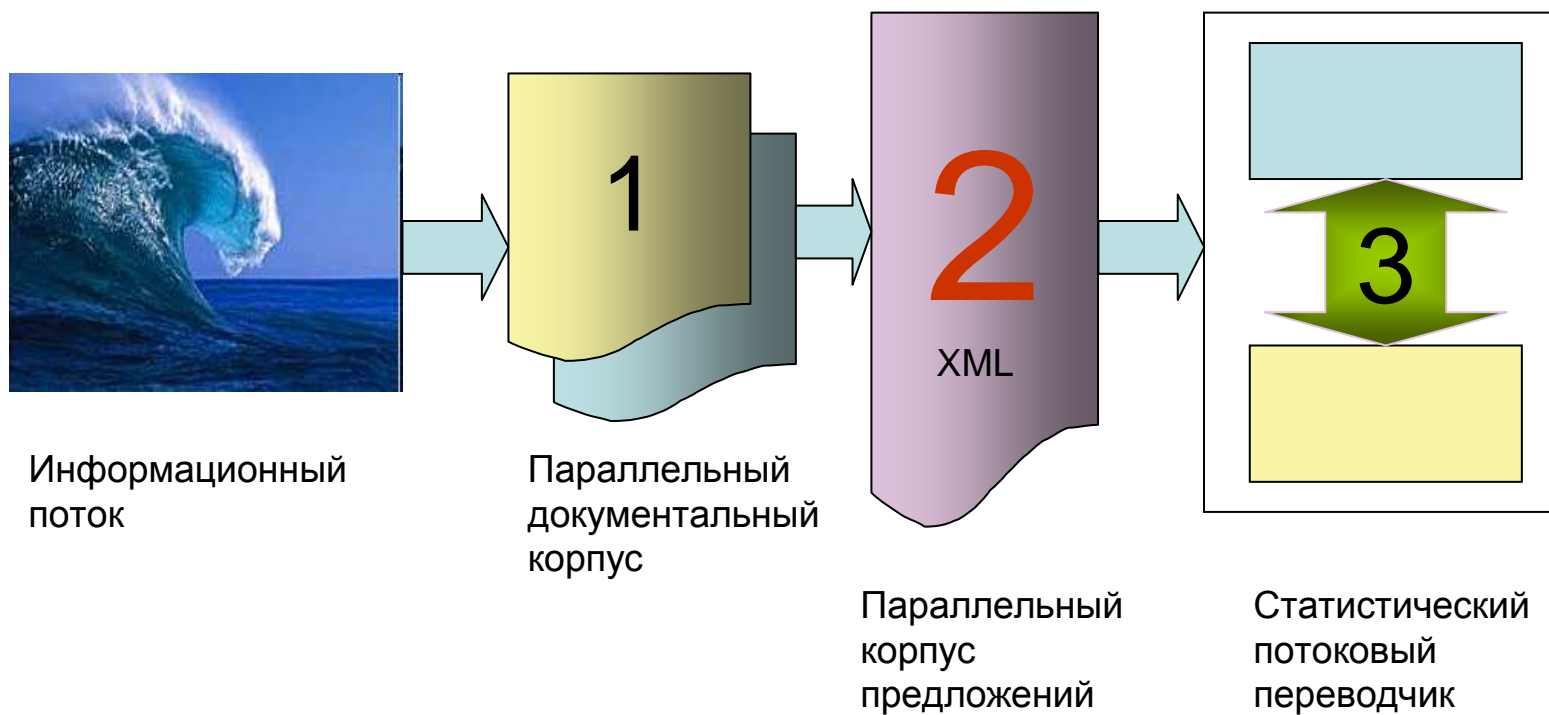


**ВЫРАВНИВАНИЕ УКРАИНСКО-РУССКОГО
КОРПУСА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ ИЗ
СЕТЕВЫХ СМИ**

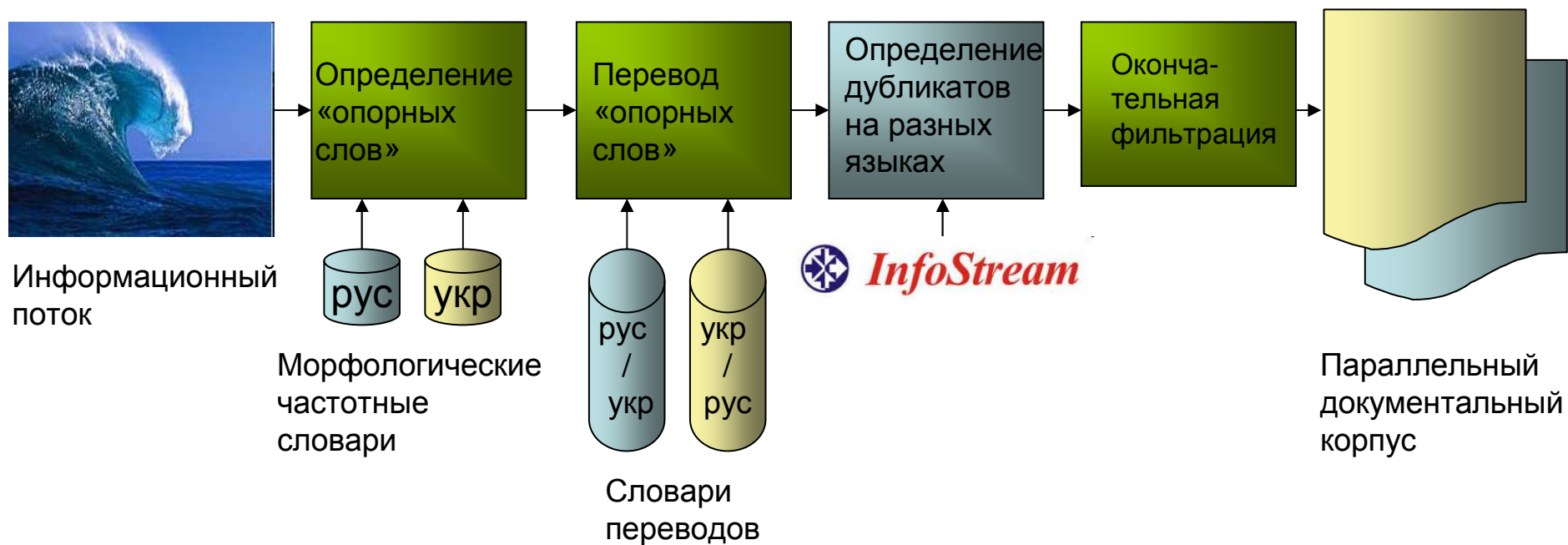
Ландэ Дмитрий Владимирович,
д.т.н., зам. директора ElVisti
Жигало Владлен Викторович,
аспирант, инж.-программист ElVisti

Партенит-2010

Три задачи – три этапа



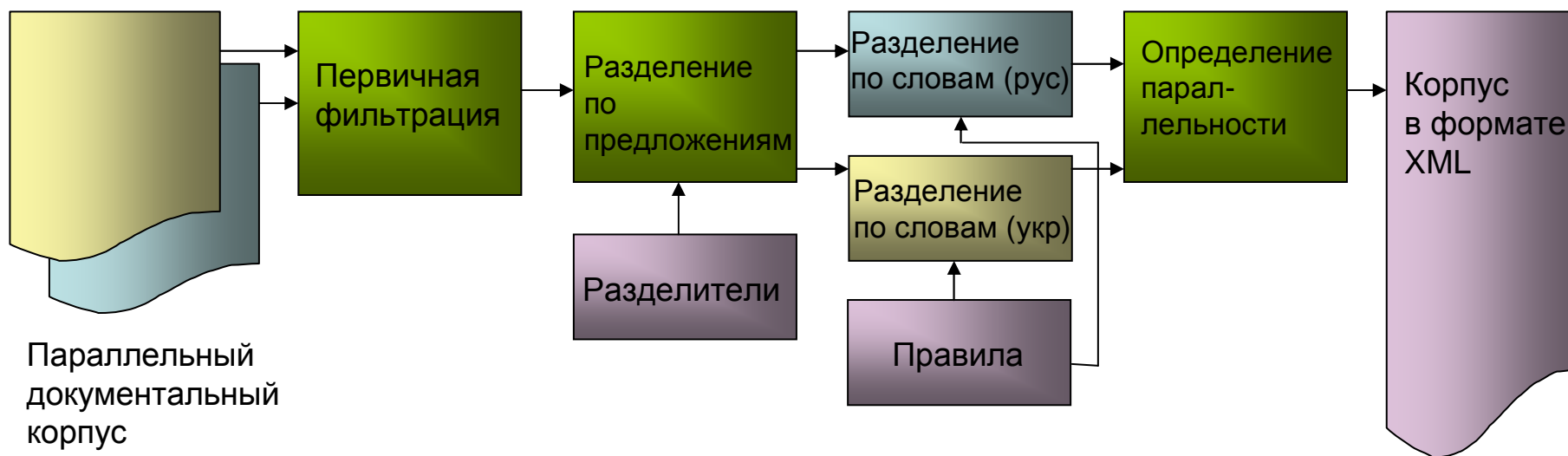
Алгоритм создания параллельного документального корпуса



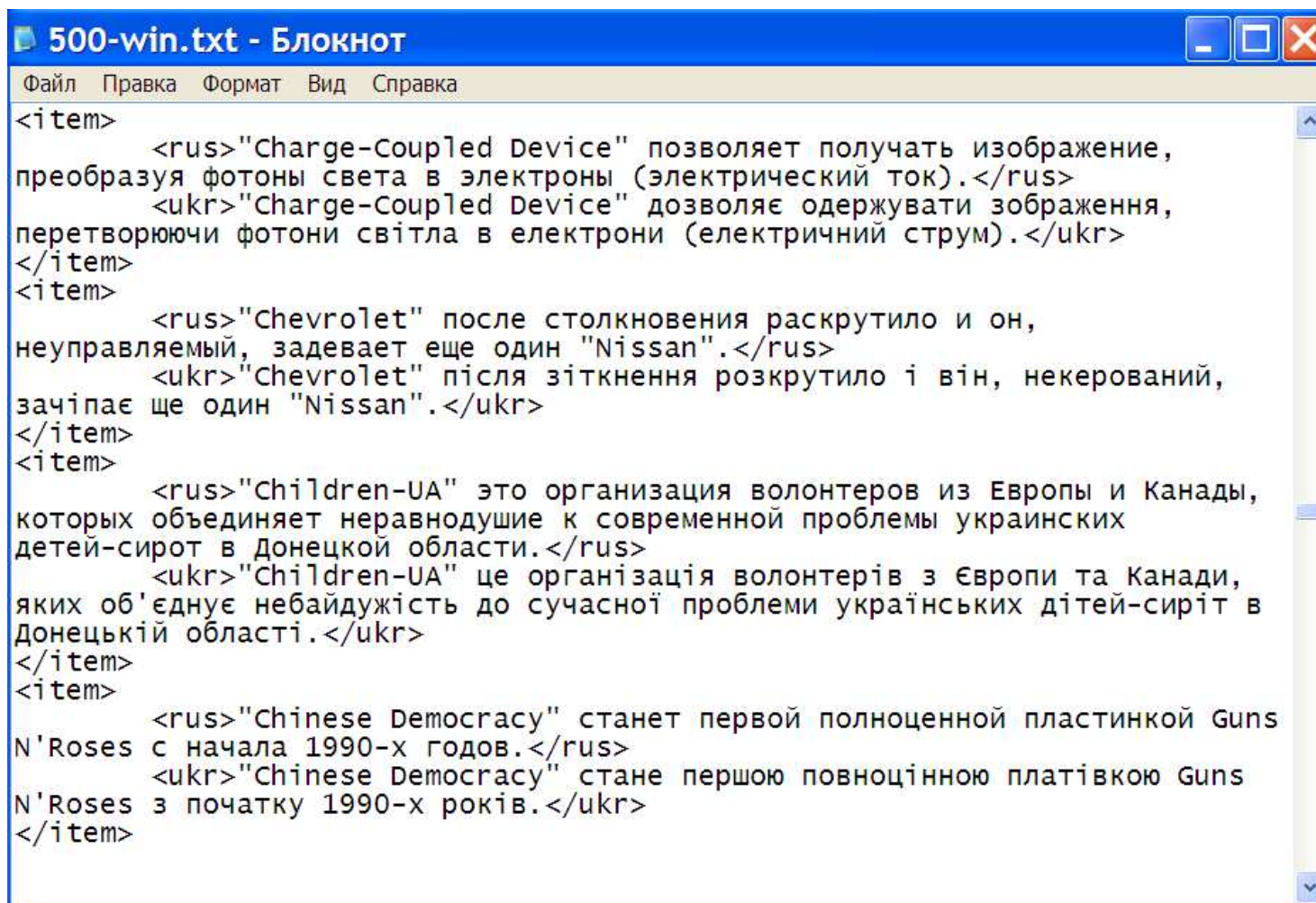
Описание процедуры создания параллельного корпуса предложений

1. Разделение параллельных документов на предложения:
 - 1.1. Определителем конца предложения были взяты символы (. ! ? ;)
 - 1.2. Если в тексте встречалось сокращение или инициалы с точкой то она не считается концом предложения.
2. Подсчет количества предложений в параллельных документах. Если данные документы по количеству предложений одинаковы, то они брались на дальнейшую обработку.
3. Разделение предложения на слова:
 - 3.1. Словом считалось любое сочетание символов отделенное от других групп символов пробелом.
 - 3.2. Накладывались дополнительные ограничения на определение слова на каждом из языков. Например, слова на украинском, в начале которых, упоминались слова: який, яка, що, котрий и т.д. условно считались одним словом.
4. Подсчет количества слов в параллельных предложениях.
В параллельный корпус предложений брались лишь те предложения которые по количеству слов не отличись более чем на одно слово.

Алгоритм создания параллельного корпуса предложений



Фрагмент параллельного корпуса

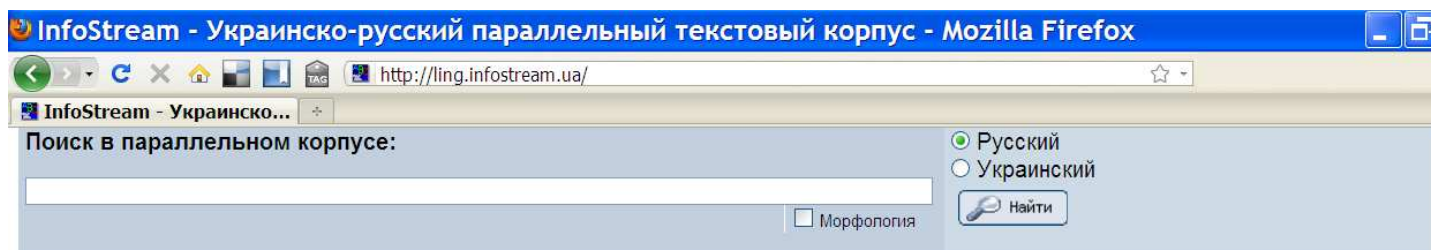


```

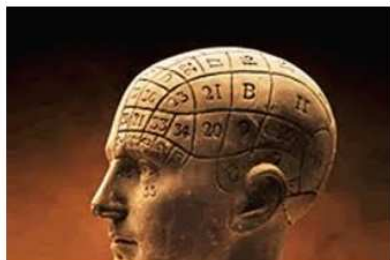
500-win.txt - Блокнот
Файл  Правка  Формат  Вид  Справка
<item>
  <rus>"Charge-Coupled Device" позволяет получать изображение,
  преобразуя фотоны света в электроны (электрический ток).</rus>
  <ukr>"Charge-Coupled Device" дозволяє одержувати зображення,
  перетворюючи фотони світла в електрони (електричний струм).</ukr>
</item>
<item>
  <rus>"Chevrolet" после столкновения раскрутило и он,
  неуправляемый, задевает еще один "Nissan".</rus>
  <ukr>"Chevrolet" після зіткнення розкрутило і він, некерований,
  зачіпає ще один "Nissan".</ukr>
</item>
<item>
  <rus>"Children-UA" это организация волонтеров из Европы и Канады,
  которых объединяет неравнодушие к современной проблемы украинских
  детей-сирот в Донецкой области.</rus>
  <ukr>"Children-UA" це організація волонтерів з Європи та Канади,
  яких об'єднує небайдужість до сучасної проблеми українських дітей-сиріт в
  Донецькій області.</ukr>
</item>
<item>
  <rus>"Chinese Democracy" станет первой полноценной пластинкой Guns
  N'Roses с начала 1990-х годов.</rus>
  <ukr>"Chinese Democracy" стане першою повноцінною платівкою Guns
  N'Roses з початку 1990-х років.</ukr>
</item>

```


Онлайн-интерфейс – сайт <http://ling.infostream.ua>



Украинско-русский параллельный текстовый корпус



В Информационном центре ElVisti создан выровненный на уровне предложений украинско-русский параллельный текстовый корпус из веб-публикаций. Объем корпуса - более 2,6 млн. пар уникальных предложений.

Метод построения корпуса базируется на использовании "опорных слов" в тестовых документах, а также средствах их

автоматического перевода. Опорные слова в рамках данного подхода выделяются с использованием русского и украинского морфологических словарей, а также словарей переводов имен существительных для русского и украинского языков. Кроме того, для вычисления весов терминов в документах используются некоторые дополнительные эмпирико-статистические правила. Для выравнивания параллельного корпуса на уровне предложений использовались преимущественно статистические методы.

Алгоритмы были реализованы в виде программного комплекса, который интегрирован с системой контент-мониторинга InfoStream, благодаря чему корпус постоянно пополняется.

Предполагается дальнейшее использование данного лингвистического ресурса для создания системы автоматического перевода новостных сообщений.

Язык запросов

Запросы состоят из поисковых слов и операторов. В качестве поисковых слов могут использоваться слова естественного языка или их правые усечения. По умолчанию, при отключенной морфологии, каждое слово воспринимается как усечение (слова менее 3 символов ищутся как точное совпадение). Для поиска по полному слову, а не усечению, необходимо дописать к нему специальный символ "J". Система не чувствительна к регистрам букв.

В системе используется следующий набор операторов:

~ - оператор контекстного следования;
@ - оператор контекстной близости;
! , ^ - логическое И-НЕТ;
& , + - логическое И;
| , - логическое ИЛИ;

Оператор контекстного следования (~) отбирает пары поисковых терминов, которые в тексте документа расположены друг за другом, причем учитывается порядок следования терминов.

Оператор контекстной близости (@) отбирает пары поисковых терминов, которые находятся рядом друг с другом, причем порядок следования не важен.

Различные уровни определяются с помощью круглых скобок.

Описание ресурса – сайт <http://ling.infostream.ua>

Для скачивания доступен [заархивированный фрагмент параллельного корпуса](#) размером в **100 тысяч** пар уникальных предложений (в ZIP-архиве ~ 9 МБ).

Формат представления данных приближен к XML:

```
<item>
  <rus>предложение</rus>
  <ukr>речення</ukr>
</item>
... 99 998 раз ;)
<item>
  <rus>предложение</rus>
  <ukr>речення</ukr>
</item>
```

Информация представлена в кодировке CP1251 (Windows).

Использование этого фрагмента корпуса в научных и учебных целях - свободное.

Подробности - в [статье](#) Д.Ландэ и В.Жигало

Препринт: [arXiv:0807.0311](#), [PDF](#)

Режим поиска – сайт <http://ling.infostream.ua>

Поиск в параллельном корпусе:

вертолет

Морфология

Русский
 Украинский

Найти

↓ вертолет

Найдено документов - **3472**, страница 1 из 348

Статистика слов:

↓ **ВЕРТОЛЕТ** - 5651,

1. **Международная аэрокосмическая выставка "Фарнборо-2010" пройдет в Лондоне**

Около 1350 участников из 52 стран примут участие в международной аэрокосмической выставке "Фарнборо-2010", которая пройдет в пригороде Лондона 19-25 июля, сообщает BBC Russia со ссылкой на пресс-службу Федеральной службы по военно-техническому сотрудничеству (ФСВТС) России.

2. **"Си Бриз-2010": украинские десантники одели немцев в тельняшки**

В рамках учений "Си бриз 2010" в Украине немецкие десантники, за плечами которых не один десяток прыжков с парашютом, впервые осуществили прыжки с украинского вертолета Ми-8 под украинскими куполами.

3. **Пэрис Хилтон учится водить вертолет**

Пэрис Хилтон платит за один урок, на котором ее учат водить вертолет, 7 тысяч долларов.

4. **Под Берлином взрываются снаряды**

Борьба с огнем пока не очень успешна: пожарные не могут зайти в лес, чтобы приступить к тушению пожара, так как снаряды продолжают взрываться.

Міжнародна аерокосмічна виставка "Фарнборо-2010" пройде в Лондоні

Близько 1350 учасників з 52 країн візьмуть участь у міжнародній аерокосмічній виставці "Фарнборо-2010", яка пройде в передмісті Лондона 19- 25 липня, повідомляє BBC Russia з посиланням на прес-службу Федеральної служби з військово-технічного співробітництва (ФСВТС) Росії.

"Сі Бриз-2010": українські десантники одягли німців у тільняки

У рамках вчень "Сі бриз 2010" в Україні німецькі десантники, за плечами яких не один десяток стрибків з парашютом, вперше здійснили стрибки з українського вертольота Мі-8 під українськими куполами.

Періс Хілтон навчатється водити гелікоптер

Періс Хілтон платить за один урок, на якому її навчать водити гелікоптер, 7 тисяч доларів.

Під Берліном вибухають снаряди

Боротьба з вогнем поки що не дуже успішна: пожежники не можуть зайти в ліс, щоб приступити до гасіння пожежі, так як снаряди продовжують вибухати.

Фрагмент целевого документа – сайт <http://ling.infostream.ua>

Документ по запросу:  вертолет



Международная аэрокосмическая выставка "Фарнборо-2010" пройдет в Лондоне

Около 1350 участников из 52 стран примут участие в международной аэрокосмической выставке "Фарнборо-2010", которая пройдет в пригороде Лондона 19-25 июля, сообщает BBC Russia со ссылкой на пресс-службу Федеральной службы по военно-техническому сотрудничеству (ФСВТС) России. "В статической экспозиции на открытых площадках будут демонстрироваться более 50 образцов авиационной техники из США, Канады, Великобритании, Италии, России, Украины.

Рядом с пассажирскими и спортивными самолетами будут демонстрироваться боевые самолеты F-18, F-15E, F-16, "Еврофайтер", военно-транспортный самолет C-27J, боевые **вертолеты** итальянской компании "Агуста Вестланд", говорится в сообщении.

 [Закреть](#)

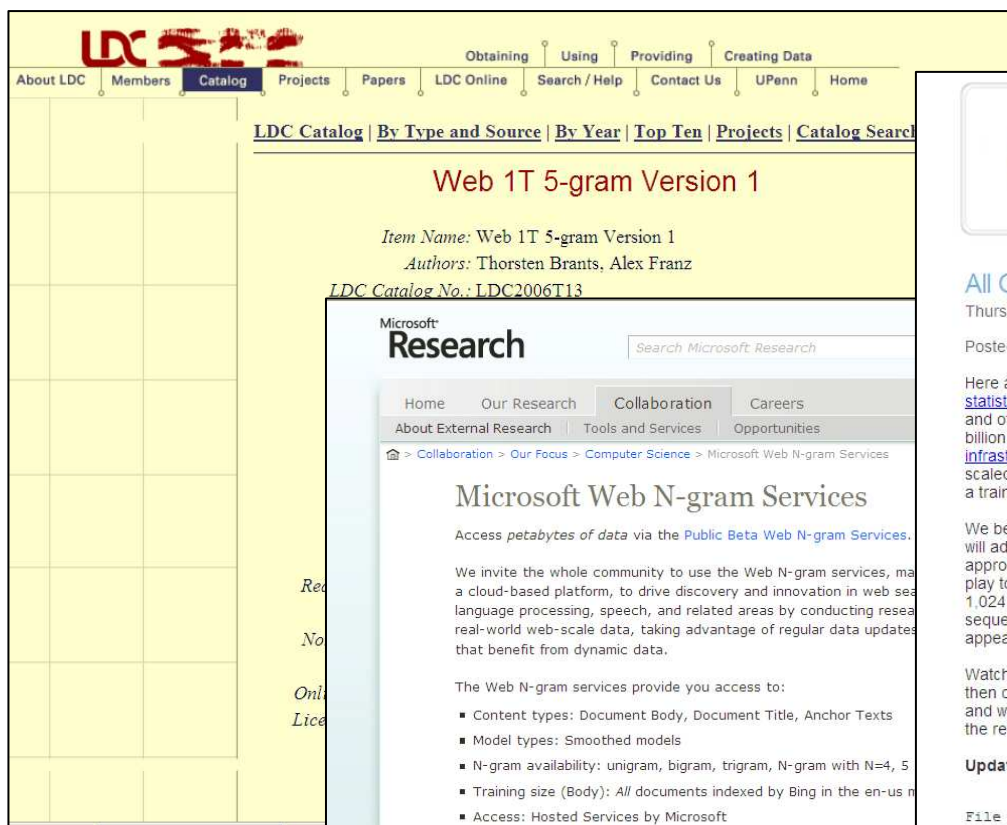
Міжнародна аерокосмічна виставка "Фарнборо-2010" пройде в Лондоні

Близько 1350 учасників з 52 країн візьмуть участь у міжнародній аерокосмічній виставці "Фарнборо-2010", яка пройде в передмісті Лондона 19- 25 липня, повідомляє BBC Russia з посиланням на прес-службу Федеральної служби з військово-технічного співробітництва (ФСВТС) Росії. "У статичній експозиції на відкритих майданчиках демонструватимуться більше 50 зразків авіаційної техніки з США, Канади, Великобританії, Італії, Росії, України ...

Поряд з пасажирськими та спортивними літаками будуть демонструватися бойові літаки F-18, F-15E, F-16, "Єврофайтер", військово- транспортний літак C-27J, бойові вертольоти італійської компанії "Агуста Вестланд", говорить в повідомленні.

 [Наверх](#)

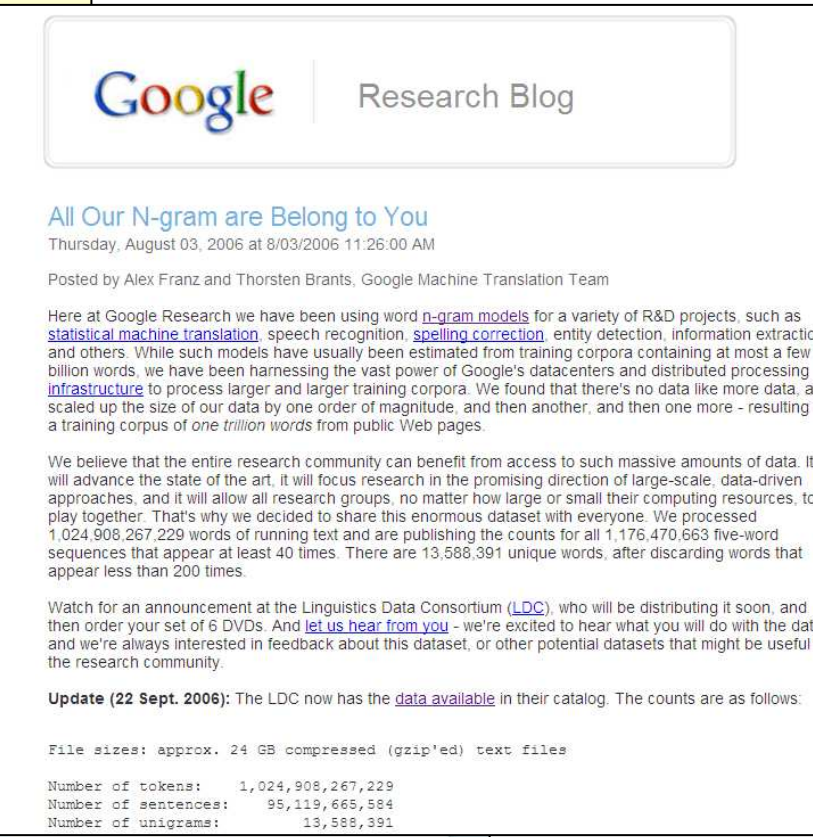
Примеры технологий n-gram



LDC Catalog | [By Type and Source](#) | [By Year](#) | [Top Ten](#) | [Projects](#) | [Catalog Search](#)

Web 1T 5-gram Version 1

Item Name: Web 1T 5-gram Version 1
 Authors: Thorsten Brants, Alex Franz
 LDC Catalog No.: LDC2006T13



Google | Research Blog

All Our N-gram are Belong to You
 Thursday, August 03, 2006 at 8/03/2006 11:26:00 AM

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word [n-gram models](#) for a variety of R&D projects, such as [statistical machine translation](#), speech recognition, [spelling correction](#), entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing [infrastructure](#) to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of *one trillion words* from public Web pages.

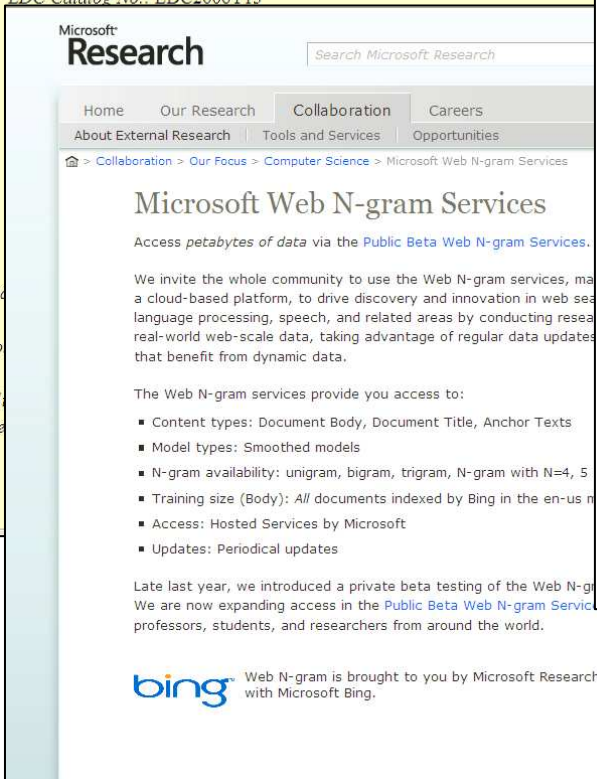
We believe that the entire research community can benefit from access to such massive amounts of data. It will advance the state of the art, it will focus research in the promising direction of large-scale, data-driven approaches, and it will allow all research groups, no matter how large or small their computing resources, to play together. That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Watch for an announcement at the Linguistics Data Consortium ([LDC](#)), who will be distributing it soon, and then order your set of 6 DVDs. And [let us hear from you](#) - we're excited to hear what you will do with the data, and we're always interested in feedback about this dataset, or other potential datasets that might be useful for the research community.

Update (22 Sept. 2006): The LDC now has the [data available](#) in their catalog. The counts are as follows:

File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,668,584
Number of unigrams:	13,588,391



Microsoft
Research

Home | Our Research | Collaboration | Careers
 About External Research | Tools and Services | Opportunities

Home > Collaboration > Our Focus > Computer Science > Microsoft Web N-gram Services

Microsoft Web N-gram Services

Access *petabytes of data* via the [Public Beta Web N-gram Services](#).

We invite the whole community to use the Web N-gram services, made available via a cloud-based platform, to drive discovery and innovation in web-scale language processing, speech, and related areas by conducting research on real-world web-scale data, taking advantage of regular data updates that benefit from dynamic data.

The Web N-gram services provide you access to:

- Content types: Document Body, Document Title, Anchor Texts
- Model types: Smoothed models
- N-gram availability: unigram, bigram, trigram, N-gram with N=4, 5
- Training size (Body): All documents indexed by Bing in the en-us n-gram corpus
- Access: Hosted Services by Microsoft
- Updates: Periodical updates

Late last year, we introduced a private beta testing of the Web N-gram services. We are now expanding access in the [Public Beta Web N-gram Services](#) to university professors, students, and researchers from around the world.

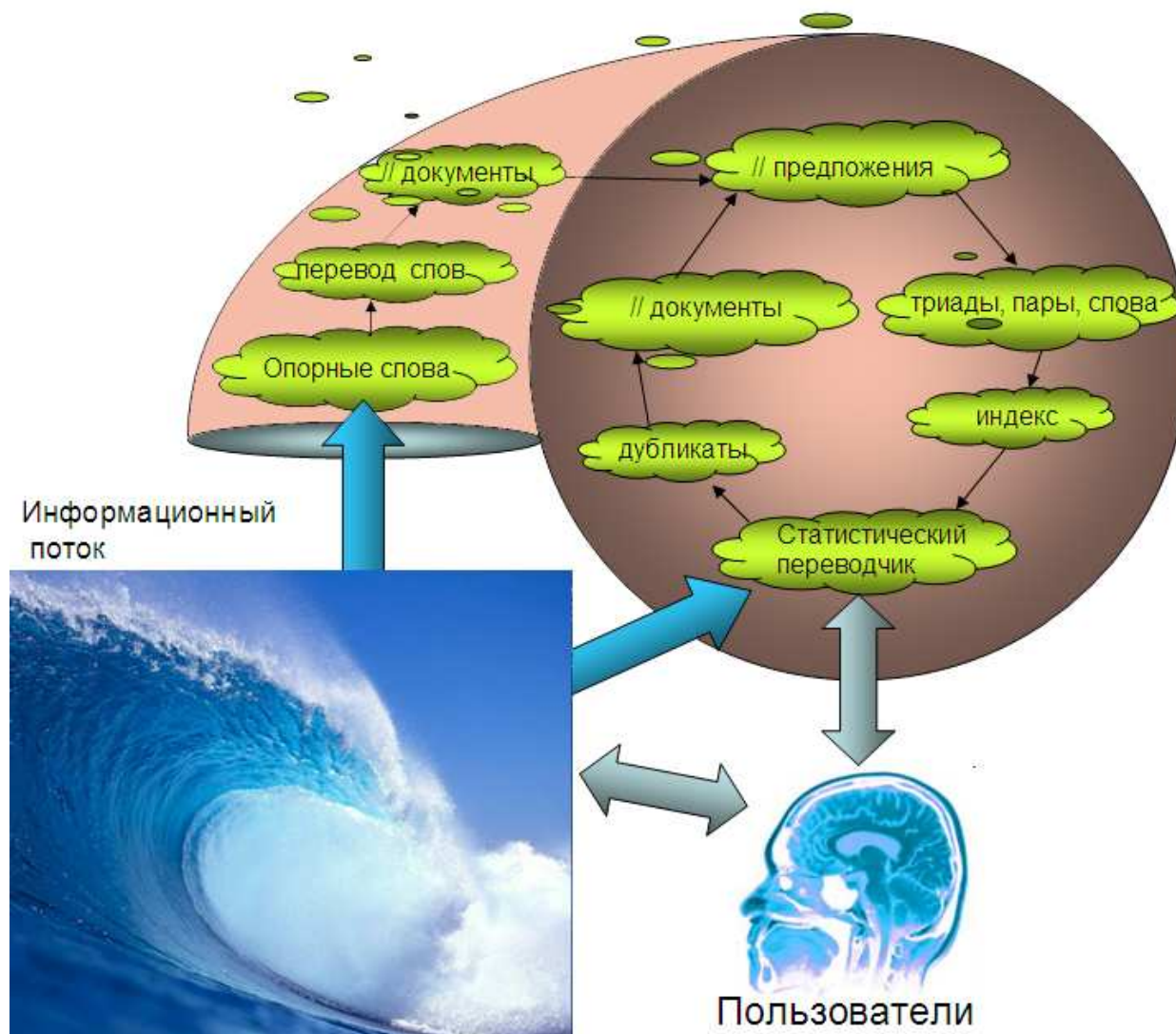
bing Web N-gram is brought to you by Microsoft Research in partnership with Microsoft Bing.

- An Overview of Microsoft Web N-gram Corpus and Applications, NAACL-HLT 2010

Learn More

- Web N-gram FAQ
- Web N-gram Community Site

Место в технологии перевода информационных потоков





*Горизонты прикладной лингвистики
и
лингвистических технологий*

СПАСИБО ЗА ВНИМАНИЕ!

Ландэ Дмитрий Владимирович,
dwl@visti.net

<http://ling.infostream.ua>

<http://dwl.visti.net>

Киев-2010