# Автоматическая обработка звучащей речи

ВВОД:  
Фонетические (и шире лингвистические) знания, наряду с такими дисциплинами, как цифровая обработка сигналов и математические основания распознавания образов создают теоретическую базу компьютерных [речевых технологий](http://lomonosov-fund.ru/enc/ru/encyclopedia:0127472:article), значимость которых для современного информационного и компьютеризованного общества трудно переоценить. Главная задача [речевых технологий (РТ)](http://lomonosov-fund.ru/enc/ru/encyclopedia:0127472:article) состоит в том, чтобы обеспечить удобное и естественное для человека общение с компьютерной техникой. Очевидно, что это должно быть голосовое общение на естественном языке, предполагающее создание средств автоматического ввода/вывода звучащей речи в компьютерные устройства.

Основные задачи автоматического анализа текста

* **Особенности автоматической обработки звучащей речи**

Потребность в создании систем автоматической обработки естественного языка возникла постольку, поскольку невозможно обучить всех пользователей программированию. Оптимальной формой диалога человека и компьютера является диалог на естественном языке. А так как естественный язык существует в двух формах – письменной и устной, то и создание систем автоматической обработки естественного языка ведется в двух направлениях: обработка устной речи и обработка письменного текста.

Под обработкой устной речи понимается разработка методов, технологий и конкретных систем, которые обеспечивают общение человека с компьютером на естественном или ограниченно естественном языке. Речевой диалог обладает рядом преимуществ по сравнению с традиционным вводом информации с помощью клавиатуры:

* устное общение не требует специальной предварительной подготовки пользователя;
* диалог освобождает руки и зрение;
* за счет системы распознавания голоса возможна защита от недозволенного доступа к объекту;
* диалоговое взаимодействие дает возможность пользоваться компьютером людям с ограниченными возможностями.

Однако связь с помощью голоса имеет и свои недостатки: подверженность шумовым помехам, невозможность неограниченного ввода данных в компьютер в течение длительного времени.

Системы автоматической обработки устной речи находят практическое применение в информационно-справочных службах, где можно получать информацию из базы данных в режиме диалога (например, в медицине или на транспорте). Кроме того, такие системы необходимы и для организации приема и озвучивания сообщений (например, получение электронной почты по телефону), а также для перевода звучащей речи в привычный текст в электронной форме. Компьютеры могут оказывать помощь и при обучении иностранному языку с помощью автоматических фонетических тренажеров.

История практического применения систем автоматической обработки звучащей речи   
началась еще в XVIII в., когда появились первые механические синтезаторы речи. Их создатели ставили целью воспроизвести процессы произнесения звуков с помощью механического устройства, имитируя строение голосового аппарата человека.

В начале XX века механические устройства сменились электрическими вокодерами. Первое устройство для распознавания речи появилось в 1952 г., оно могло распознавать произнесённые человеком цифры. В 1964 г. на ярмарке компьютерных технологий в Нью-Йорке было представлено устройство *IBM Shoebox*.

Коммерческие программы по распознаванию речи появились в начале 90-х годов. Обычно их используют люди, которые из-за травмы руки не в состоянии набирать большое количество текста. Эти программы (н-р, *Dragon NaturallySpeaking, VoiceNavigator*) переводят голос пользователя в текст, таким образом, разгружая его руки. Надёжность перевода у таких программ не очень высока, но с годами она постепенно улучшается.

Увеличение вычислительных мощностей мобильных устройств позволило и для них создать программы с функцией распознавания речи. Среди таких программ стоит отметить приложение *Microsoft Voice Command*, которое позволяет работать со многими приложениями при помощи голоса. Например, можно включить воспроизведение музыки в плеере или создать новый документ.

Прогресс, однако, не стоит на месте и в последнее время в телефонных интерактивных приложениях все чаще стали использоваться системы автоматического распознавания и синтеза речи. В этом случае общение с голосовым порталом становится более естественным, так как выбор в нем может быть осуществлен не только с помощью тонового набора, но и с помощью голосовых команд. При этом системы распознавания являются независимыми от дикторов, то есть распознают голос любого человека.

На сегодняшний день существует два типа систем распознавания речи – 1) работающие по принципу «клиент-сервер» (*client-server*), 2) «на клиенте» (*client-based*). При использовании клиент-серверной технологии речевая команда вводится на устройстве пользователя и через Интернет передается на удаленный сервер, где обрабатывается и возвращается на устройство в виде команды (*Google Voice, Vlingo*); ввиду большого количества пользователей сервера система распознавания получает большую базу для обучения.

Первый вариант работает на иных математических алгоритмах и встречается редко (*Speereo Software*) – команда вводится на устройстве пользователя и обрабатывается в нем же. Плюс обработки «на клиенте» в мобильности, независимости от наличия связи и работы удаленного оборудования. Так, система, работающая «на клиенте» кажется надежнее, но иногда ограничивается мощностью устройства на стороне пользователя.

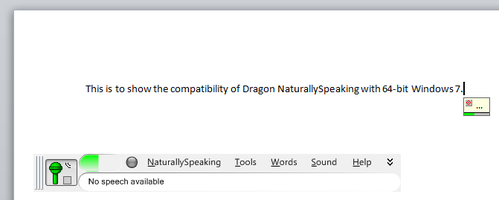
Структура программ распознавания и синтеза звучащей речи

Современные синтезаторы речи включают два блока: блок лингвистической обработки текста, с помощью которого строится полная фонетическая транскрипция синтезируемого текста, а также блок акустического синтеза, который генерирует речевой сигнал.

Блок лингвистической обработки текста имеет достаточно сложную структуру, поскольку создание транскрипции включает несколько этапов: определение языка входного текста, устранение возможных орфографических ошибок, проведение морфологического анализа словоформ для постановки ударения. Самая трудная задача этапа лингвистической подготовки текста – формирование интонации и просодических характеристик фразы. Во многих случаях для этого необходим значительно более сложный семантический и синтаксический анализ фразы. Последний этап работы блока лингвистической подготовки текста – создание фонетической транскрипции. На этом этапе применяются стандартные правила чтения, при этом сложность и трудоемкость этого этапа определяется соотношением между орфографией и произношением каждого конкретного языка.

После создания фонетической транскрипции начинает работу второй блок синтезатора блок акустического синтеза. Его задача – перевод транскрипции в цифровой сигнал, который, в свою очередь, преобразуется в звуковые колебания при помощи обычного цифро-аналогового преобразователя.

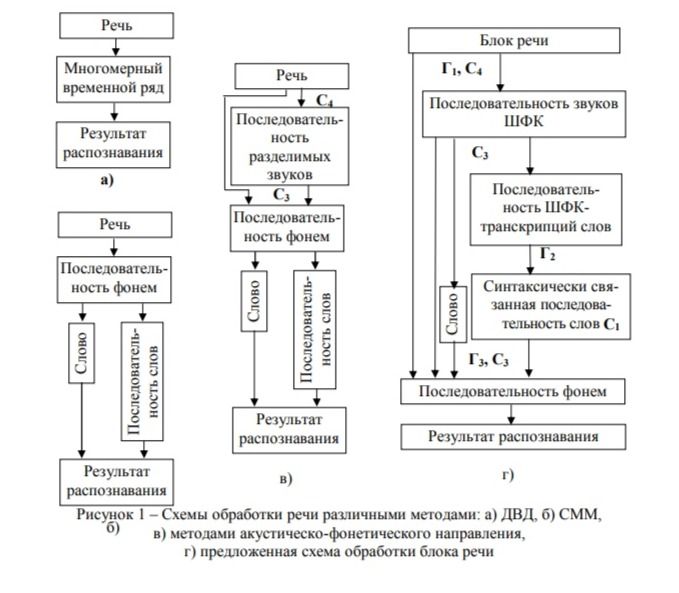
Обзор некоторых программ распознавания и синтеза звучащей речи

***Dragon Naturally Speaking***  – это мировой лидер в программном обеспечении по распознаванию человеческой речи. Программа дает большие возможности при использовании компьютера. Пользователь может диктовать тексты в микрофон, и программа будет писать их сама, например, в текстовом процессоре. 

Программные решения синтеза русской и английской речи, а также программные комплексы распознавания английской речи предлагаются следующими компаниями:   
  


***Sakrament TTS (Text-to-Speech) Engine*** – система нового поколения, осуществляющая качественный речевой синтез. Она может использоваться как отдельное приложение для озвучивания электронных текстов, в качестве речевого движка для других приложений, а также для интеграции с различными информационными системами.

***Sakrament ASR Engine*** – разработка компании «Сакрамент», рассчитана на применение в различных аппаратных системах и программных приложениях, использующих технологии распознавания речи, таких как: *IVR*-системы, мобильные электронные устройства, бытовая техника и т.д. *Sakrament ASR Engine* может быть легко перенесена на любую существующую программную или аппаратную платформу.

Пофонемное   
распознавание выполняется путём анализа множества составных гипотез (последовательностей фонем, последовательностей слов и т.д.), соответствующих произнесению. Новые методы пофонемного распознавания могут повлиять на развитие методов принятия решений на основе анализа составных гипотез. Кроме того, в речи используется многоуровневая система дублирующих друг друга знаков, что обеспечивает значительный запас помехоустойчивости при передаче сообщений между людьми.   
  
Изучение этой многоуровневой системы знаков, способов взаимодействия между знаками различных уровней тесно связано с моделированием интеллектуальной деятельности человека, требует привлечения средств компьютерной лингвистики и других направлений искусственного интеллекта, а успех в этой области будет способствовать развитию средств интеллектуализации компьютерных интерфейсов, ускорению и упрощению процесса человеко-машинного общения.   
В последнее время в области речевых технологий можно отметить использование статистических методов описания речи на акустическом, лексическом и более высоких уровнях. Это оправдано, если в основании лежит адекватная модель объекта. Кроме того, неявное использование свойств объекта (что происходит при его статистическом описании) в практических целях не развивает науку о самом объекте. С этой точки зрения актуальной является разработка методов пофонемного распознавания, использующих явно заданные свойства языка и речи. Объект исследования – речь как средство передачи сообщений. Предмет исследования – модели и методы представления и распознавания речи. Методы исследования. Методы теории информации и кодирования, положения теорий речеобразования и восприятия речи, акустической фонетики, синтаксиса для анализа речи как объекта.  
Целью работы является разработка методов пофонемного распознавания речи, использующих явно заданные свойства языка и речи.  
   
Поставленная цель определила задачи исследования:  
  
1. Выявить важные для распознавания свойства языка и речи.   
  
2. Проанализировать существующие методы распознавания речи, положенные в их основу схемы анализа речи, с точки зрения использования выделенных свойств.   
  
3. Разработать схему анализа речи и методы распознавания речи, использующие явно заданные свойства речи.  


## [Распознавание речи — Технологии Яндекса](https://tech.yandex.ru/speechkit/cloud/doc/guide/concepts/asr-overview-technology-docpage/" \t "_blank)

## Распознавание речи (speech-to-text — stt) — это процесс преобразования речи в текст. SpeechKit Cloud позволяет распознавать спонтанную речь на нескольких языках.

## Языки

* русский
* английский
* украинский
* турецкий

## Языковые модели

SpeechKit решает задачу распознавания в два этапа. На первом этапе в аудиосигнале выделяются наборы звуков, которые могут быть интерпретированы как слова. Для каждого набора звуков обычно существует несколько вариантов слов — то есть несколько гипотез.

На втором этапе подключается языковая модель, которая позволяет проверить каждую гипотезу с точки зрения структуры языка и контекста — насколько данное слово согласуется со словами, распознанными ранее. Система распознавания проверяет гипотезы, пользуясь языковой моделью как словарем. Создание такого словаря — это сложная вычислительная задача, здесь используется машинное обучение нейронных сетей.

Нейронная сеть обучается на речи, которая обычно используется в той или иной области. Поэтому языковые модели специализируются на распознавании речи определенной тематики. Например, для распознавания номера телефона лучше всего подходит модель Числа, а для того чтобы распознать имя и фамилию абонента, следует использовать модель Имена.

*Короткие запросы* (queries) — фразы (3—5 слов) на различные темы, в том числе запросы в поисковых системах (на сайтах).

Например:

* + [покажи следующий поворот]
  + [соединить с отделом продаж]
  + [еще чашку кофе и две мягких французских булочки]
  + [какая погода во владивостоке]
  + [напомни купить овощей и фруктов по дороге домой]

*Адреса*(maps) — адреса, названия организаций и географических объектов.

Например:

* + [поехали на улицу кирпичные выемки пять]
  + [сколько ехать от льва толстого до новой земли]
  + [покажи маршрут до музея маяковского]

*Даты* (dates) — названия месяцев, порядковые и количественные числительные.

Например:

* + [второго ноль седьмого две тысячи первого]
  + [двадцать седьмое апреля тысяча девятьсот девятнадцатого года]

*Имена* (names) — имена и фамилии, просьбы соединить по телефону.

Например:

* + [щукин платон]
  + [соедините с людчиком]
  + [переговорить с васей васиным]

*Числа* (numbers) — количественные числительные от 1 до 999 и разделители — точка, запятая, тире. Модель подходит для диктовки номеров телефонов, счетов, документов.

Например:

* + [два двенадцать восемьдесят пять ноль шесть]
  + [сто пятьдесят семь запятая пятнадцать сорок три]

*Музыка* (music) — названия музыкальных произведений и исполнителей. Модель не предназначена для распознавания музыкальных фрагментов. Подходит только для распознавания названий, имен авторов и исполнителей песен.

Например:

* + [третий концерт рахманинова для фортепиано с оркестром]
  + [алла пугачева любовь похожая на сон]

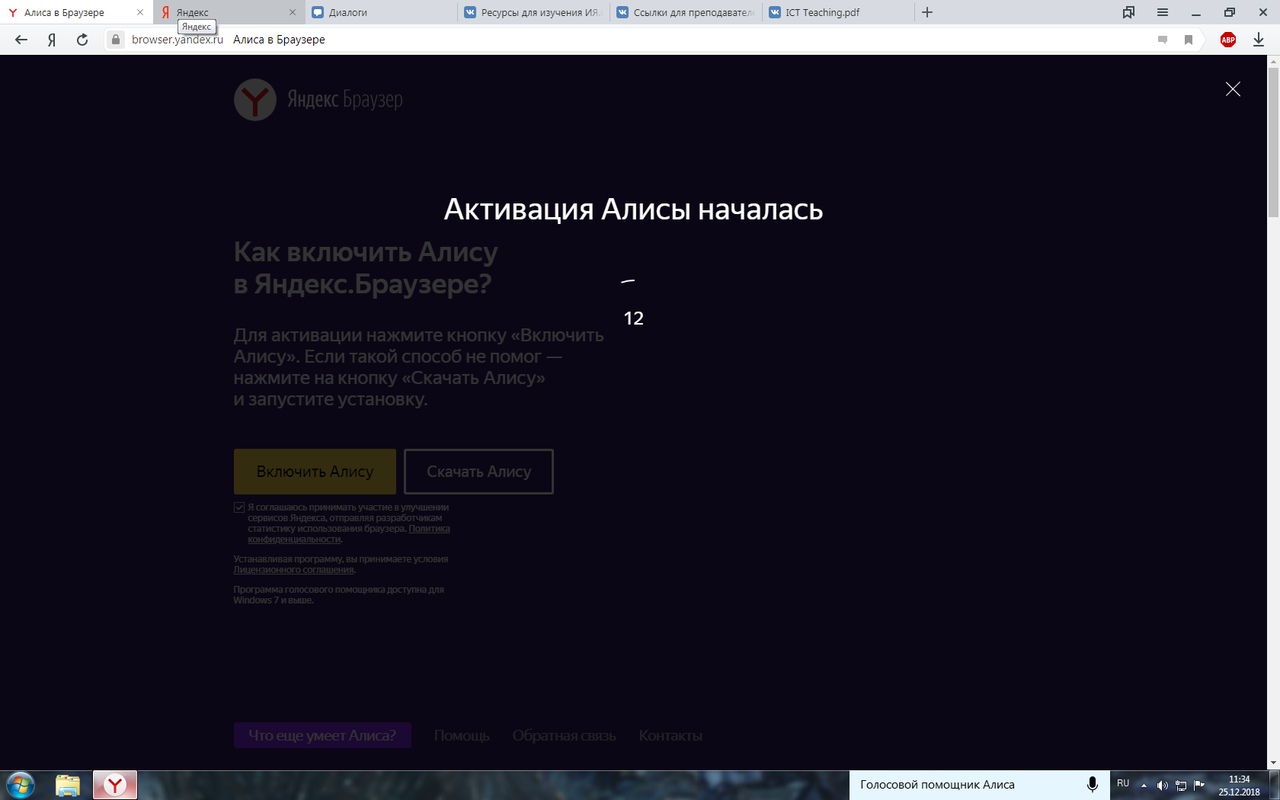
*Заказы* (buying) — фразы, связанные с оформлением заказов в интернет-магазинах (подтверждение заказа и форма доставки).

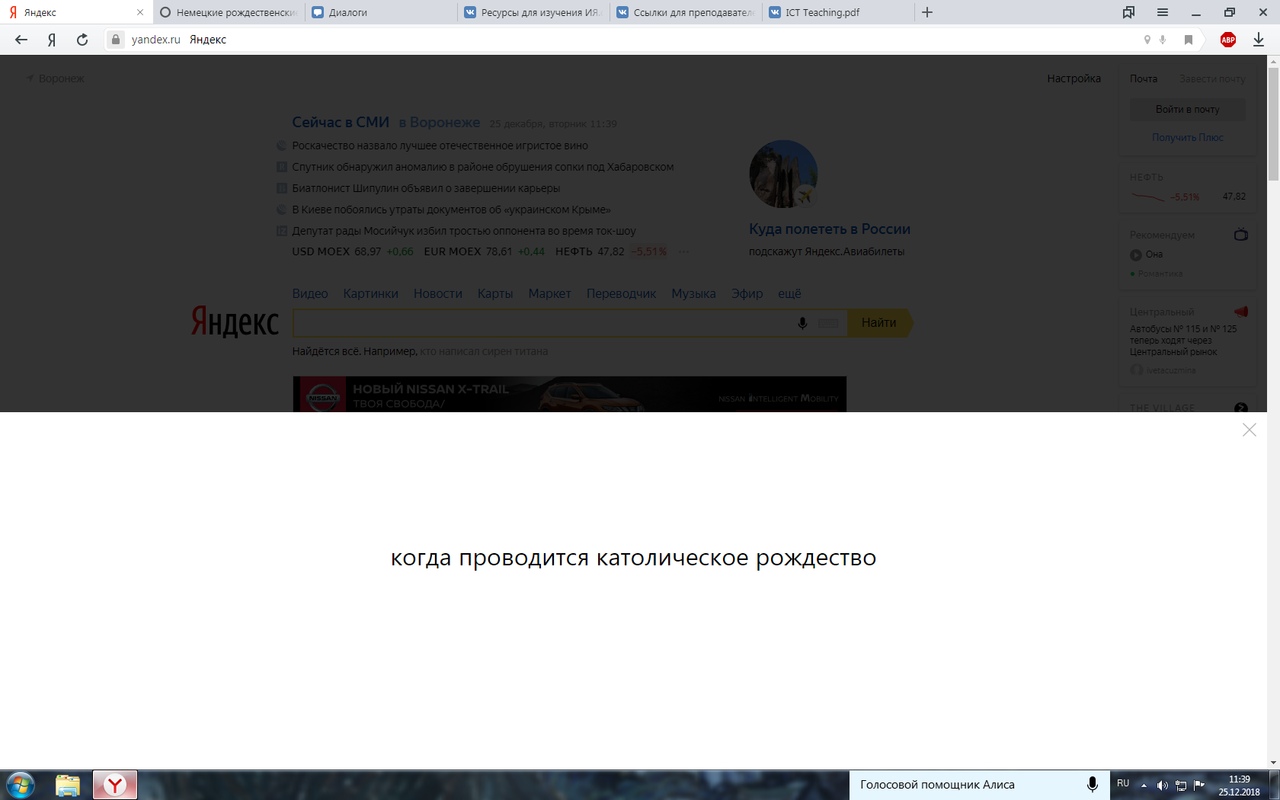
Например:

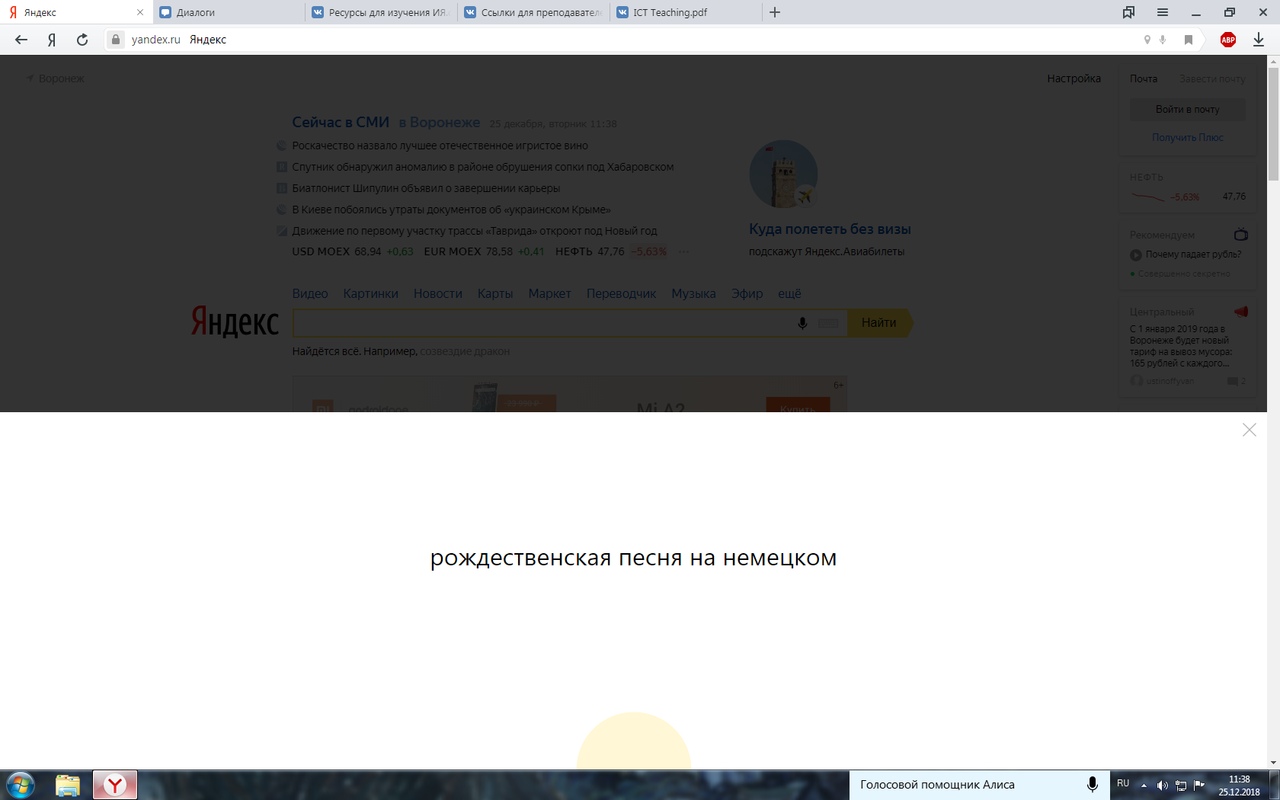
* + [верно хочу купить]

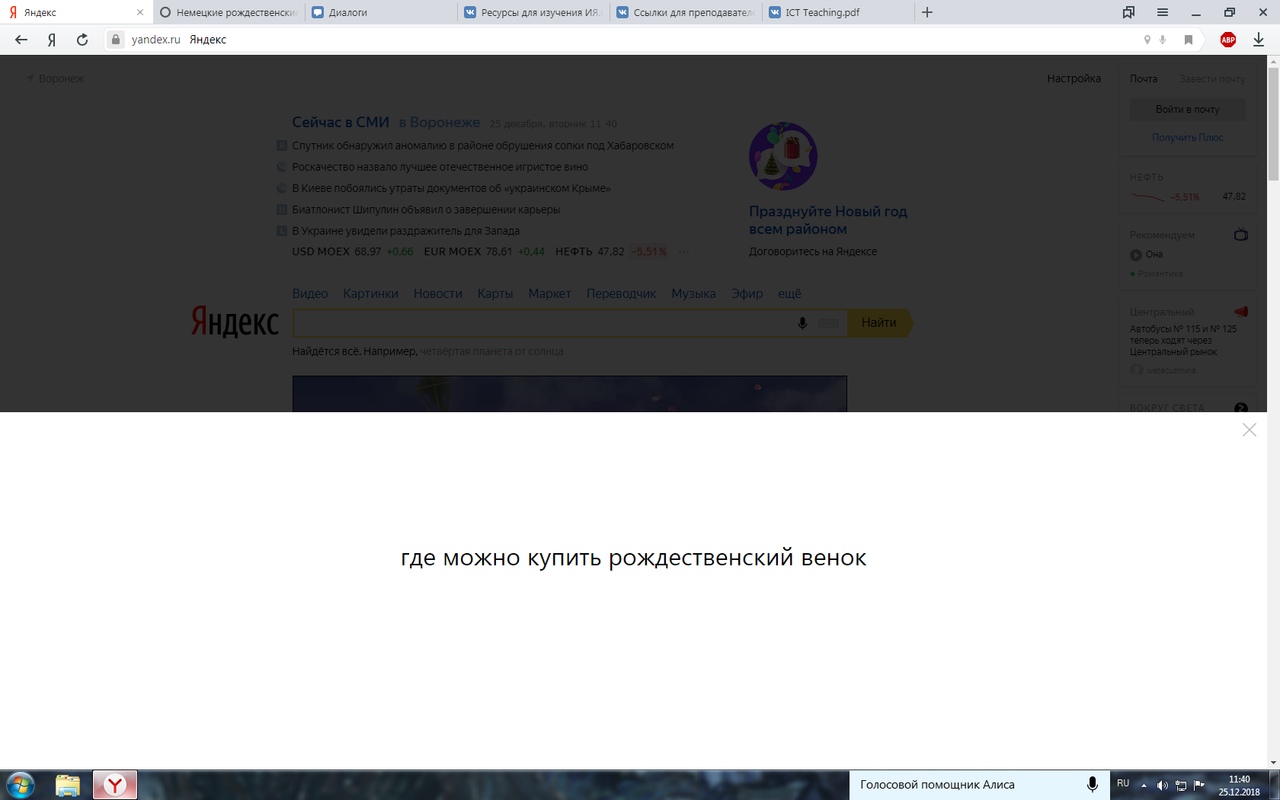
Лабораторная работа:

Используя поисковую систему Яндекс и распознавание речи, подготовьтесь к празднованию Рождества в ВГУ:  
 1) Установите «Алису»:



2) Задавайте вопросы с помощью автоматической обработки звучащей речи, используя разную громкость голоса и разную четкость произношения:   






Полезные ресурсы

* [ACL](http://www.aclweb.org/index.php) (Ассоциация по компьютерной лингвистике): подразделяется на две ветви: Европейскую и Североамериканскую
* Международная русскоязычная конференция по компьютерной лингвистике "[Диалог](http://www.dialog-21.ru/)"
* Лаборатория компьютерной лингвистики Института проблем передачи информации РАН <http://proling.iitp.ru/ru/node/1>
* Журнал «Computational linguistics», выходит он-лайн в открытом доступе <http://www.mitpressjournals.org/loi/coli>
* Langauge-technology world - портал, посвященной лингвистическим технологиям <http://www.lt-world.org/>
* GATES – программное обеспечение для автоматического анализа в открытом доступе <http://gate.ac.uk/>
* Форум "Оценка методов автоматического анализа текста морфологические парсеры русского языка"
* Ресурсы по компьютерной лингвистике в России <http://uisrussia.msu.ru/linguist/_B_comput_ling.jsp>

## Рекомендуемая литература

* Баранов А.Н. Введение в прикладную лингвистику. - М.: Эдиториал УРСС, 2001. - 360 с.
* Бектаев К. Б, Пиотровский Р. Г. Математические методы в языкознании. // Ч. I. Алма-Ата, 1973; ч. II. Алма-Ата, 1974.
* Коваль С. А. Лингвистические проблемы компьютерной морфологии.- СПб.: Изд-во С.-Петерб. ун-та, 2005. - 151 с.
* Марчук Ю.Н. Основы компьютерной лингвистики: Учебное пособие. - М., 1999. - 225 с.
* Новое в зарубежной лингвистике. Вып. XXIV: Компьютерная лингвистика.//М., 1989.
* Поспелов Д.А. Искусственный интеллект. Справочник. Книга 2. Модели и методы
* 1990, 304 с.
* Структурная и прикладная лингвистика. Вып. 1. // Под ред. А. С. Герда. Л, 1978.
* Jurafsky, Daniel, and James H. Martin. 2009.[Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics](http://www.cs.colorado.edu/~martin/slp.html). 2nd edition. Prentice-Hall.
* Manning, Chris, and Schütze, Hinrich Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.