применяет правила, чтобы определить, каким словам какие теги должны быть назначены. Но, в отличии от них, TBL является методом машинного обучения, в котором применяемые правила автоматически извлекаются из размеченных данных.[12]

В основном, алгоритм TBL изучает правила для разметки в три этапа [Juravsky D., 2008:152]:

- прежде всего алгоритм маркирует каждое слово наиболее вероятным тегом;
- затем он исследует все возможные преобразования и выбирает то, которое дает наиболее лучшую маркировку;
- наконец, он повторно помечает данные в соответствии с этим правилом;
- TBL повторяет последние два этапа, пока не достигнет некоторого критерия остановки, такого как недостаточное улучшение по сравнению с предыдущим проходом.

12 Чтобы лучше понять, как работает теггер, читатель отсылается к статям автора (Brill Eric, 1992, 1994, 1995).

Результатом процесса TBL является упорядоченный список преобразований; затем они составляют «процедуру маркировки», которая может быть применена к новому тексту. Текст сначала маркируется в соответствии с самым широким правилом, то есть таким, которое применяется в большинстве случаев. Затем выбирается более конкретное правило, которое меняет некоторые исходные теги (некоторые из которых могут быть ранее измененными тегами).

Ограничением использования этого теггера для СТ является отсутствие корпуса. Несмотря на это, идея предварительного присвоения тега словоформе и последующей его замены серией преобразований может быть полезна для преодоления некоторых трудностей при работе с СТ, что и обсуждается далее в этой главе.

5.3 Стохастические теггеры

В принципе, любую модель, которая включает частоту или вероятность (статистику), можно назвать стохастической. Стохастические теггеры учитывают распределение вероятностей появления POS-тегов в последовательности предложения. Они получают высокую степень точности, не полагаясь на чистый синтаксический анализ входных данных, но требуют наличие корпуса для вычисления вероятностей.

Самые простые стохастические теггеры применяют один из следующих подходов к POS-тегам:

- подход основан на частоте слова: стохастические теггеры устраняют неоднозначность слов на основе вероятности того, что слово встречается с конкретным тегом. Тег, который чаще всего встречается со словом в обучающем наборе - это тег, присвоенный неоднозначному экземпляру этого слова. Основная проблема такого подхода заключается в том, что он может привести к недопустимой

последовательности тегов. Другой недостаток этой модели заключается в отсутствии вероятности для слов, которых нет в корпусе.

- вероятности последовательности тегов: теггер вычисляет вероятность появления заданной последовательности тегов. Его также называют подходом «N-грамма». Он называется так, потому что лучший тег для данного слова определяется вероятностью, с которой он встречается с n предыдущими тегами.

Преимущество стохастических методов заключается в простоте получения статистических данных по сравнению с правилами, заданными вручную. Например, модель N-грамм для частей речи может быть подходящим решением для моделирования последовательности тегов в СТ.

5.4 Скрытая Марковская модель

Скрытая Марковская Модель (СММ) – это стохастический классификатор последовательности, который широко и успешно используется для частеречной разметки во многих языках. В основном, модель сочетает в себе оба статистических подхода из предыдущего раздела: вероятности последовательности тегов и вероятности частоты слова.

СММ может быть определена как стохастическая модель с двойным вложением, в которой лежащий в основе стохастический процесс скрыт. Этот скрытый случайный процесс можно наблюдать только с помощью другого набора случайных процессов, который производит последовательность наблюдений.

Задача СММ – найти для любой строки словоформ (наблюдаемых состояний) наиболее вероятную последовательность частей речи (скрытые состояния).

Из множества последовательностей тегов алгоритм выбирает такую последовательность, которая наиболее вероятна при наблюдаемой последовательности слов [Jurafsky D., 2008:139]:

$$\hat{t}_1^n \approx argmax\, P(t_1^n | w_1^n) \quad (1)$$

где из всех последовательностей тегов длины *n* алгоритм ищет конкретную последовательность тегов, которая максимизирует правую часть уравнения. Применяя правило Байеса можно разбить любую условную вероятность на три другие вероятности, которые легче вычислить:

$$\hat{t}_1^n \approx argmax \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)} \quad (2)$$

где $P(t^n_1)$ априорная вероятность появления тега t_1 и $P(w^n_1|t^n_1)$ вероятность наступления слова при истинности тега и $P(w^n_1)$ полная вероятность наступления слова.

Поскольку знаменатель является константой, которая не меняется для каждой последовательности тегов, его можно опустить:

$$\hat{t}_1^n \approx argmax\, P(w_1^n | t_1^n) P(t_1^n) \quad (3)$$

где первый элемент представляет собой вероятность слова $P(w^n_1|t^n_1)$, а второй – априорную вероятность последовательности тегов $P(t^n_1)$.

Подводя итог, наиболее вероятная последовательность тегов $\hat{t}^n_1$ для некоторой строки слов w^n_1 вычисляется путем умножения двух вероятностей для каждой последовательности тегов и выбора последовательности тегов, для которой это произведение является наибольшим. Два термина – это: априорная вероятность последовательности тегов $P(t^n_1)$ и вероятность строки слов $P(t^n_1|t^n_1)$.

Чтобы сделать вычисления возможными, модель делает два упрощающих предположения. Первое предположение состоит в том, что вероятность появления тега зависит только от предыдущих *n* тегов, а не от всей последовательности тегов. В случае триграммной моделью оно представляет собой вероятность тега *t* с учетом предыдущих 2 тегов.

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i | t_{i-2}, t_{i-1}) \quad (4)$$

Этот параметр называется вероятностью перехода и вычисляется на основе корпуса следующим образом:

$$P(t_i | t_{i-2}, t_{i-2}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})} \quad (5)$$

где числитель указывает сколько раз в корпусе встречается цепочка $t_{i-2}\, t_{i-1}\, t_{i-1}$ и знаменатель, который обозначает количество раз, когда в корпусе встречается цепочка из идущих подряд тегов $t_{i-2}\, t_{i-1}$. Вероятности перехода соответствуют модели 3-грамм, описанной в предыдущем разделе.

Второе предположение CMM состоит в том, что вероятность появления слова зависит только от его собственного тега части речи, то есть она не зависит от других слов вокруг него и от других тегов вокруг него.

$$P(w_1^n | t_1^n) \approx \prod_{i=1}^{n} P(w_i | t_i) \quad (6)$$

Этот параметр называется вероятностью результата и представляет собой вероятность того, что данный тег будет связан с данным словом. В корпусе вычисляется:

$$P(w_i | t_i) = \frac{C(t_i, w_i)}{C(t_i)} \quad (7)$$

где числитель указывает на то сколько раз тег t_i отмечает слово w_i ; а знаменатель показывает сколько раз тег t_i появляется в корпусе. В основном, это эквивалентно подходу, основанному на частоте слова из предыдущего раздела.

Объединение этих двух параметров приводит к следующему уравнению, по которому теггер оценивает наиболее вероятную последовательность тегов [Jurafsky D., 2008:141]:

$$\hat{i}_1^n = argmax\, P(t_i^n | w_i^n) \approx argmax\, [\prod_{i=1}^{n} P(w_i|t_i)\, P(t_i|t_{i-1}, t_{i-2})]\, P(t_{n+1}\, t_n) \qquad (8)$$

где последний элемент - это маркер конца последовательности для t_{n+1}.

Процесс обнаружения последовательности скрытых состояний с учетом последовательности наблюдений известен как декодирование. Для декодирования обычно используется алгоритм Витерби [Jurafsky D., 2008:184].

СММ обучаются на больших корпусах. Если обучающий корпус недостаточно большой, не только разреженность данных будет влиять отрицательным образом на вычисленные вероятности, но также в ней будет представлена только часть словарного запаса языка. Поскольку СТ не имеет большого корпуса для обучения модели, частеречная разметка на основе СММ не может быть реализована для этого языка.

5.5 Гибридный теггер

С одной стороны, недостаток знания о языке для формулирования правил для снятия неоднозначности вручную, а с другой - отсутствие корпуса для применения чистого стохастического метода, заставляет выбирать средний путь между обоими подходами для построения POS-теггера.

Предлагаемый теггер является гибридным. Он использует лексикон и подходы, которые основаны на правилах. В лексиконе пара слов и их возможные теги хранятся следующим образом:

$w_1 \leftarrow t_1, t_2, t_3, t_4$

$w_2 \leftarrow t_3,$

$w_3 \leftarrow t_1, t_2, t_3$

$w_4 \leftarrow t_2, t_4$

…

где элемент слева стрелки — уникальная словоформа, а справа — набор тегов, которые данная словоформа может принимать. Если пара словоформа-часть речи однозначна (как у w_2), вероятность появления тега данной словоформы равна 1. Теги, которые не указаны для данной словоформы, имеют вероятность 0 и, следовательно, они не могут появиться с ней. Но если словоформа имеет более одного назначенного тега, то каждому тегу дается вероятность $p(t|w)$, где совместная вероятность равна 1. Способ вычисления этой вероятности поясняется в разделе «Вероятность словоформ / тегов» в конце этой главы.

В отличие от подхода, основанного на правилах, предлагаемый гибридный теггер устраняет неоднозначность части речи не в соответствии с правилами, а с использованием модели POS тегов 3-грамм. В рамках этой работы предполагается, что размер корпуса не критичен для вычисления 3-грамм тегов. 3-граммы могли быть вычислены даже на очень скромном наборе данных, размеченных вручную, с приемлемыми результатами при условии, что он охватывает наиболее распространенные последовательности частей речи языка. Очевидно, что чем больше обучающих данных, тем лучше будут работать модели, но предполагается, что можно достичь работающей модели даже в случае небольшого набора.

В итоге, гибридный теггер полагается на лексикон для присвоения тегов заданным словоформам и на 3-граммную модель, обученную на небольшом наборе предложений для устранения неоднозначности данных тегов в контексте предложения.

Чтобы найти более вероятную последовательность тегов, гибридный теггер использует ту же формулу, что и СММ, с той разницей, что вероятность результатов *p(w|t)* заменена на вероятность *p(t|w)* из лексикона.

$$\hat{i}_1^n = argmax\, P(t_i^n | w_i^n) \approx argmax\, [\prod_{i=1}^{n} P(t_i|w_i)\, P(t_i|t_{i-1}, t_{i-2})]\, P(t_{n+1}\, t_n) \qquad (8)$$

где $p(t_i|w_i)$ — вероятность, что тег t_i встречается при условии словоформа w_i. Декодирования также при применении алгоритма Витерби.

5.6 Составление лексикона

Задача составления лексикона для СТ представляет некоторые трудности. Помимо наличия ограниченных надежных источников для извлечения пар слов-тегов, существуют трудности с составлением лексикона (какие именно слова следует включать в него). Во-первых, трудно решить, является ли одно слово заимствованием или может считаться частью словарного запаса СТ. Словарный запас СТ невелик [Sebba M., 1984:36], поэтому говорящие на СТ чаще используют слова из нидерландского, чтобы заполнить лексические пробелы. Следовательно, любое слово из нидерландского может потенциально вступать в СТ, даже предлоги и союзы. Во-вторых, несмотря на попытки стандартизировать язык, по-прежнему существуют многие варианты написания слов. По этим причинам и независимо от размера лексикона ожидается большое количество словоформ, которые в него не входят.

В языках обычно есть относительно небольшой набор слов, которые часто встречаются и действуют как функциональные слова. Артикли, предлоги, местоимения принадлежат к этому классу слов. Вполне вероятно, что любой конкретный говорящий будет их знать, или любой корпус будет содержать общий набор слов этого типа. Они называются словами закрытого класса и обычно выполняют структурирующую роль в грамматике, определяя окружающие слова [Jurafsky D., 2008:125]. Что касается слов закрытого класса, даже небольшой лексикон способен охватить большинство из них.

Существительные, глаголы, прилагательные, наречия и междометия относятся к так называемому открытому классу слов, потому что новые элементы постоянно создаются или заимствуются из других языков. По тому же определению этого понятия, невозможно создать лексикон (независимо от того, насколько большой), который мог бы содержать все слова из открытого класса. Поэтому надо иметь решение, чтобы разобраться со случаями, когда слова нет в лексиконе.

Самый простой способ справиться с неизвестными словами - это принять, что каждое неизвестное слово неоднозначно среди всех возможных тегов с равной вероятностью. Тогда теггер должен полагаться исключительно на контекстные POS 3-граммы, чтобы предложить правильный тег [Jurafsky D., 2008:158]. Но в случае CT потенциальное количество слов вне лексикона настолько велико, что это может испортить способность модели POS 3-грамм снимать неоднозначности.

5.7 Присвоение тегов

Чтобы избежать проблемы с неизвестными словами, алгоритм применяет ряд предположений. Во-первых, все слова рассматриваются как находящиеся вне лексикона. Как и в случае Transformation-based learning (TBL), они предварительно размечаются с наиболее вероятным вариантом, в

то время как следующие процедуры будут выполнять необходимые преобразования, чтобы пытаться найти более подходящий тег. Однако, в отличие от метода TBL, который назначает наиболее вероятный тег, здесь применяется список, содержащий теги открытого класса слов: существительное, глагол, прилагательное и наречие. Предполагается, что слово, которого нет в лексиконе, скорее всего, принадлежит к одной из этих частей речи. Предварительное назначение этих тегов - это попытка (хотя и очень упрощенная) имитировать феномен многофункциональности слова.

После предварительного присвоения тегов, алгоритм пробует найти словоформу в лексиконе. Если словоформа найдена, то предварительно заданный список тегов заменяется списком тегов, указанных в лексиконе. Ожидается, что слова закрытого класса, которые играют структурирующую роль в грамматике, будут идентифицированы лексиконом.

Если поиск не прошел успешно, алгоритм проверяет, начинается ли слово с заглавной буквы. В случае, если это условие не выполняется, алгоритм принимает предварительный список тегов как самый лучший вариант и возвращает его. Но, если слово начинается с заглавной буквы, алгоритм смотрит на его положение в цепочке предложений. Если слово не находится в самом начале предложения, тогда подразумевается, что оно является именем собственным и, следовательно, предварительный список тегов заменяется тегом для имен собственных. Но, если слово оказывается в самом начале предложения, тогда алгоритм не может принять заглавную букву исключительно как признак имен собственных. В этом случае, алгоритм возвращает предварительный список тегов плюс тег имен собственных.

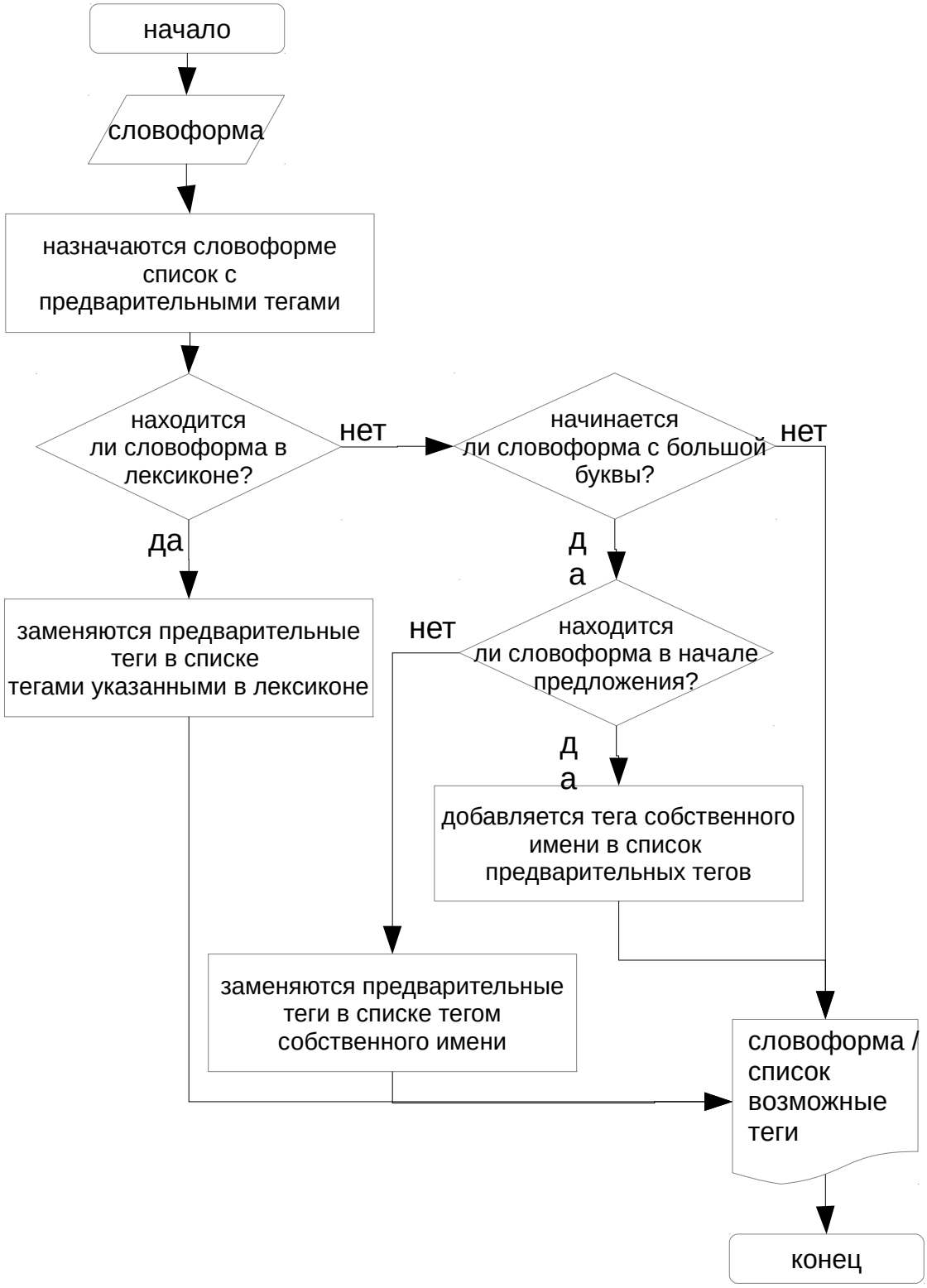


Рисунок 2: Блок схема присвоения POS тегов

5.8 Лексическое знание

Лексикон - это не просто набор слов и соответствующих возможных им тегов, это также это способ хранения лексических знаний СТ и последующего использования их в модели. Путем определения того, какие теги могут появляться с данной словоформой, записи лексикона заменяют более расплывчатые предположения из предварительного списка тегов. Если одна словоформа, например, “ripoliku” - «республика», более вероятно встречается в текстах СТ только с определенными тегами (в этом случае с существительным), то это можно закодировать в лексиконе следующим образом:

“ripoliku” ← существительное

Мультифункциональность слов является распространенным явлением в СТ, но не совсем продуктивным [Sebba M., 1986:116]: есть некоторые ограничения в зависимости от конкретного слова / группы слов. Например, хотя существует подмножество слов как “singi” «песня/петь» и “lobi” «любовь/любить», которые могут действовать в качестве существительных и глаголов, не все глаголы могут выступать как существительные. Слова как “go” «идти» и “kon” «приходить» являются только глаголами. Это можно закодировать в лексиконе следующим образом:

“singi” ← существительное, глагол
“go” ← глагол

Такие слова как маркеры времени и вида, союзы, несмотря на тот факт, что они с меньшей вероятностью будут выполнят другие функции в синтаксисе (если это возможно), все же могут иметь омонимы. СТ допускает

большое количество омонимов, возникающих из-за потери фонематических различий при передаче с лексификатора языка. Например, словоформа “sa” чаще всего встречается в качестве модальной частицы («был бы», из нидерландского “zal”), но есть омонимы “sa” (из английского “saw”, “to saw”) - «пила» и «пилить». Следовательно, омонимия должна быть представлена в лексиконе таким же способом:

“sa” ← модальный глагол, существительное, глагол

5.9 Вероятность словоформ / тегов

Как было указано ранее, вероятность того, что словоформа w_i соответствует тегу t_i, возможно вычислять только в большом корпусе. Но, даже в небольших данных ожидается (при условии, что набор включает разные синтаксические конструкции), что она будет содержать репрезентативное распределение тегов для частей речи. Следовательно, можно подсчитать теги на основе обучающей выборки и каким-то образом использовать значения этих подсчетов для вычисления вероятностей появления присвоенных тегов для одной словоформы. Для этого предлагаются три метрики, принцип которых объясняется ниже, а его влияние в результате частеречной разметки проверяется в следующей главе.

Первая метрика (A) переводит частоту тегов в обучающих данных в контекст списка данной словоформы. Теги с более высоким числом имеют более высокую вероятность и наоборот.

$$P(t_{i,n}) = \frac{tf_{i,d}}{\sum_{i=1}^{n} tf_{i,d}} \qquad \text{(A)}$$

где n — количество тегов в списке и $tf_{i,d}$, число вхождений тега t_i в обучающем выборке d. В знаменателе представлена сумма всех частот тегов

$tf_{i,d}$ в списке n. Например, если список для данной словоформы w содержит t_1, t_2 и t_3, t_1 встречается в обучающих данных 100 раз, а t_2 и t_3 всего 20 и 10 соответственно, то вероятности тегов будут 0,77, 0,15 и 0,08.

Альтернативная метрика (B) в целом почти такая же, как и (A), но с использованием натурального логарифма, сглаживающего разницу между счетчиками тегов:

$$P(t_{i,n}) = \frac{\ln(tf_{i,d})}{\sum_{i=1}^{n} \ln(tf_{i,d})} \qquad \text{(B)}$$

где значение натурального логарифма частоты тега в числителе нормируется на сумму всех натуральных логарифмов частоты тегов в списке. Как результат, вероятности тегов t_1, t_2 и t_3 из предыдущего примера теперь находятся ближе друг к другу: 0,47, 0,30, 0,23.

Третья метрика (C) снижает частоту, делая теги с более низкими показателями более вероятными:

$$P(t_{i,n}) = \frac{\sum_{i=1}^{n} tf_{i,d} - tf_{i,d}}{\sum_{i=1}^{n} tf_{i,d}\, n - 1} \qquad \text{(C)}$$

где числитель - это разница между суммой всех частот тегов $tf_{i,d}$ в списке и частоты $tf_{i,d}$, в данном случае теги с меньшим подсчетом получают более высокие значения. Знаменатель содержит нормирующую постоянную. Теги t_1, t_2 и t_3 из двух предыдущих примеров получат вероятности 0,12, 0,42 и 0,46 соответственно.

В таблице 19 показаны теги, которые может принимать словоформа “moro” «больше», их частота в обучающем наборе из 3890 тегов и вероятности, рассчитанные на основе предложенных метрик:

Таблица 19: Пример вероятностей тегов для словоформы "moro" на основе разных метрик

тег	частота	часть речи	(A)	(B)	(C)	(D)
RB	129	наречие	0.66	0.42	0.16	0.59
AB	48	квантификатор	0.24	0.33	0.37	0.28
COMP	16	сравнительная степень	0.08	0.24	0.45	0.11

Глава 6. Тестирование модели

В этой главе представлены план и результаты эксперимента, предназначенного для оценки производительности POS-теггера для CT, обученного на небольших выборках с использованием ограниченного лексикона. Целью эксперимента является оценка вклада данных и метрик на предсказания модели.

6.1 Набор POS тегов

Набор POS тегов, используемых в эксперименте, состоит из 31 элемента в соответствии с критериями, описанными в четвертой главе.

Таблица 20: Набор POS тегов

закрытый класс слов		открытый класс слов	
AB	квантификатор	JJ	прилагательное (атрибутивное)
AP	спецификатор	NN	существительное
AT	артикль	NP	имя собственное
COMP	сравнительная степень	NUMB	числительное
COMPL	комплементатор	RB	наречие
COP	связка	RN	номинальное наречие
DT	указательное местоимение	UH	междометие
EX	экзистенциальная связка	CC	союз
FOC	маркер фокуса	CS	подчиняющий союз
IN	предлог	VB	глагол
INL	предлог места		
LOC	местный предлог		
MD	модальный		
NEG	отрицание		
OF	притяжательная связка		
PRN	местоимение		
PPL	возвратное местоимение		
PP$	притяжательное местоим.		
REL	относительное местоим.		
TAM	маркер времени и вида		
WP	вопросительное слово		

6.2 Лексикон

Лексикон был составлен на основе лексических записей из словарей Wilner J. [2007] (см. раздел A1. «Лексический ресурс, использованный для построения лексикона» в приложении). Метод построения следует трем правилам:

1. лексикон стремится содержать все словоформы из слов закрытого класса. Таким образом, лексикон включает все артикли, местоимения, маркеры времени и вида и т.д.;
2. лексикон также старается включить наиболее употребляемые слова из подчиненных союзов, междометий и номинальных наречий;
3. словоформы из открытого класса слов, такие как прилагательные, существительные, наречия и глаголы включаются в лексикон при условии, что для них есть омоним в словоформах закрытого класса. Например, у модального “musu” «должен» есть омоним в существительных “musu” «берет». Следовательно, лексикон содержит “musu” в качестве модального и существительного.

Составленный лексикон для эксперимента состоит из 346 словоформ с 500 тегами. Хотя словари Wilner J. включают около 3231 разных словоформ, в рамках данного эксперимента интересно проверить производительность алгоритма теггера с минимальным лексиконом. В разделе приложения A1. «Лексический ресурс, использованный для построения лексикона» находится общее описание словарей Wilner J.

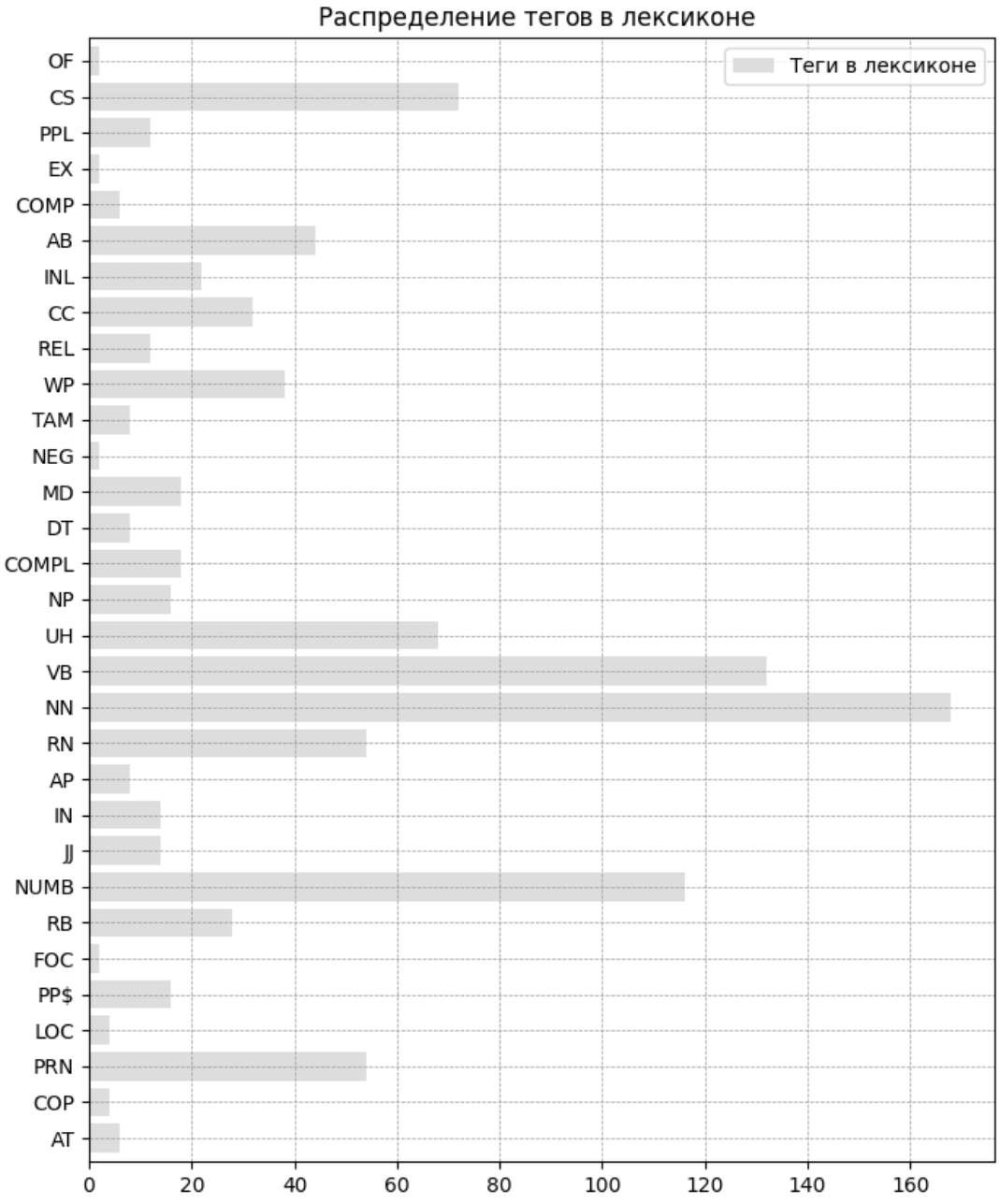


Рисунок 3: Распределение тегов в лексиконе

6.3 Обучающая выборка

Обучающая выборка состоит из предложений/примеров, извлеченных из набора данных APiCS [Winford D. et al., 2013] и описания языка [Nickel M. et al., 1984] и размеченных вручную на основе вышеупомянутого набора тегов. Это решение основано на том факте, что и работа Winford D. et al. и работа Nickel M. et al. являются систематическим описанием языка и, следовательно, они должны отображать большинство его грамматических особенностей путем изучения относительно небольшого набора примеров.

Обучающая выборка была разделена на четыре части, чтобы создать на основе их четыре стадии обучения модели. Они предназначены для наблюдения за развитием производительности модели с добавлением данных.

Первые две части содержат примеры из базы данных APiCS [Winford D. et al, 2013] и последние два из [Nicket M. et al, 1984]:

1. TD1: база данных APiCS: 165 предложений, 1381 токен
2. TD2: база данных APiCS: 164 предложения, 1472 токена
3. TD3: Nickel et al.: 111 предложений, 827 токенов
4. TD4: Nickel et al.: 110 предложений, 833 токена

Из рисунка 3 и таблиц 21 и 22 (см. внизу) видно, что хотя образцы похожи по содержанию, такие теги как «AB» (квантификатор), «JJ» (атрибутивное прилагательное), «NP» (имя собственное) и «OF» (притяжательная связка) имеют относительно больший вес в образцах из Nickel M. et al.. В образцах из базы данных APiCS, которых в среднем в 1.71 раз больше чем таких из Nickel M. et al., такие теги как «NN» (существительно), «AT» (артикль) имеют относительно больший вес. Это говорит о том, что такие части речи, которые чаще встречаются в предложениях растут быстрее с большим количеством данных.

Таблица 21: Процент POS тегов с большой частотой по образцу

	Токены	VB	NN	PRN	AT
TD1	1381	216 (16%)	203 (15%)	151 (11%)	131 (9%)
TD2	1472	225 (15%)	217 (15%)	148 (10%)	137 (9%)
TD3	827	121 (15%)	106 (13%)	101 (12%)	70 (8%)
TD4	833	135 (16%)	107 (13%)	111 (13%)	51 (6%)

Таблица 22: Процент POS тегов с небольшой частотой по образцу

	Токены	AB	JJ	NP	OF
TD1	1381	10 (0,7%)	18 (1,3%)	29 (2%)	10 (0,7%)
TD2	1472	15 (1%)	20 (1,3%)	32 (2,1%)	10 (0,6%)
TD3	827	10 (1,2%)	19 (2,2%)	29 (3,5%)	9 (1%)
TD4	833	13 (1,5%)	20 (2,4%)	24 (2,8%)	11 (1,3%)

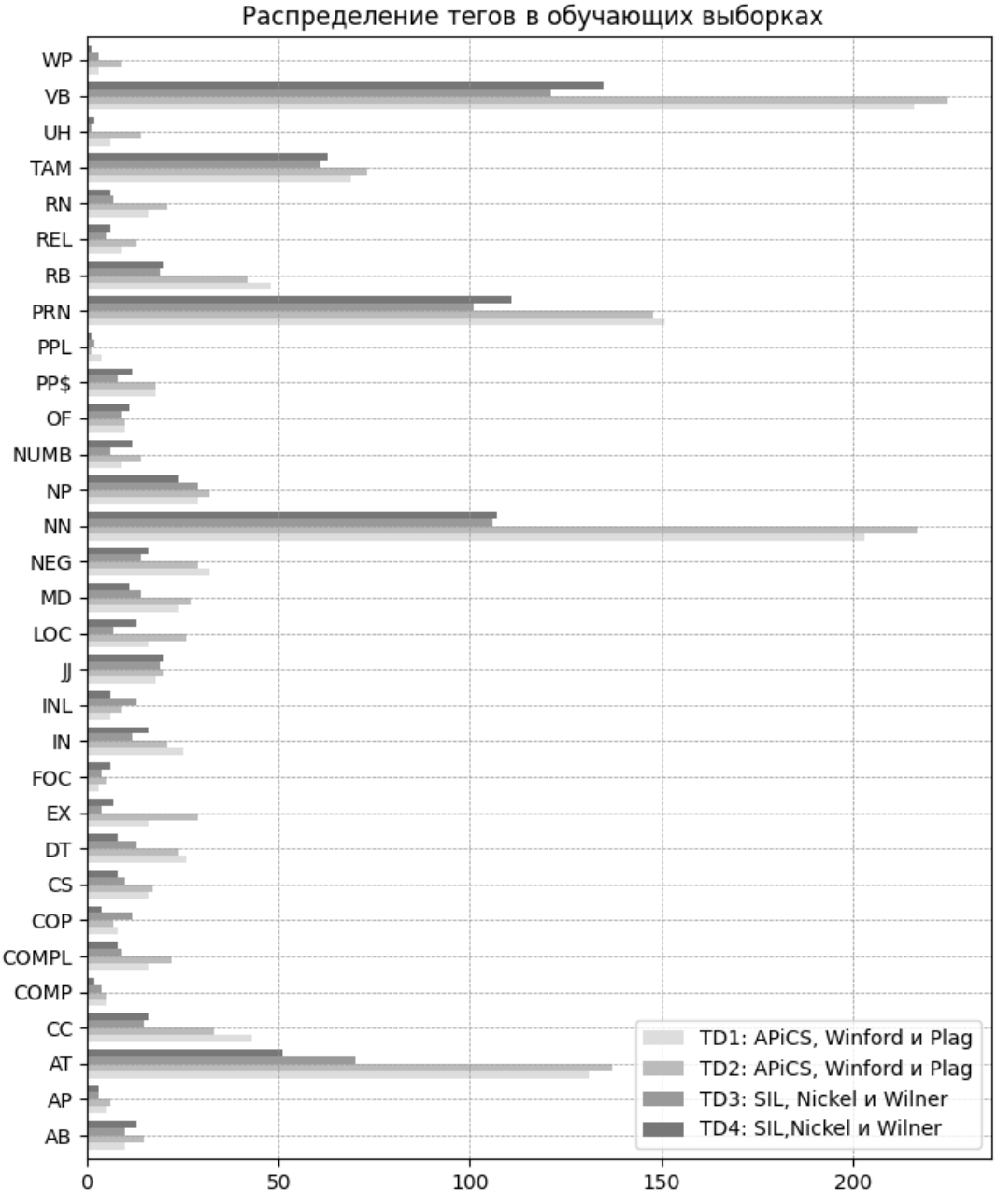


Рисунок 4: Распределение тегов в обучающих выборках

6.4 Тестовая выборка

Тестовая выборка состоит из 70 предложений (613 токенов), извлеченных из примеров в записях словарей Wilner J. [2007] и размеченных вручную. Извлеченные примеры были специально отобраны, чтобы содержать все элементы из набора тегов и словоформы с разными интерпретациями. Например, словоформа “baka” в качестве существительного, глагола, наречия и предлога; “moro” как квантификатор, маркер сравнительной степени и наречие; “taki” как глагол и комплементатор и т. д.

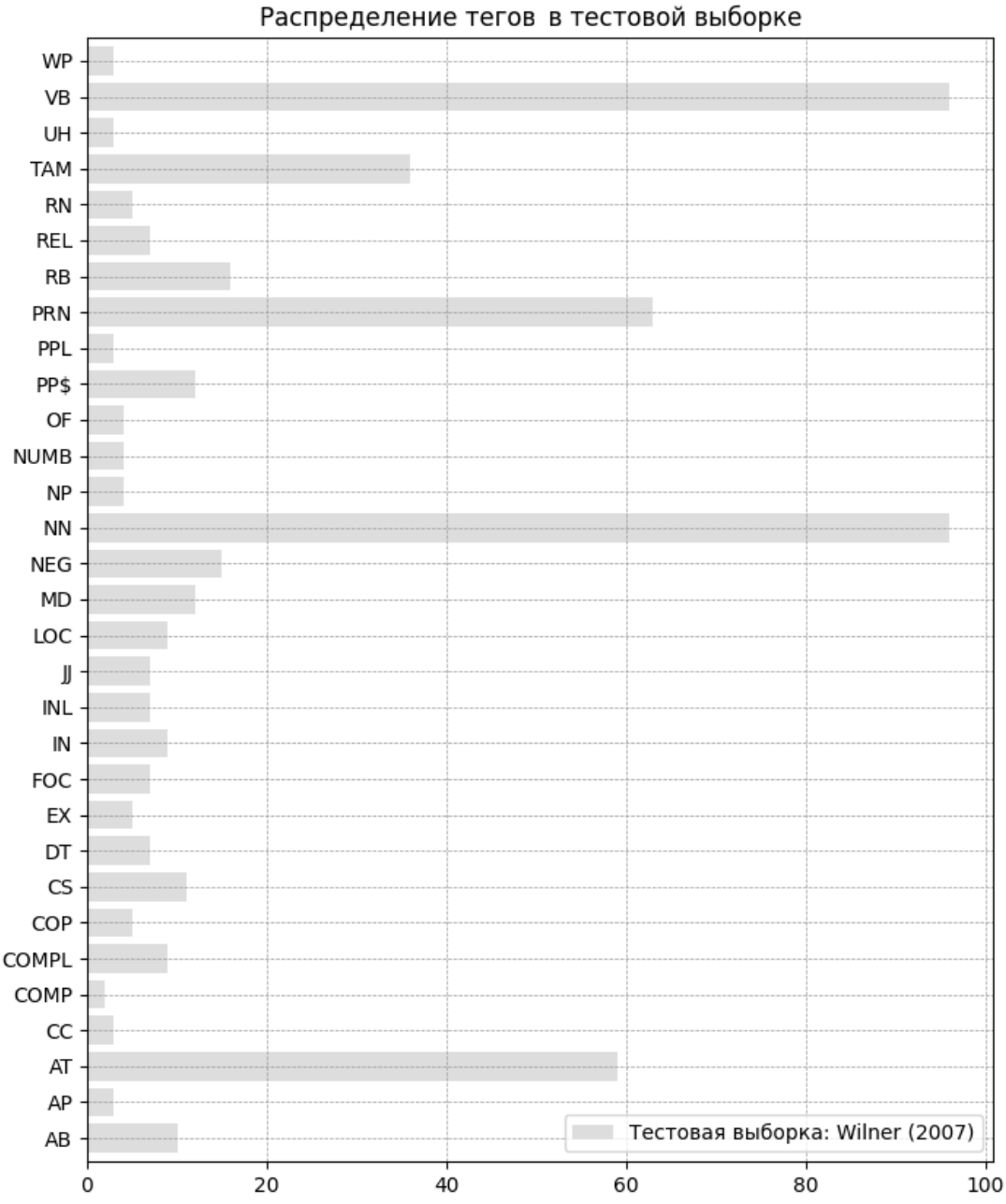


Рисунок 5: Распределение тегов в тестовой выборке

6.5 Обучение и тестирование модели

Обучение модели проходит в четыре этапа. Во время первого используется только один образец, например TD1. Остальные образцы (TD2, TD3, TD4) будут добавлены один за другим на следующих этапах. Это означает, что на четвертом этапе модель обучается на всех данных. Процесс повторяется еще три раза, каждый раз начиная с другого образца и заканчивая следующим. Это сделано для попытки оценить влияние количества данных на результаты, уменьшая при этом любые смещения, вносимые в порядок добавления данных.

Производительность модели оценивается по данным тестирования после каждого этапа для каждой из четырех предложенных метрик. В эксперименте не учитывались результаты разметки знаков препинания, которые искусственно улучшили бы все показатели.

Результаты экспериментов представлены в следующих таблицах, где:

- x - обучающие данные, инициализированные в столбце (x =) с первой обучающей выборкой и расширенные на последующих этапах (x +). Эти столбцы содержат значения и результаты для каждого из этапов эксперимента;
- (K) - это константа (все теги для одной словоформы получают такую же вероятность), так что предсказания модели зависят исключительно от вероятностей переходов. Это представляет собой базовый уровень, который, как ожидается, будет улучшен;
- (A), (B) и (C) показывают производительность модели при применении метрик для замены вероятностей результатов. Серым цветом выделены максимальные значения каждого этапа обучения модели.

Таблица 23: Кумулятивные образцы DT1 + DT2 + DT3 + DT4

кумулятивные образцы		x = TD1	x + TD2	x + TD3	x + TD4
кумулятивные предложения		165	329	440	550
кумулятивные токены		1381	2853	3680	4513
(K)	точность	0.75	0.79	0.80	0.78
	полнота	0.74	0.78	0.80	0.76
	F-мера	0.72	**0.77**	**0.78**	0.76
(A)	точность	0.79	0.79	0.81	0.80
	полнота	0.72	0.73	0.74	0.72
	F-мера	**0.73**	0.73	0.75	0.73
(B)	точность	0.78	0.80	0.79	0.78
	полнота	0.73	0.76	0.77	0.75
	F-мера	**0.73**	**0.77**	0.76	0.75
(C)	точность	0.71	0.79	0.81	0.81
	полнота	0.74	0.80	0.81	0.81
	F-мера	0.70	**0.77**	**0.78**	**0.79**

Таблица 24: Кумулятивные образцы TD2 + TD3 + TD4 + TD1

кумулятивные образцы		x = TD2	x + TD3	x + TD4	x + TD1
кумулятивные предложения		164	275	385	550
кумулятивные токены		1472	2299	3132	4513
(K)	точность	0.76	0.78	0.76	0.78
	полнота	0.73	0.76	0.74	0.76
	F-мера	**0.72**	0.75	0.74	0.76
(A)	точность	0.80	0.79	0.79	0.80
	полнота	0.69	0.70	0.70	0.72
	F-мера	0.70	0.72	0.72	0.73
(B)	точность	0.78	0.75	0.75	0.78
	полнота	0.72	0.73	0.72	0.75
	F-мера	**0.72**	0.73	0.72	0.75
(C)	точность	0.77	0.78	0.81	0.81
	полнота	0.75	0.78	0.79	0.81
	F-мера	**0.72**	**0.76**	**0.78**	**0.79**

Таблица 25: Кумулятивные образцы TD4 + TD3 + TD1 + TD2

кумулятивные образцы		x = TD3	x + TD4	x + TD1	x + TD2
кумулятивные предложения		111	221	386	550
кумулятивные токены		827	1660	3041	4513
(K)	точность	0.73	0.77	0.78	0.78
	полнота	0.67	0.72	0.77	0.76
	F-мера	0.65	0.71	**0.75**	0.76
(A)	точность	0.76	0.75	0.78	0.80
	полнота	0.65	0.68	0.72	0.72
	F-мера	0.66	0.68	0.73	0.73
(B)	точность	0.75	0.78	0.79	0.78
	полнота	0.69	0.71	0.75	0.75
	F-мера	**0.68**	**0.72**	**0.75**	0.75
(C)	точность	0.70	0.76	0.78	0.81
	полнота	0.70	0.74	0.78	0.81
	F-мера	0.66	0.71	**0.75**	**0.79**

Таблица 26: Кумулятивные образцы TD4 + TD1 + TD2 + TD3

кумулятивные образцы		x = TD4	x + TD1	x + TD2	x + TD3
кумулятивные предложения		110	275	439	550
кумулятивные токены		833	2214	3686	4513
(K)	точность	0.71	0.76	0.80	0.78
	полнота	0.66	0.73	0.77	0.76
	F-мера	0.65	0.72	0.77	0.76
(A)	точность	0.79	0.80	0.78	0.80
	полнота	0.67	0.70	0.71	0.72
	F-мера	**0.69**	0.73	0.73	0.73
(B)	точность	0.78	0.78	0.79	0.78
	полнота	0.68	0.73	0.75	0.75
	F-мера	**0.69**	**0.74**	0.76	0.75
(C)	точность	0.71	0.72	0.79	0.81
	полнота	0.70	0.75	0.80	0.81
	F-мера	0.66	0.71	**0.78**	**0.79**

6.6 Обсуждение

Представленные здесь выводы основаны на наблюдении за значениями F-меры:

- сначала отмечается, что влияние размера обучающих данных на предсказания модели оказалось не так линейно как ожидалось. Если рассматриваются только значения из (K), где метрики не использованы, только в таблице 25, F-мера достигла самого высокого значения на последнем этапе. В других таблицах (23, 24 и 26) этот максимум достигается раньше и следующие этапы уменьшают его.
- если бы модель была обучена только на одном образце (первый этап), наилучшие результаты были бы достигнуты на образце DT1 (F-мера = 0.73, таблица 23). В этом случае подойдет любая метрика (A) или (B). Метрика (C) действует хуже чем базовая степень (K). Наоборот, самый неудачный вариант для обучения модели на одном образце был бы DT3: там метрика (B) получает наивысшие значение (F-мера = 0.68, таблица 25);
- самое высокое значение достигается в четвертом этапе при помощи метрики (C).
- метрика (A) достигает самого высокого значения только в первом этапе в таблицах 23 и 26, в обоих случаях вместе с метриками (A) и (B). На следующих этапах значения метрики (A) всегда среди самых низких.

Для облегчения анализа, в таблице 27 представлены средние значения показателей из предыдущих четырех таблиц.

Таблица 27: Средние значения показателей из таблиц 23, 24, 25 и 26

кумулятивные образцы		1 этап	2 этап	3 этап	4 этап
(K)	точность	0.7375	0.775	0.785	0.78
	полнота	0.7	0.7475	0.77	0.76
	F-мера	0.685	0.7375	0.76	0.76
(A)	точность	0.785	0.7825	0.79	0.8
	полнота	0.6825	0.7025	0.7175	0.72
	F-мера	0.695	0.715	0.7325	0.73
(B)	точность	0.7725	0.7775	0.78	0.78
	полнота	0.705	0.7325	0.7475	0.75
	F-мера	**0.705**	**0.74**	0.7475	0.75
(C)	точность	0.7225	0.7625	0.7975	0.81
	полнота	0.7225	0.7675	0.795	0.81
	F-мера	0.685	0.7375	**0.7725**	**0.79**

Из таблицы 27 видно, что метрика (B) работает лучше с меньшим количеством данных, а метрика (C) наоборот.

В основном, метрики (A) и (B) делают более вероятными теги с более высоким счетом. По мере увеличения объема данных более распространенные теги повышают свою вероятность по сравнению с менее частыми. Этот факт отрицательно влияет на предсказание модели. Метрика (B) работает чуть лучше чем (A), потому что она применяет натуральный логарифм, чтобы уменьшать эффект крайних значений, сохраняя при этом связь между более и менее частыми тегами.

Метрика (C) отдает предпочтение менее частым тегам. По мере увеличения объема данных вес редких тегов имеет тенденцию быть меньше по отношению к общему количеству тегов (см. рисунок 5). Но, чаще всего, редкие теги оказываются более предсказательным. Например, хотя частота существительных в обучающей выборке скорее всего выше чем модальных глаголов, в рамках предложения словоформа “bo” встречается скорее всего в

качестве маркера времени и вида (используемого для выражения нереализованной возможности в прошлом или гипотетической ситуации в будущем) чем существительного “ƅo” «лук» (лук и стрела). Метрика (C) понижает вес тегов с более высокой частотой в обучающих данных, и поэтому кажется, что она работает лучше при увеличении объема данных. Но с обучающими выборками большего размера, чем использованные здесь, можно было ожидать, что метрики из (C) начнут вести себя как метрики из (A) в противоположном направлении, делая существительные и глаголы очень маловероятными.

Заключение

Исследования в области автоматической обработки текстов на языке СТ находятся в начальной стадии. Данная работа призвана стать небольшим вкладом в данном направлении. Малоресурсные языки как СТ нуждаются в корпусах и инструментах для работы с ними. Разработка ресурсов для них также выгодна для расширения горизонтов применения уже проверенных методов, доступных для более исследованных языков. Хотя работа посвящена СТ, описанный здесь метод может быть экстраполирован на языки с аналогичной структурой.

Подход, примененный здесь, состоял в исследовании электронных ресурсов, уже доступных для СТ и их использовании на основе уже существующих методов с целью создания POS-теггера. Работа началась с этапа исследования морфологических и синтаксических особенностей языка и продолжалась сбором данных для компиляции небольшого вручную размеченного исследовательского корпуса.

Непосредственно с самого начала было решено отказаться от реализации POS-теггера, основанного исключительно на правилах из-за отсутствия опыта пользования этим языком. Создание чистого стохастического POS-теггера также не было возможно, потому что для СТ не существует большого корпуса. По той же причине были отброшены методы, основанные на нейросетях.

Представленный POS-теггер является гибридным. Он сочетает в себе подходы: первый, основанный на правилах, и второй – стохастический. Данный теггер основан на широком допущении языка (многофункциональность слов), минимальном лексиконе (содержащим почти только слова закрытого класса) и паре простых правил назначения возможных тегов POS для словоформы. Вероятности присвоенных тегов

оцениваются простым подсчетом тегов, затем устраняется неоднозначность с использованием модели POS 3-грамм, обученной на маленьком наборе предложений.

В проведенном эксперименте показано, как далеко может зайти гибридный POS-теггер с очень ограниченными ресурсами, а именно минимальным лексиконом с 346 словоформами и 550 предложениями для обучения модели. Эксперимент с небольшим объемом данных — это идеальная установка для проверки ограничений текущей модели и обеспечения основы для будущих разработок.

Хотя полученные показатели производительности теггера невысоки, есть еще возможности улучшения достижений POS-теггера без изменения принципа его работы, а путем добавления большего количества обучающих данных. Также использование более широкого лексикона могло бы ограничить влияние предварительно назначенных тегов, облегчая задачу устранения неоднозначности и, следовательно, обеспечить лучшие результаты. Имея это в виду, архитектура теггера позволяет пользователям расширить, модифицировать или составлять свой собственный лексикон и переобучать теггер с помощью собственных обучающих данных. Также у пользователей есть возможность перенастроить теггер, чтобы использовать его с совершенно другим набором тегов. Использованные теги влияют на производительность теггера: если они очень широки, они будут терять предсказательную силу для определения окружающих тегов, но, если они слишком специфичны, они будут приводить к разреженности данных в процессе обучения. Поэтому, используемые теги в рамках этой работы должны рассматриваться подробнее в будущих экспериментах. Дальнейшие эксперименты могут предлагать также лучшие метрики для оценки вероятностей тегов для слова.

В заключении, стоит подчеркнуть, что представленный здесь POS-теггер следует рассматривать не как определенное решение для СТ, а как предварительный инструмент для мотивации создания корпусов и стимулировать создание конкурирующих вариантов.

Список литературы

1. Arends J. Syntactic developments in Sranan. Creolization as a gradual process / Jacques Arends. –Phd. Thesis, Katholieke Universiteit Nijmegen, 1989.
2. Bickerton D. The language bioprogram hypothesis / Derek Bickerton // The Behavioral and Brain Sciences. –1984. –Vol. 7. –P. 173-188.
3. Blanker G. Prisma Woordenboek Sranantongo / Gracia Blanker, Jaap Dubbeldam. (eds.). – Het Spectrum, Utrecht. 2010.
4. Bruyn A. Grammaticalization in creoles: Ordinary and not-so-ordinary cases / Adrienne Bruyn // Studies in language, 2009. –Vol. 33. –P. 312-337.
5. Defares J. Enkele problemen rond de lexicale beschrijving van het Sranan / John Defares // OSO. Tijdschrift voor Surinaamse Taalkunde, Letterkunde en Geschiedenis, 1982. –Vol. 1. –P. 47-52.
6. Donicie A. De creolentaal van Suriname / Antoon Donicie. – Radhakishun Paramaribo, 1956.
7. Haspelmath M. Functions of reduplication / Martin Haspelmath, the APiCS Consortium // The atlas of pidgin and creole language structures / Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath, Magnus Huber (eds.). – Oxford, Oxford University Press, 2013.
8. Haspelmath M. Inclusive/exclusive distinction in independent personal pronouns / Martin Haspelmath, Susanne Maria Michaelis, the APiCS Consortium // The atlas of pidgin and creole language structures / Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath, Magnus Huber, Magnus (eds.). – Oxford University Press, Oxford, 2013.
9. Huber M. Order of possessor and possessum / Huber Magnus, APiCS Consortium // The atlas of pidgin and creole language structures / Susanne

M. Michaelis, Philippe Maurer, Martin Haspelmath, Maguns Huber (eds.). – Oxford University Press, Oxford, 2013.

10. Jurafsky D. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition / Daniel Jurafsky, James H. Martin. – Prentice Hall, New Jersey, 2008. –2nd edition.
11. Lefebvre C. Issues in the Study of Pidgin and Creole Languages / Claire Lefebvre. John Benjamins Publishing Company, Amsterdam; Philadelphia, 2004.
12. Lichtveld U. Creole drum. An Anthology of Creole Literature in Surinam / Ursy M. Lichtveld, Jan Voorhoeve (eds.). – Yale University Press, New Haven; London, 1975. –1st ed.
13. Manning C. Foundations of statistical natural language processing / Christopher Manning, Hinrich Schütze. –Massachusetts Institute of Tecnology, 2003.
14. Migge B. Creole learner varieties in the past and in the present: implications for Creole development / Bettina Migge, Margot van den Berg // Acquisition et interaction en langue étrangère [Online], Aile... Lia, 2009. –Vol. 1.
15. Migge B. On the emergence of new language varieties: the case of the Eastern Maroon Creole in French Guiana / Bettina Migge, Isabelle Léglise // Variation in the Caribbean: From Creole continua to individual agency / Lars Hinrichs, Joseph Farquharson (eds.). – John Benjamins, 2011. –P. 207-229.
16. Moetete. Poëzie + proza / Hugo Pos et al. – 1968. –1st ed.
17. Nickel M. Papers on Sranan Tongo / Marilyn Nickel, John Wilner. –Summer Institute of Linguistics, 1984. URL: https://archive.org/details/rosettaproject_srn_morsyn-1 (Last access 05/04/2021).

18. Plag I. The syntax of some locative expressions in Sranan. Prepositions, postposition or noun? / Ingo Plag // Journal of Pidgin and Creole Languages, 1998. –Vol. 13. –P. 335-353.
19. Plag I. Creoles as interlanguages: Word-formation / Ingo Plag // Journal of Pidgin and Creole Languages, 2009. –Vol. 24. –P. 339-362.
20. Radke H. Niederländisch und Sranantongo in Surinamischer Onlinekommunikation / Henning Radke // Taal en Tongval. – University Press, Amsterdam, 2017. Vol. 69. –P. 113-136.
21. Sebba M. Derivational regularities in a Creole lexicon: the case of Sranan / Mark Sebba // Linguistics An Interdisciplinary Journal of the Language Sciences. –De Gruyter Mouton, 1981.
22. Sebba M. Serial verbs: a boost for a small lexicon / Mark Sebba // OSO. Tijdschrift voor Surinaamse Taalkunde, Letterkunde en Geschiedenis. – Nijmigen, 1984. –Vol. 3. –P. 35-38.
23. Sebba M. Adjectives and copulas in Sranan Tongo. / Mark Sebba // Journal of Pidgin and Creole Languages. –1986. –Vol. 1. –P. 109–121.
24. Seuren P. Tense and aspect in Sranan / Pieter A. M. Seuren // Linguistics. – 1981. – Vol. 19. – P. 1043-1076.
25. van den Berg M. “Mi no sal tron tongo” Early Sranan in court records: 1667-1767 / Margot van den Berg. – Master thesis, University of Nijmegen, 2000.
26. Van der Hilst E. De spelling van het Sranan: hoe en waarom zo / Eddy Van der Hilst. – Paramaribo, 2008.
27. Voorhoeve J. Structureel onderzoek van het Sranan / Jan Voorhoeve // De West-Indische Gids, 1956. – Vol. 37. – P. 189-211.
28. Voorhoeve J. De oorsprong van het Sranan Tongo / Jan Voorhoeve // Forum der Letteren. 1977. –P. 139-149.
29. Wilner J. Wortubuku fu Sranan Tongo. Sranan Tongo-English Dictionary / John Wilner (ed.), Ronald Pinas, Lucien Donk, Hertoch Linger Arnie Lo-

Ning-Hing, Tieneke MacBean, Celita Zebeda-Bendt, Chiquita Pawironadi-Nunez, Dorothy Wong Loi Sing. – SIL International, 2007. –5th ed.

30. Wilner J. Wortubuku fu Sranan Tongo. Sranan Tongo-Nederlands Woordenboek / John Wilner (ed.), Ronald Pinas, Lucien Donk, Hertoch Linger Arnie Lo-Ning-Hing, Tieneke MacBean, Celita Zebeda-Bendt, Chiquita Pawironadi-Nunez, Dorothy Wong Loi Sing. – SIL International, 2007. – 5th ed.
31. Winford D. Tense and Aspect in Sranan and the Creole Prototype / Donald Winford // Language change and language contact in pidgins and creoles / John McWhorter (ed.). – John Benjamins B.V., 2000. – P. 391-440.
32. Winford D. Irrealis in Sranan: mood and modality in a radical creole / Donald Winford // Journal of Pidgin and Creole Languages. – John Benjamins B.V., Amsterdam. – Vol. 15:1. – P. 63-125.
33. Winford D. Atlantic creole syntax / Donald Winford // The Handbook of Pidgin and Creole Studies / Silvia Kouwenberg. John Victor Singler (eds.). – Wiley-Blackwell, 2008. –P. 19-47.
34. Winford D. Sranan structure dataset / Donald Winford, Ingo Plag // Atlas of Pidgin and Creole Language Structures Online / Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, Huber Magnus (eds.). – Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. URL: http://apics-online.info/contributions/2 (last access 21/05/2021)
35. Yakpo K. Transatlantic patterns: The relexification of locative constructions in Sranan / Kofi Yakpo, Adrienne Bruyn // Surviving the Middle Passage: The West Africa-Surinam Sprachbund / Pieter Muysken, Norval Smith (eds.). – De Gruyter Mouton, Berlin, 2015. – P. 135–175.
36. Yakpo K. Creole in transition: Contact with Dutch and typological change in Sranan / Kofi Yakpo // Boundaries and Bridges. – De Gruyter Mouton, 2017.

Приложения

А. Лексический ресурс, использованный для построения лексикона

Б. Параметры настроек для обучения модели

В. Реализация программного кода алгоритма POS теггера

Программный код находится по ссылке:
https://github.com/nicolascortegoso/pos-tagger-sranan-tongo/tree/main/code

Обучающие и тестовые данные находятся по ссылке:
https://github.com/nicolascortegoso/pos-tagger-sranan-tongo/tree/main/corpus

A. Лексический ресурс, использованный для построения лексикона

Важным лексическим ресурсом для современного СТ являются словари под редакцией John Wilner и опубликованных SIL Internacional (2007): “Wortubuku fu Sranan Tongo. Sranan Tongo-English Dictionary”, именуемые в дальнейшем как “СТ-АН” (сранан тонго-нидерландский) и “Wortubuku fu Sranan Tongo. Sranan Tongo-Nederlands Woordenboek”, именуемый в дальнейшем как “СТ-НИ” (сранан тонго-нидерландский). Оба словаря доступны для скачивания по лицензии “Creative Commons” и имеют онлайн-версии на веб-сайте “The languages of Suriname. Publishing the research of SIL in Suriname”.

A1. Словари сранан-английского и сранан-нидерландского

Оба словаря аналогичные по объему и по структуре и включают такое же количество записей: каждая запись, существующая в одном, присутствует также в другом. Записи включают примеры употребления с параллельным переводом на целевой язык (соответственно английский или нидерландский). Орфография словарей соответствует рекомендациям официального написания для СТ. При составлении словарей использовались существующие списки слов, письменные тексты, в особенности те, что были опубликованы SIL в Суринаме, а также другие данные, собранные составителями. Слова, которые вышли из употребления, либо не включаются, либо помечаются как архаичные.

Пользователь может скачать словари на формате PDF или прямо использовать их электронные версии на сайте “The languages of Suriname. Publishing the research of SIL in Suriname” (Языки из Суринама. Публикация исследования SIL в Суринаме). Содержание словарей доступный по лицензии “Creative Commons”.

В рамках этой работы рассматриваются только записи словарей сранан-тонго на целевой язык, а не обратное направление (английский или нидерландский на СТ).

A2. Структура записей словарей сранан тонго - целевой язык

Словари включают три типа записей: основная запись и записи второго и третьего уровня. Основная запись - это полная трактовка главного слова. Записи второго уровня включены в предыдущие и содержат фразеологизмы или словосочетания, в которых встречается главное слово. Записи третьего уровня - это сокращенная запись, которая указывает на основную запись, где есть дополнительная информация. Главное слово выделено жирным шрифтом. Если за записью следует небольшой номер в нижнем индексе, это означает, что существует омоним.

Часть речи задается после слова. Если слово может функционировать более чем одним способом, например как существительное и как глагол. Это обозначается пронумерованными значениями после словоформы. Например часть речи словоформы “ɓaka” может быть существительным, предлогом, наречием и глаголом. Таким образом, если слово принадлежит к двум или более частям речи, для каждой части есть свой перевод и пример употребления с соответствующим переводом. Поскольку словари предназначены для говорящих на английском или нидерландском, то соответственно, используются термины на английском или на нидерландском при указании частей речи слова на СТ. Поэтому, части речи указывают на то, как говорящий на этих языках может понимать или использоваться слово в своем языке. Это приводит к некоторым различиям в спецификации частей речи в двух словарях. Для некоторых записей, где не применяется часть речи, встречаются другие теги. Например теги: “фразеологизм”, “поговорка”, “выражения” для словосочетаний и “префикс”, “суффикс” для частиц.

Большинство основных записей включают предложения в качестве примера использования слова в различных смыслах и его соответствующий переводов на целевой язык. Семантически связанные слова как антонимы и синонимы обозначаются ссылками. В некоторых случаях выделяется этимология слова или научная номенклатура для животных или растений.

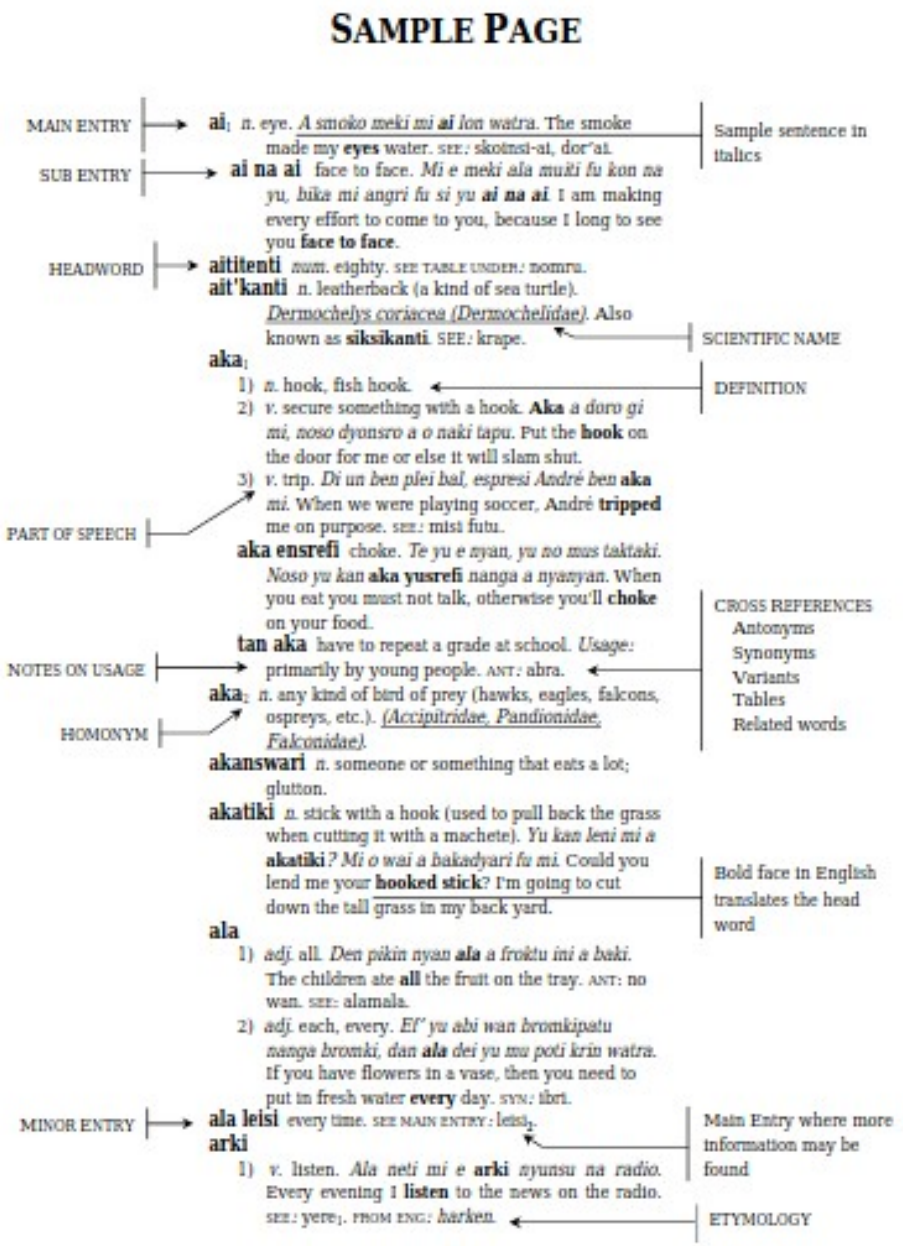


Иллюстрация 1: Шаблон словаря CT-AH (Wilner J., 2007:9)

А3. Электронные версии словарей

На сайте “The languages of Suriname. Publishing the research of SIL in Suriname”, под разделам “So you want to learn Sranan?” (на английском: “Хочешь изучать сранан тонго”) или “Wil je Sranan lernen?” (на нидерландском: “Итак, хочешь ли ты изучать сранан-тонго?”) находятся интерактивные электронные словари. Интерфейс пользователя простой: в левой части страницы размещен словарь сранан-тонго целевых языков (английского или нидерландского) а в правдой словарь целевого языка-сранан тонго. Между ними, в центре, располагается набор 50 электронных текстов с народными рассказами. Слова текстов связаны с словарём в левой части: нажав в тексте на незнакомое слово можно сразу показать его значение в словаре.

Электронные словари также эквивалентны по объему и содержанию по отношению к самим себе и их соответствующим PDF-версиям. В некоторых случаях набор тегов в электронных словарях является более специфичным, чем у PDF-версий. Например, для одного из значений слова “taki”, PDF CT-AH указывает “союз” (conj.), тогда как онлайн-словарь выделяет “конъюнктура” (compl.):

- taki **conj.** that. Yu denki taki yu o man du en? Do you think that you can do it?
- taki **compl.** that. Yu denki taki yu o man du en? Do you think that you can do it?

Однако, в онлайн словарях есть некоторый несоответствия и опечатки в названии частей речи, которые придется обрабатывать. Форматом представления электронных словарей является XML, поэтому довольно проще извлечь их содержания при помощи парсера XML для дальнейшей обработки и анализа данных, чем работать прямо над PDF-версиями. Тем не

менее, словари в формате PDF будут полезны для уточнения или проверки ошибок, которые могут появиться после автоматического извлечения записей.

Ссылка на словари (последний доступ 21.05.2021):
http://www.suriname-languages.sil.org/Sranan/Sranan.html

Б. Настройка параметров для обучения модели

Полезность набора тегов зависит от того, сколько информации требуется для выполнения конкретной задачи. Поскольку POS-теггер CT предназначен для исследовательских целей, пользователь должен иметь возможность применять свои категории для определения частей речи и соответствующих тегов, модифицировать/расширять лексикон или обучать модель своими аннотированными данными.

Б1. О формате файлов

Пользователь производит изменения в файлах в формате данных «csv» (comma-separated values, значения, разделённые запятыми). Формат «csv» подходит для обработки в любом приложении для работы с электронными таблицами, например LibreOffice Calc.

Файлы «csv» создаются вручную, но компилируются при помощи скриптов написаных на Python3. Скрипты не только превращают данные в нужный формат, но и заботятся о целостности данных. После введения изменений в файлах «csv», придется производить компиляцию заново.

Таблица 28: Этапы для определения параметров и обучения модели

этап	процедура	скрипт	вход	выход
1	компилация набора тегов	compile_tagset.py	tagset.csv	tagset.json
2	компилация лексикона	compile_lexicon.py	tagset.json	lexicon.json composed_tokens.txt
3	обучение модели	train_model.py	tagset.json	transition_probabiities.json tags_frequency.json

Б2. Определение набора тегов

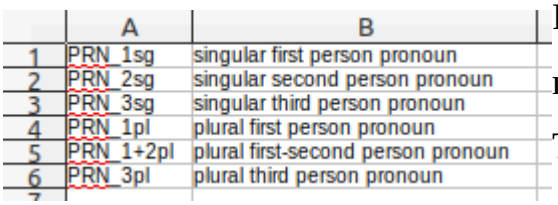
В POS-теггере части речи задаются тегом, поэтому, первый шаг состоит в определении набора тегов в файле “tagset.csv”. Файл должен соответствовать следующим спецификациям:

- каждая строка файла содержит один тег и соответствующее ему описание;
- тег и его описание разделяются запятой, без пробела между ними.

Например:

```
PRN_1sg,singular first person pronoun
PRN_2sg,singular second person pronoun
PRN_3sg,singular third person pronoun
PRN_1pl,plural first person pronoun
PRN_1+2pl,plural first-second person pronoun
PRN_3pl,plural third person pronoun
```

Файл может быть создан с помощью приложения для работы с электронными таблицами.



	A	B
1	PRN_1sg	singular first person pronoun
2	PRN_2sg	singular second person pronoun
3	PRN_3sg	singular third person pronoun
4	PRN_1pl	plural first person pronoun
5	PRN_1+2pl	plural first-second person pronoun
6	PRN_3pl	plural third person pronoun

Иллюстрация 2: Пример таблицы при обработке с помощью приложения

Размер таблицы должен быть *nx2* (*n* строк и *2* столбца), где *n* равно количество тегов. Теги перечисляются в строках: в первом столбце задается тег а, в во втором описание тега.

При сохранении файла необходимо выбрать формат «csv» и удостовериться, что в параметре «разделитель полей» (field delimiter) указано запятая «,» а в параметре «разделитель строк» (string delimiter) - ничего не указано.

Номенклатура тега может содержать любую комбинацию символов. Символ «_» используется специально, чтобы отделить в номенклатуре основной и дополнительный элементы тега.

Основный элемент тега используется для вычисления вероятностей 3-грамм. Символ «_» вводит дополнительную информацию после основного элемента тега. Например, в случае категории «личное местоимение» информация о лице и числе данного местоимения может быть установлена отдельно. Таким образом, при вычислении вероятностей 3-грамм всех нижеупомянутых тегов в таблице 29, используется общая форма “PRN”. Необходимо ввести только основной элемент тега, а дополнительный элемент введенный символом “_” может отсутствовать, как, например, у тегов в таблице 32. Это позволяет избежать проблем разреженных данных в маленьком корпусе и сократить количество тегов при вычислении вероятностей 3-грамм.

Таблица 29: Местоимения

<table>
<tr><td></td><td>ед. число</td><td colspan="2">мн. число</td></tr>
<tr><td>1-ое лицо</td><td>mi
PRN_1sg</td><td>wi
PRN_1pl</td><td rowspan="2">unu
PRN_1+2pl</td></tr>
<tr><td>2-ое лицо</td><td>yu
PRN_2sg</td><td></td></tr>
<tr><td>3-ое лицо</td><td>a, en
PRN_3sg</td><td colspan="2">den
PRN_3pl</td></tr>
</table>

Таблица 30: Притяжательные местоимения

	ед. число	мн. число	
1-ое лицо	mi PP$_1sg	wi PP$_1pl	unu PP$_1+2pl
2-ое лицо	yu PP$_2sg		
3-ое лицо	en PP$_3sg	den PP$_3pl	

Таблица 31: Артикли

	ед. число	мн. число
определенная	a AT_sg	den AT_pl
неопределенная	wan AT_ind	∅

Таблица 32: Связка

экзистенциальная	de EX
зкквативная	a, na COP

Содержание из таблицы «tagset.csv» конвертируется в файл в формате «json» при помощи скрипта «compiler_tagset.py». Данные в файле «tagset.json» необходимы для этапов компилирования лексикона и обучения модели.

Б3. Компилация лексикона

Лексикон языка также сохраняется в формате «csv». Файл «csv» должен соответствовать следующим спецификациям:

- каждая из строк содержит один лексический элемент и ему соответствующий тег;
- лексический элемент и его тег разделяются запятой без пробела;
- если лексический элемент имеет более чем один тег, то тогда создается новая строка для этого лексического элемента с каждым из тегов;
- лексический элемент может состоять из более чем одного слова (токена). Например, отрицательная связка “Sranan Tongo”.

Здесь показан пример создания файла «csv» по спецификациям описанным в предыдущего разделе и информации из таблиц 29, 30, 31 и 32. Содержание файла «example.csv»:

```
a,AT_sg
a,COP
a,PRN_3sg
de,EX
den,AT_pl
den,PRN_3pl
den,PP$_3pl
en,PRN_3sg
en,PP$_3sg
na,COP
mi,PRN_1sg
mi,PP$_1sg
unu,PRN_1+2pl
unu,PP$_1+2pl
wan,AT_ind
wi,PRN_1pl
wi,PP$_1pl
yu,PRN_2sg
yu,PP$_2sg
```

Такое же содержание из файла “example.csv” можно передать в разных «csv» файлах. Это не влияет на процесс составления словаря и позволяет сохранять данные лексикона более организованными образом. Например:

Таблица 33: Данные сохранены в многих файлах

«pers_pronouns.csv»	«poss_pronouns.csv»	«articles.csv»	«copulas.csv»
mi,PRN_1sg	mi,PP$_1sg	a,AT_sg	a,COP
yu,PRN_2sg	yu,PP$_2sg	den,AT_pl	na,COP
a,PRN_3sg	en,PP$_3sg	wan,AT_ind	de,EX
en,PRN_3sg	wi,PP$_1pl		
wi,PRN_1pl	unu,PP$_1+2pl		
unu,PRN_1+2pl	den,PP$_3pl		
den,PRN_3pl			

Так как при наборе тегов, файлы «csv» могут быть созданы с помощью приложении для работы с электронными таблицами. Размер таблицы должен быть *nx2* (*n* строк и *2* столбца), где *n* равно количеству лексических элементов. Например:

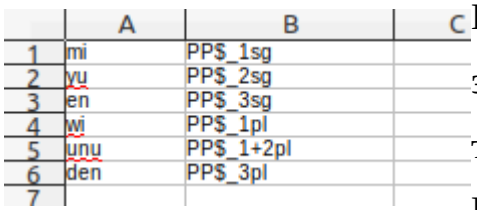


Иллюстрация 3: Пример таблицы при обработке с помощью приложений

Первый столбец должен содержать лексические элементы, а второй – ему соответствующие теги.

При сохранении файла необходимо учитывать ту же процедуру, что и в предыдущем разделе.

Скрипт “compile_lexicon.py” принимает на вход все «csv» файлы из папки “lexicon”. В итоге создаются два файла:

1. скомпилированный лексикон;
2. список лексических элементов, состоящих из более чем одного токена.

Скомпилированный лексикон состоит из ассоциативного массива, где:

- ключом является каждая уникальная словоформа;
- значением является список всех тегов, которые данная словоформа может принимать.

Омонимы собираются вокруг одного и того же ключа. Например, словоформа “a” из примеров принадлежит к классу слов «артикль», «личное местоимение» и «связка». Следовательно, ключ “a” содержит список всех вышеупомянутых частей речи:

```
"a": ["AT_sg", "COP", "PN_3sg"]
```

Список лексических элементов, состоящих из более чем одного токена принимается после процесса токенизации для объединения таких токенов, которые составляют лексический элемент в данном списке. Например, последовательность токенов “no” (нет) и “wan” (один), должна превратиться в одну “no wan” (никто), поскольку, следуя грамматике CT, отрицание может предшествовать только глаголам.

Скрипт также проверяет наличие недопустимых тегов. Он будет компилироваться только в том случае, если теги из «csv» файлов лексикона также существуют в файле «tagset.json».

Б4. Обучение модели

Аннотированные данные принимаются в формате xml.

Его нужно разделить на предложения при помощи меток <sentence> </sentence>.

Токены приложения передаются между метками <token> </tokens>.

Каждый токен должен содержать следующие атрибуты:

- word = словоформа;
- type = сокращенная форма тега;

- tag = полная форма тега.

  Все остальное содержимое между тегами необязательно.

  Пример шаблона:

```
<sentence id="14">
        <source>Nickel, Marilyn; Wilner, John. Papers on Sranan Tongo.  SIL. Example 9a</source>
        <attribution></attribution>
        <feature>Active clause. Indicative</feature>
        <text>San yu tyari kon gi mi?</text>
        <gloss>What did you bring for me?</gloss>
        <tokens>
                <token id="96" word="San" en="interrogative word" type="WP" tag="WP">San</token>
                <token id="97" word="yu" en="singular second person pronoun" type="PN" tag="PN_2sg">yu</token>
                <token id="98" word="tyari" en="transitive verb" type="VB" tag="VB_tran">tyari</token>
                <token id="99" word="kon" en="verb of movement" type="VB" tag="VB_mov">kon</token>
                <token id="100" word="gi" en="dative preposition" type="DAT" tag="DAT">gi</token>
                <token id="101" word="mi" en="singular first person pronoun" type="PN" tag="PN_1sg">mi</token>
                <token id="102" word="?" en="punctuation sign" type="PNCT" tag="PNCT">?</token>
        </tokens>
</sentence>
```

Для обучения модели алгоритм вычисляет вероятность появления тега t_i при условии того, что перед ним находятся теги t_{i-2}, t_{i-1} для набор тегов, определенных в файле «tagset.json».

Скрипт «train_model.py» в папке «code» принимает как вход все «xml» файлах в папке «corpus». Далее пользователь должен указать которые из них он будет использовать, чтобы обучать модель.

На выход алгоритм создает два файла с разными форматами («csv» и «json»), но с одинаковым содержанием: вероятностями 3-грамм. Файл в формате «json» необходим для передачи в POS-теггер, а аналогичный файл в формате «csv» предназначен для пользователей.

В. Реализация программного кода алгоритма POS теггера

Разбор текста происходит в три этапа:

1. Сегментирование текста на предложения и значимые единицы
2. Присвоение словоформам тегов и их вероятности
3. Определение более вероятной последовательности тегов

Например, приложение: “a boi lobi a umapikin” «мальчик любит девочку» разбирается следующим образом:

Таблица 34 пример разбора предложений

1	[[('A', 'word'), ('boi', 'word'), ('lobi', 'word'), ('a', 'word'), ('umapikin', 'word'), ('.', 'period')]]
2	[('A', [('PRN_3sg', 0.31558865253327417), ('AT_sg', 0.30178422597116733), ('LOC', 0.2088517598930887), ('COP', 0.17377536160246984)]), ('boi', [('NN', 0.6729078087289168), ('UH', 0.3270921912710832)]), ('lobi', [('VB_fact', 0.5037052050616948), ('NN', 0.4962947949383052)]), ('a', [('PRN_3sg', 0.31558865253327417), ('AT_sg', 0.30178422597116733), ('LOC', 0.2088517598930887), ('COP', 0.17377536160246984)]), ('umapikin', [('VB', 0.29488863984469343), ('NN', 0.2905502972188513), ('RB', 0.21890185326752043), ('JJ', 0.19565920966893477)]), ('.', [('PNCT', 1)])]
3	[('A', 'AT_sg'), ('boi', 'NN'), ('lobi', 'VB_fact'), ('a', 'AT_sg'), ('umapikin', 'NN'), ('.', 'PNCT')]

В1. Сегментирование текста на предложения и значимые единицы

До процесса теггирования надо разделить письменный текст на значимые единицы, такие как предложения, слова, знаки препинания и т.д. Такой процесс называется токенизация. Алгоритм токенизации делит строку письменного языка на приложения и их элементы. Реализация алгоритма выполняет три шага:

1. выделение токенов
2. сегментация предложений
3. объединение составных токенов

Токены выделяются на основе регулярных выражений. В основном, токены классифицируются на четыре общих класса: “word” (слово), “acronym” (акроним), “number” (цифр) и знаки препинания. Класс “acronym” включает строки как “U.S.A.” и “T.M.”.

На этом этапе можно выделять также токены с аффиксами. Хотя аффиксы могут помогать при процессе классификации частей речи, пока это функция не реализована в коде.

Текст разделяется на предложения на основе новой строки, точки, восклицательного или вопросительного знаков.

После сегментации текста применяется словарь, чтобы объединить составленные токены внутри цепочки предложений.

Алгоритм токенизации реализован в классе “Tokenizer” из файла “pos_tagger.py”. При инициализации объекта класса передается текстовой файл из составных токенов в качестве аргумента.

Метод “tokenize” объекта класса “Tokenizer” принимает тестовую строку в качестве аргумента и возвращает список списков (один для каждого сегментированного предложения) с кортежами токена и их класса.

B2. Присвоения словоформам тегов и их вероятности

После токенизации алгоритм присвоит возможный набор тегов для каждого токена и их вероятности. Реализация алгоритма выполняется в три шага:

1. фильтрация слов от других токенов
2. определение возможных тегов для одной словоформы
3. вычисление вероятностей для каждого из присвоенных тегов

Токены, которые был классифицирован как “number” или как любые знаки препинания на этапе токенизации, получают сразу их соответствующие теги и вероятность 1. Если токен был классифицирован как “word” или

“acronym”, алгоритм принимает подход на основе лексикона для присвоения тегов.

Сначала слову присваивается предварительный набор тегов открытого класса слов. Потом алгоритм ищет слово в лексиконе. Если словоформа найдена, то предварительный набор тегов заменяется тегами, указанными в лексике. Но если поиск прошел неуспешно, алгоритм проверяет начинается ли слово с заглавной буквы. В случае, если это условие не выполняется, алгоритм принимает предварительный список тегов как самый лучший вариант и возвращает его. Но, если слово начинается с заглавной буквы, алгоритм смотрит на его положение в цепочке предложений. Если слово не находится в начале предложения, тогда подразумевается, что оно является именем собственным и, следовательно, предварительный набор тегов заменяется на набор тегов для имен собственных. Но, если слово оказывается в самом начале предложения, тогда алгоритм не может принять заглавную исключительно как признак имен собственных. В этом случае, алгоритм возвращает предварительный набор тегов плюс набор тегов для имен собственных.

Если для данной словоформы присвоен только один тег, тогда его вероятность равна 1. Иначе, вероятности присвоенных тегов вычисляются используя указанную метрику. Если метрика не указана, тогда все присвоенные теги получают такую же вероятность. Сумма всех вероятностей тегов для одной словоформы всегда равняется 1.

Алгоритм присвоения тегов реализован в классе “Emission” из файла “pos_tagger.py”. При инициализации объекта класса передается следующие аргументы в указанном порядке:

1. файл “tag_frequencies.json” с подсчетом тегов в обучающей выборке
2. файл "lexicon.json", содержащий пары словоформ и их возможные теги
3. список с тегами для предварительного назначения

4. список с тегами собственных имен
5. тег для указания цифр
6. тег для указания знаков препинания

Метов “get_emission_probabilities” объекта класса “Emission” принимает список с токенизированным предложением и метрикой для вычисления вероятностей тегов данной словоформы в качестве аргументов. Метод возвращает список, содержащий набор тегов и их соответствующие вероятности для каждой словоформы в предложении.

ВЗ. Определение более вероятной последовательности тегов

После присвоения тегов словоформам алгоритм принимает их вероятности и вероятности 3-грамм для определения более вероятной последовательности тегов. Код является реализацией Витерби алгоритма.

Входными данными для алгоритма является список со всеми возможными тегами для одной словоформы. POS-теггер разрешает эти неоднозначности путем выбора правильного тега с помощью контекста.

Алгоритм выбирает в качестве наиболее вероятной последовательности тегов ту, которая максимизирует произведение двух терминов: вероятность того, что тег присвоен словоформе и последовательности тегов.

На выходе получается один тег для каждой словоформы.

Алгоритм реализован в классе “Transition” из файла “pos_tagger.py”. При инициализации объекта класса передается файл “transition_probabilities.json” с вероятностями 3-грамм и тегов для указания знаков препинания в качестве аргументов.

Метод “get_sequence” объекта класса “Transition” принимает в качестве аргумента список, содержащий набор тегов и их соответствующие вероятности для каждой словоформы в предложении. Метод возвращает

список с кортежами, содержащие предсказанный тег для каждой словоформы в предложении.