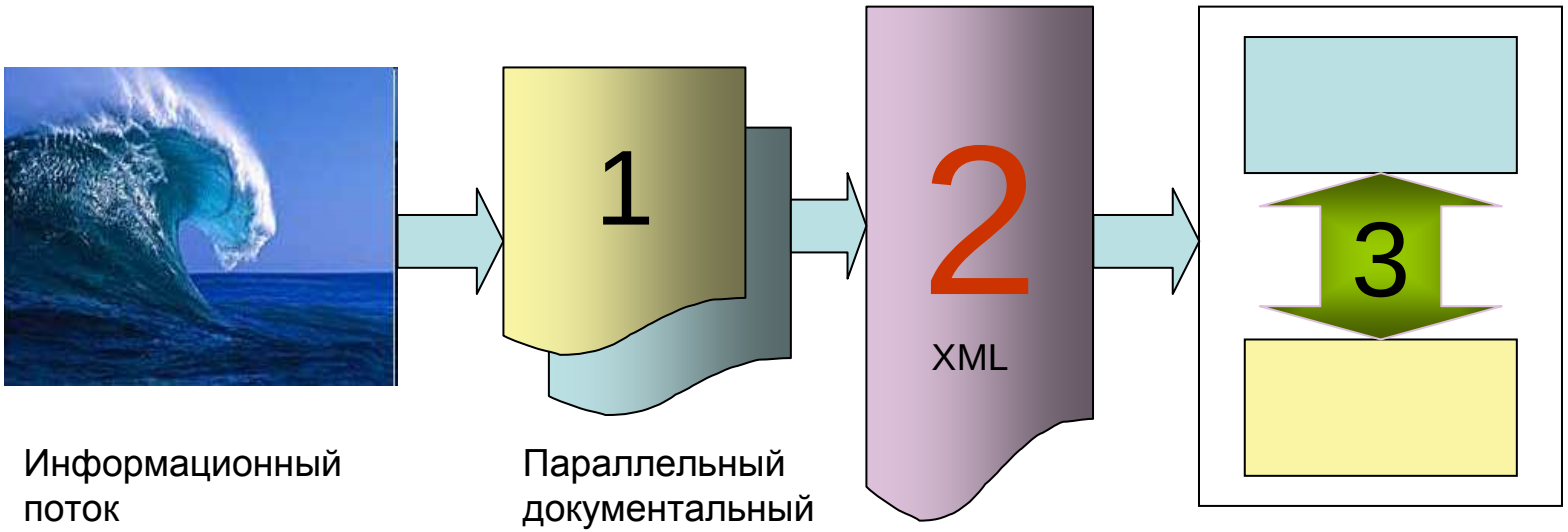
Megaling
Горизонты прикладной лингвистики
и
лингвистических технологий

# 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



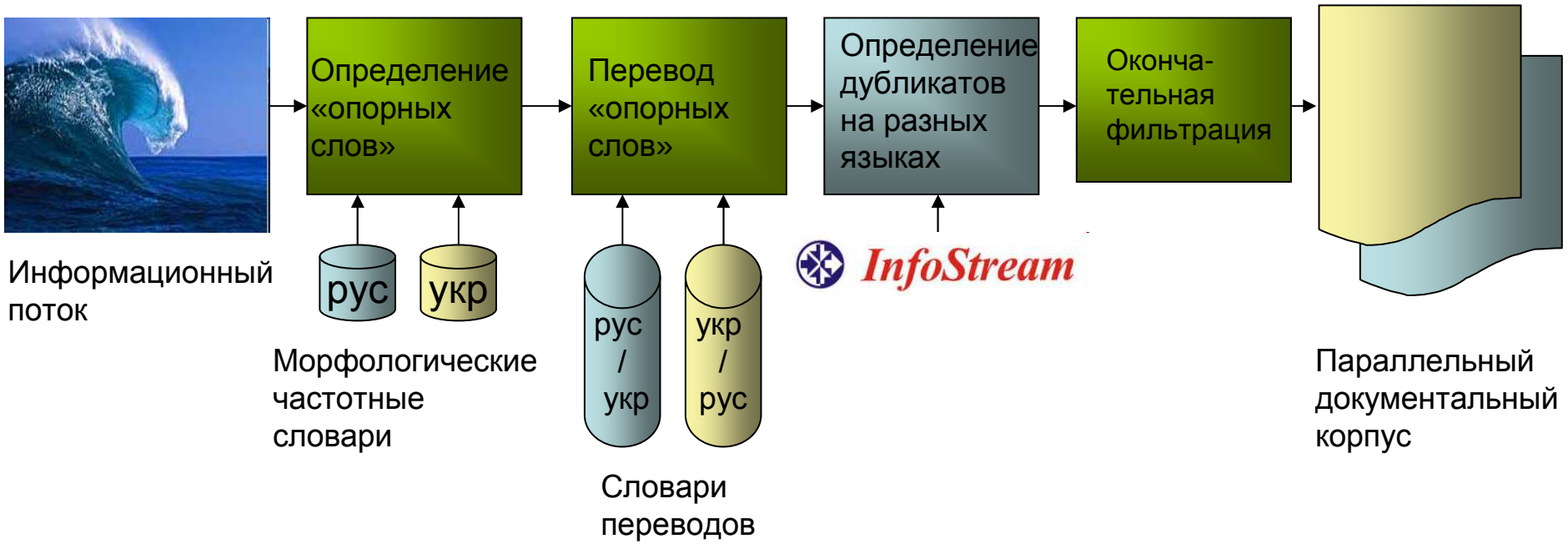

Информационный поток

Параллельный документальный корпус

Параллельный корпус предложений

Статистический потоковый переводчик

# Алгоритм создания параллельного документального корпуса

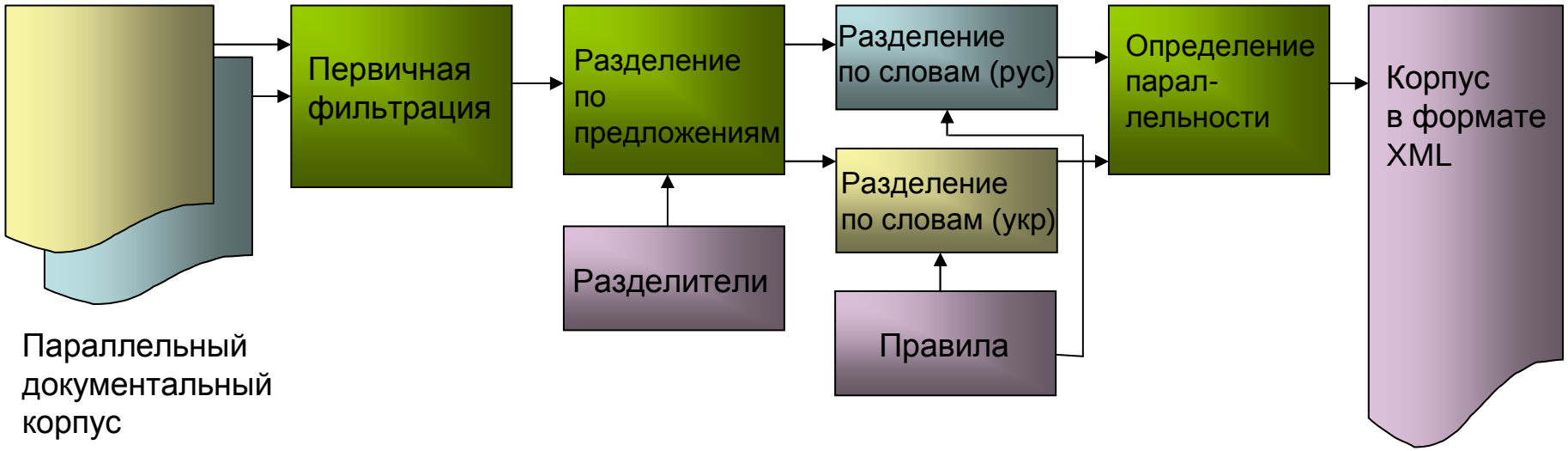


# Описание процедуры создания параллельного корпуса предложений

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

InfoStream - Украинско-русский параллельный текстовый корпус - Mozilla Firefox

http://ling.infostream.ua/

InfoStream - Украинско...

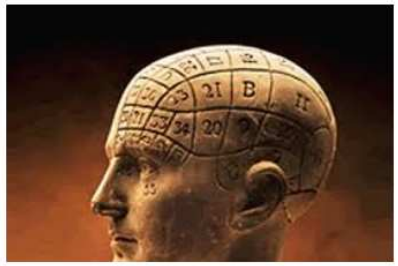
**Поиск в параллельном корпусе:**

Русский
Украинский

Морфология

Найти

**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 "**]**". 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.

   **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.

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.

   **"Сі Бриз-2010": українські десантники одягли німців у тільники**
   У рамках вчень "Сі бриз 2010" в Україні німецькі десантники, за плечами яких не один десяток стрибків з парашутом, вперше здійснили стрибки з українського вертольота Мі-8 під українськими куполами.

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

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

4. **Под Берлином взрываются снаряды**
   Борьба с огнем поке не очень успешна: пожарные не могут зайти в лес, чтобы приступить к тушению пожара, так как снаряды продолжают взрываться.

   **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

InfoStream Online

## 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.

## Міжнародна аерокосмічна виставка "Фарнборо-2010" пройде в Лондоні

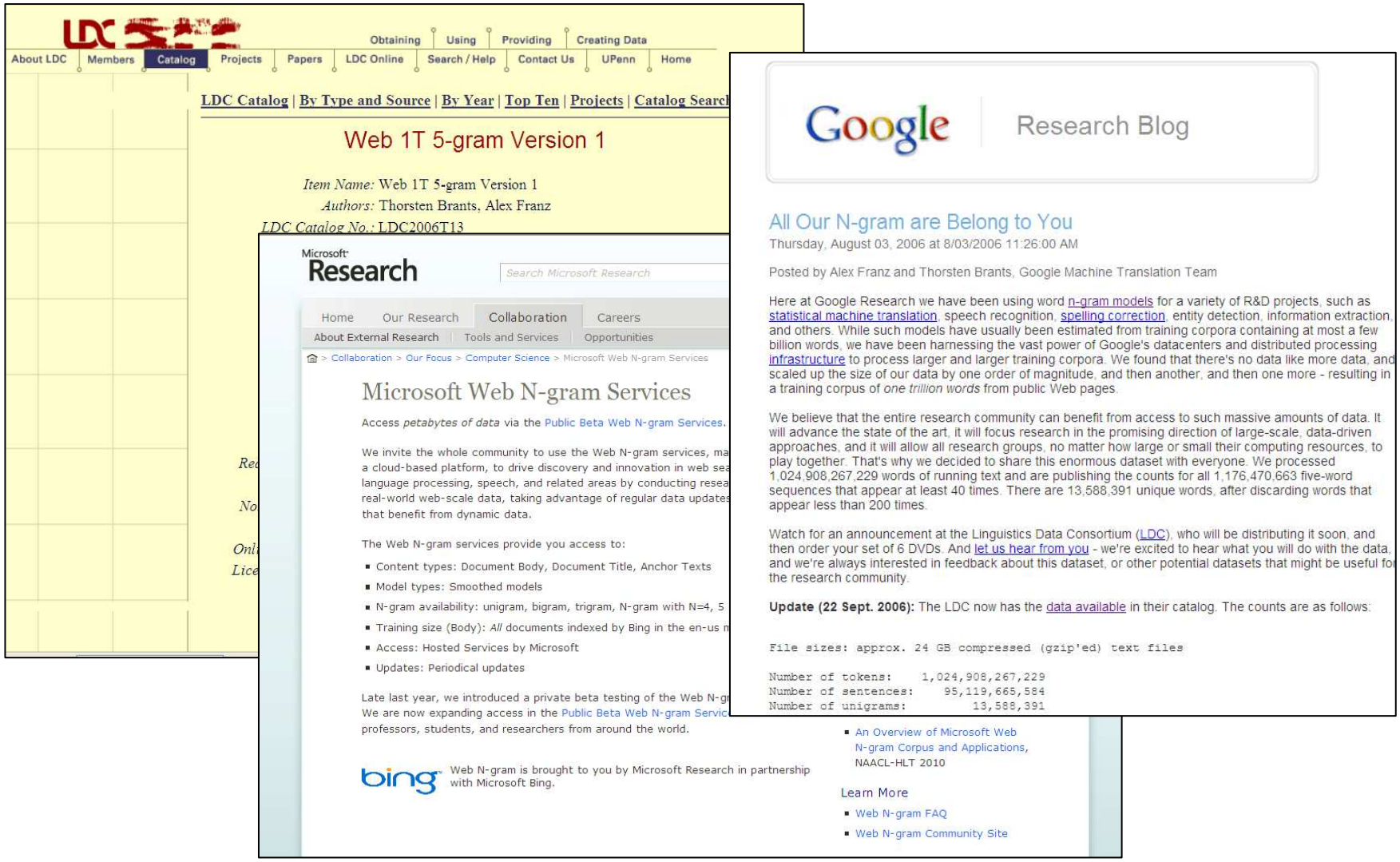
Близько 1350 учасників з 52 країн візьмуть участь у міжнародній аерокосмічній виставці "Фарнборо-2010", яка пройде в передмісті Лондона 19- 25 липня, повідомляє BBC Russia з посиланням на прес-службу Федеральної служби з військово-технічного співробітництва ( ФСВТС) Росії. "У статичній експозиції на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

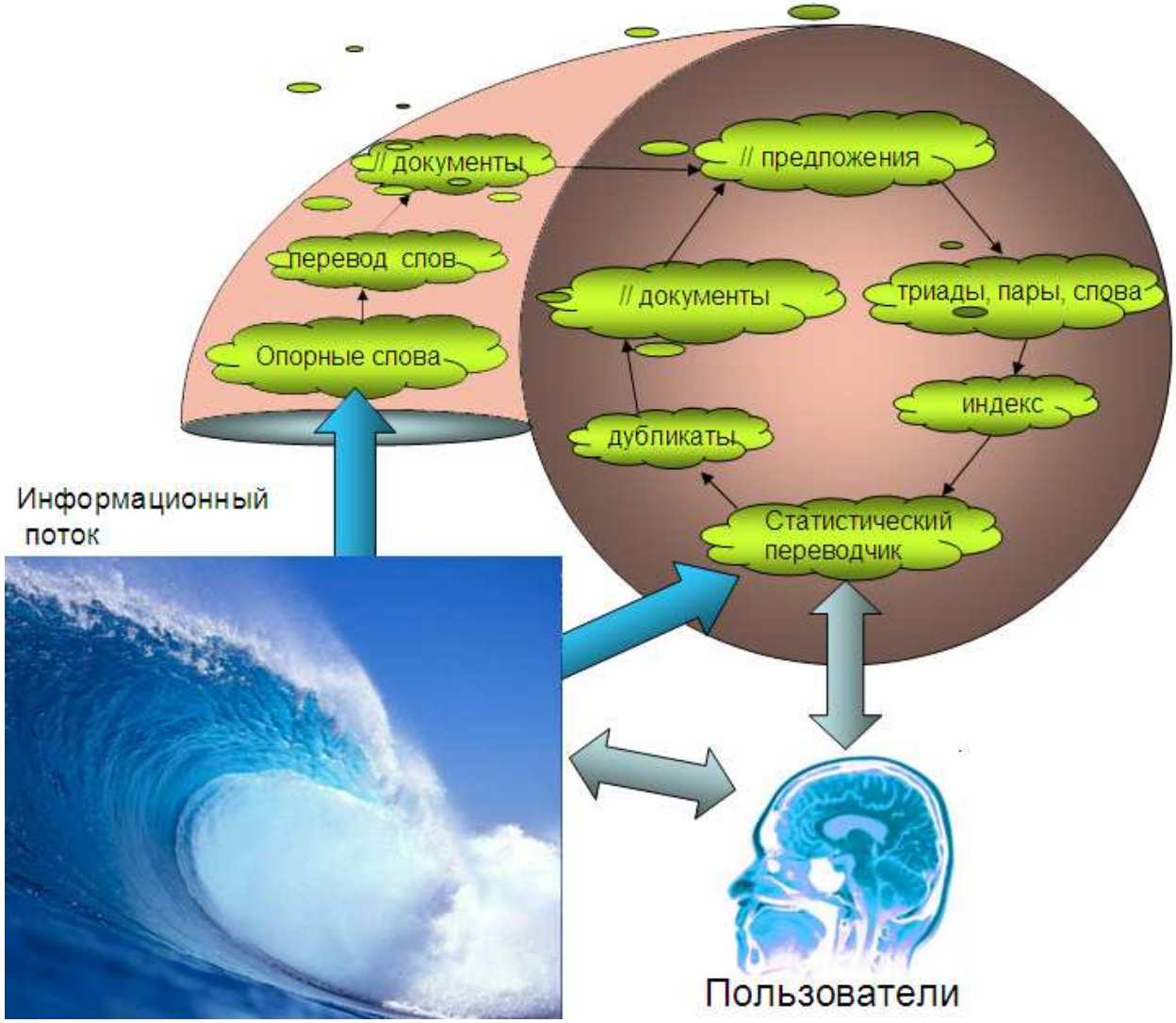
Закрыть

Наверх

# Примеры технологий n-gram



# Место в технологии перевода информационных потоков



СПАСИБО ЗА ВНИМАНИЕ!

Ландэ Дмитрий Владимирович,
dwl@visti.net

http://ling.infostream.ua
http://dwl.visti.net

Киев-2010