

10th Workshop on Building and  
Using Comparable Corpora  
at ACL'17

Vancouver, Canada



A parallel collection of clinical trials in Portuguese and English

*Mariana Neves*

*August 3rd, 2017*

# Parallel corpora of documents

- Parallel collections of documents are valuable resources for training, tuning and evaluating machine translation (MT) tools.
- However, these are not available for some domains, e.g., biomedicine, and manually creating such collections is an expensive task.

# Parallel corpora for news vs. biomedical domains

Corpora	cs-en	de-en	es-en	fr-en	hu-en	pl-en	ro-en	sv-en
CESTA	-	-	-	3,617	-	-	-	-
ECDC	2,324	2,379	2,357	2,377	2,306	2,202	2,363	2,345
EMA (OpenSubtitles)	445,365	481,443	487,901	493,933	462,541	459,225	424,904	466,108
EMA (new crawl)	687,635	615,256	-	-	-	652,336	621,490	-
Medical Web Crawl	-	-	148,982	-	-	-	-	-
Medical Web Texts from CzEng 1.6	7,029	-	-	-	-	-	-	-
MuchMore	-	28,919	-	-	-	-	-	-
PatTR Medical	-	1,830,647	-	2,191,537	-	-	-	-
Subtitles	3,140	77,937	151,675	120,841	-	3,010	116,335	96,575
<b>Total Parallel Segments</b>	<b>1,145,493</b>	<b>3,036,581</b>	<b>790,915</b>	<b>2,812,305</b>	<b>464,847</b>	<b>1,116,773</b>	<b>1,165,092</b>	<b>565,028</b>
<b>Total Parallel Segments (after 'sort   uniq')</b>	<b>819,697</b>	<b>2,662,810</b>	<b>631,087</b>	<b>2,634,229</b>	<b>351,336</b>	<b>800,662</b>	<b>852,800</b>	<b>444,777</b>
<b>Total Words (target language/en)</b>	<b>14M/15M</b>	<b>84M/94M</b>	<b>9M/10M</b>	<b>89M/100M</b>	<b>5M/5M</b>	<b>14M/14M</b>	<b>14M/15M</b>	<b>6M/5M</b>

## Out-of-Domain

We also included general domain data in the release. The following table summarizes the general purpose corpora included in the UFAL Medical Corpus collection:

Corpora	cs-en	de-en	es-en	fr-en	hu-en	pl-en	ro-en	sv-en
Cordis	-	-	-	-	-	168,067	-	-
EUbookshop	428,339	9,011,774	5,103,274	10,225,247	412,618	509,105	310,653	1,877,976
EUROPARL	643,361	1,918,724	1,964,134	2,006,305	621,328	627,367	396,882	1,852,450
Hunglish	-	-	-	-	2,083,159	-	-	-
JRC-Acquis	1,113,649	642,797	720,201	720,747	449,361	1,412,095	428,618	708,759
MultiUN	-	153,545	7,734,469	11,840,859	-	-	-	-
News Commentary	146,135	200,534	193,665	182,645	-	-	-	-
OpenSubtitles	44,618,012	12,815,341	75,947,825	49,035,989	44,612,969	34,926,913	59,732,934	17,840,535
PatTR Other	-	9,302,172	-	10,957,584	-	-	-	-
Rapid	-	-	-	-	-	132,156	-	-
<b>Total Parallel Segments</b>	<b>46,949,496</b>	<b>34,034,887</b>	<b>91,663,568</b>	<b>84,969,376</b>	<b>48,179,435</b>	<b>37,775,703</b>	<b>60,869,087</b>	<b>22,279,720</b>
<b>Total Parallel Segments (after 'sort   uniq')</b>	<b>38,065,775</b>	<b>31,638,916</b>	<b>75,421,729</b>	<b>74,045,053</b>	<b>39,499,594</b>	<b>31,786,926</b>	<b>47,829,602</b>	<b>19,447,606</b>
<b>Total Words (target language/en)</b>	<b>276M/333M</b>	<b>716M/817M</b>	<b>874M/889M</b>	<b>1,392M/1,490M</b>	<b>340M/262M</b>	<b>288M/229M</b>	<b>402M/377M</b>	<b>221M/195M</b>
Dictionaries	cs-en	de-en	es-en	fr-en	hu-en	pl-en	ro-en	sv-en
DBpedia	148,181	681,494	544,686	44,977	139,329	549,600	-	297,913
Linguee	-	51,571	-	-	-	-	-	-

# Sources for biomedical documents

- Clinical discharge summaries
  - Not available due to privacy issues
  - Usually monolingual

Discharge Summary for Ian TEST      MRN: 123432



LIVERPOOL HOSPITAL

SYDNEY SOUTH WEST  
AREA HEALTH SERVICE  
NSW HEALTH

Department of Newborn Care  
Locked Mailbag 7103, Liverpool BC, NSW 1871  
Telephone: 9828 5678 Facsimile: 9828 5582

## Neonatal Discharge Summary

11/04/2008

Dear Dr Jamie Smith

**RE:** Ian TEST      DOB: **28/02/2008**      MRN: **123432**  
(Male infant of Ann MRN: 123456)  
Unit 1 / 25 The Address  
LIVERPOOL 2170      Tel: (02) 9828 3000

Date of Admission: 28/02/08      **Consultant:** Dr Ian Callander  
Date of Discharge: 01/04/08

Thank you for accepting Ian TEST who was managed in the Newborn Care Centre for the following problems:

1. Extreme prematurity (26 weeks)
2. Very low birth weight (1070 gms)
3. Low Apgar scores
4. Respiratory Support
5. Jaundice
6. Infection
7. Necrotising enterocolitis
8. Patent ductus arteriosus

### MATERNAL HISTORY

Ann is a 28 year old G2 P1 (now) woman whose blood group is O positive. She was booked to deliver at Campbelltown Hospital under the care of Kaisher however delivered at Liverpool Hospital under the care of Dr Peter Hammill. She had a history of essential hypertension. This pregnancy was complicated by hypertension of pregnancy, fetal growth restriction, Bilateral Renal Pelvis dilatation 5 - 10mm, GBS +ve swab, fever, abnormal Dopplers, prolonged rupture of membranes for 2 days, clinically suspected chorioamnionitis. Ann was treated with antenatal steroids, tocolytics, and antihypertensive drugs. Following the spontaneous onset of labour, she proceeded to a vaginal delivery. Antibiotics were given before delivery.

### PERINATAL HISTORY

Ian was born at 13:00 hours with a birth weight of 1070 grams (76th centile). Apgars were 3 at 1 minute and 7 at 5 minutes respectively treated with intubation and ventilation. The



# Sources for biomedical documents

- Scientific publications
  - Many are freely available, but frequently monolingual (English)
  - There are some exceptions, e.g., Scielo, EDP

Biota Neotropica  
versão On-line ISSN 1676-0611

## Resumo

[WIDMER, Cynthia Elisa](#); [PERILLI, Miriam Lúcia Lages](#); [MATUSHIMA, Eliana Reiko](#) e [AZEVEDO, Fernando Cesar Cascelli](#). **Captura de jaguatiricas (*Leopardus pardalis*): armadilhas, iscas, ferimentos, imobilização e custos.** *Biota Neotrop.* [online]. 2017, vol.17, n.1, e20150125. Epub 16-Jan-2017. ISSN 1676-0611. <http://dx.doi.org/10.1590/1676-0611-bn-2015-0125>.

A captura de animais selvagens é capaz de proporcionar informações importantes acerca da estrutura da comunidade, dinâmica populacional, tamanho das áreas de vida, uso dos habitats, locais de toca, comportamento social e estado de saúde. Este estudo teve como objetivo descrever o método de captura enfatizando as iscas utilizadas, ferimentos, capturas de espécies não-alvo, anestesia e custos, para avaliar o sucesso de captura como parte de um programa de avaliação de saúde de jaguatiricas numa reserva de Mata Atlântica no Brasil. De um de esforço total de 1.011 armadilhas-noite em 86 dias, nós tivemos 68 eventos de captura compostos de jaguatiricas (22%, n= 15) e espécies não-alvo (78%, n= 53). Nós capturamos 10 indivíduos diferentes em 15 eventos de captura, correspondendo a 5,7 dias para capturar uma jaguatirica. A eficiência de captura foi de 14,8 jaguatiricas/1.000 armadilhas-noite. Nós sugerimos que os métodos de captura deveriam ser selecionados e implementados com base nos seguintes critérios: (i) alta eficiência de captura; (ii) alta seletividade; (iii) baixa taxa de ferimentos; (iv) alta adequação de imobilização; e (v) baixos custos, de forma a viabilizar comparações de estudos de diferentes grupos e diferentes áreas, permitindo a escolha do melhor método.

**Palavras-chave :** Brasil; custo de captura; eficiência de captura; Mata Atlântica; seletividade de captura; taxa de ferimentos.

Biota Neotropica  
versão On-line ISSN 1676-0611

## Resumo

[WIDMER, Cynthia Elisa](#); [PERILLI, Miriam Lúcia Lages](#); [MATUSHIMA, Eliana Reiko](#) e [AZEVEDO, Fernando Cesar Cascelli](#). **Live-trapping Ocelots (*Leopardus pardalis*): traps, baits, injuries, immobilization and costs.** *Biota Neotrop.* [online]. 2017, vol.17, n.1, e20150125. Epub 16-Jan-2017. ISSN 1676-0611. <http://dx.doi.org/10.1590/1676-0611-bn-2015-0125>.

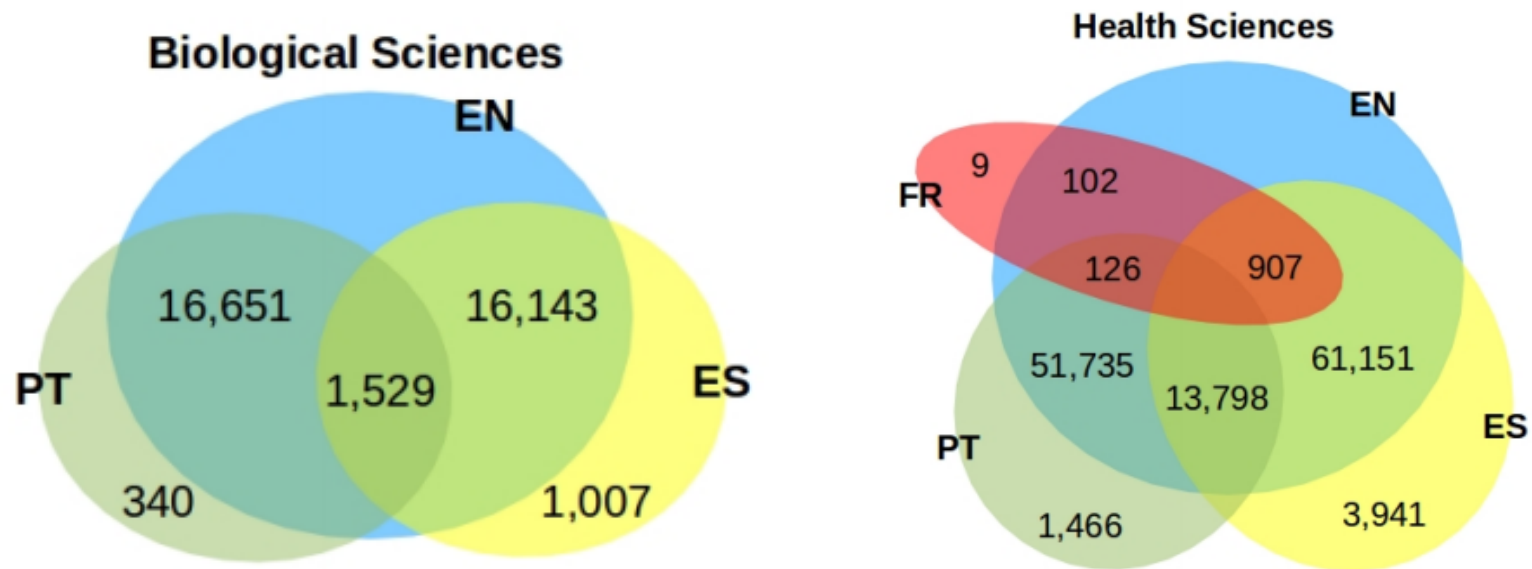
The capture of wild animals can provide important information on community structure, population dynamics, home range size, activity patterns, habitat use, denning, social behavior and health status. The objective of this study was to describe the method of capture with details on baits, injuries, non-target captures, anesthesia and costs, to evaluate its success as part of a health evaluation program of ocelots in a Brazilian Atlantic Forest Reserve. From a total of 1,011 trap-night effort in 86 days, we had 68 capture events composed of ocelots (22%, n=15) and non-target species (78%, n = 53). We captured 10 individual ocelots in 15 capture events, corresponding to 5.7 days to capture one ocelot. Capture efficiency was 14.8 ocelots/1,000 trap-nights effort. We suggest capture methods should be selected and implemented based on the following criteria: (i) high capture efficiency; (ii) high selectivity; (iii) low injury rate; (iv) high immobilization suitability; and (v) low costs, in order to enable comparisons of studies from different research groups and from different study areas, allowing a deliberate choice of the best method.

**Palavras-chave :** Atlantic forest; Brazil; capture cost; capture efficiency; capture selectivity; injury rate.

([http://www.scielo.br/scielo.php?script=sci\\_abstract&pid=S1676-06032017000100201&lng=pt&nrm=iso&tlng=en](http://www.scielo.br/scielo.php?script=sci_abstract&pid=S1676-06032017000100201&lng=pt&nrm=iso&tlng=en))  
([http://www.scielo.br/scielo.php?script=sci\\_abstract&pid=S1676-06032017000100201&lng=pt&nrm=iso&tlng=pt](http://www.scielo.br/scielo.php?script=sci_abstract&pid=S1676-06032017000100201&lng=pt&nrm=iso&tlng=pt))

# Sources for biomedical documents

- Scientific publications
  - Scielo corpus: the first parallel (comparable) corpus of scientific publications



# Sources for biomedical documents

- Scientific publications
  - Datasets from Scielo and EDP currently being used in the WMT Biomedical Translation Task

**EMNLP 2017  
SECOND CONFERENCE ON  
MACHINE TRANSLATION (WMT17)**

**September 7-8, 2017  
Copenhagen, Denmark**

**Shared Task: Biomedical Translation Task**

# Sources for biomedical documents

- Clinical trials
  - Freely (publicly) available but usually monolingual (English)



▶ Purpose

RATIONALE: Vitamin D may help prevent breast cancer.

PURPOSE: This randomized clinical trial is studying vitamin D and breast cancer biomarkers in female patients.

<u>Condition</u>	<u>Intervention</u>	<u>Phase</u>
<b>Breast Cancer</b>	Dietary Supplement: vitamin D Other: placebo	Phase 3

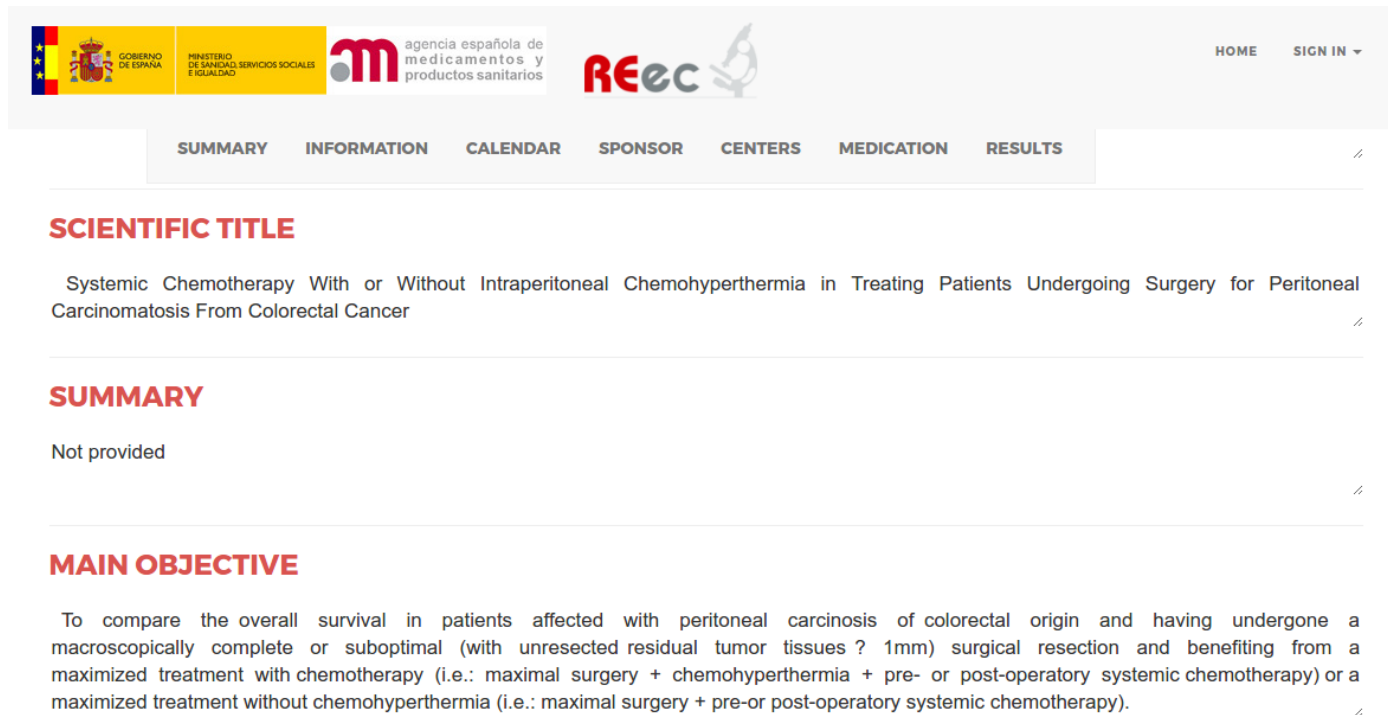
Study Type: Interventional  
 Study Design: Allocation: Randomized  
 Intervention Model: Parallel Assignment  
 Masking: Double Blind (Participant, Investigator)  
 Primary Purpose: Prevention


Official Title: Vitamin D and **Breast Cancer** Biomarkers



# Sources for biomedical documents

- Clinical trials
  - Even sometimes in countries whose native language isn't English...




HOME SIGN IN ▾

SUMMARY INFORMATION CALENDAR SPONSOR CENTERS MEDICATION RESULTS

**SCIENTIFIC TITLE**

Systemic Chemotherapy With or Without Intraperitoneal Chemohyperthermia in Treating Patients Undergoing Surgery for Peritoneal Carcinomatosis From Colorectal Cancer

**SUMMARY**

Not provided

**MAIN OBJECTIVE**

To compare the overall survival in patients affected with peritoneal carcinosis of colorectal origin and having undergone a macroscopically complete or suboptimal (with unresected residual tumor tissues ? 1mm) surgical resection and benefiting from a maximized treatment with chemotherapy (i.e.: maximal surgery + chemohyperthermia + pre- or post-operative systemic chemotherapy) or a maximized treatment without chemohyperthermia (i.e.: maximal surgery + pre-or post-operative systemic chemotherapy).

# Deutschen Register Klinischer Studien (DRKS)

- Clinical trials
  - Sometimes parallel documents are available but license doesn't allow its distribution

## Trial document

[Back to search results](#) | [Change history](#) |  PDF

DRKS-ID: **DRKS00000021**

### Trial Description

#### Title

Olanzapine augmentation therapy in t

#### Trial Acronym

[---]\*

#### URL of the Trial

[---]\*

#### Brief Summary in Lay Language

[---]\*

#### Brief Summary in Scientific Lan

The study is using a randomized double-blind, placebo-controlled design. Patients will be excluded by a score of 15 or higher on the HAM-D score at baseline. Patients will be randomized to either 10 mg/d Olanzapine or placebo. The primary endpoint is the reduction of the initial HAM-D score of responders for further 60 days.

## Studiendokument

DRKS-ID der Studie: **DRKS00000021**

### Studienbeschreibung

#### Titel der Studie

Olanzapin-Augmentationstherapie bei therapierefraktärer Depression: eine doppelblinde, plazebo-kontrollierte Studie

#### Studienakronym

[---]\*

#### Internetseite der Studie

[---]\*

#### Allgemeinverständliche Kurzbeschreibung

In dieser randomisierten, doppelblinden, plazebo-kontrollierten Studie sollen 60 Patienten mit therapierefraktärer Depression 2 Wochen mit Olanzapin 10 mg/Tag oder Plazebo behandelt werden. In einem beschreibenden Prä/Post-Vergleich soll untersucht werden, ob sich Hinweise auf eine mögliche augmentierende Wirkung in dieser Indikation finden.

#### Wissenschaftliche Kurzbeschreibung

In einer doppel-blinden, Plazebo-kontrollierten klinischen Studie soll eine mögliche augmentierende Wirkung des atypischen Antipsychotikums Olanzapin bei therapierefraktärer Depression untersucht werden. Eingeschlossen werden depressive Patienten, die nach 2 oder mehr ausreichend hoch und lange dosierten antidepressiven Behandlung keine ausreichende Besserung der Depression erreichen konnten. Response wird als Besserung des initialen Hamilton Depression Rating-Scale -Wertes um mehr als 50 % definiert. Bei Non-Respondern wird die Studie nach 14 Tagen beendet. Responder werden für weitere 60 Tage doppelblind mit Studienmedikation weiterbehandelt.  
Charakteristika

[Zurück zu den Suchergebnissen](#) | [Anderungshistorie](#) |  PDF

# Brazilian Clinical Trials Registry / Registro Brasileiro de Ensaios Clínicos (ReBEC)

**RBR-48pb9h**

**Estudo de Fase II, Randomizado, Duplo-cego para avaliar Carboplatina / Paclitaxel / CT-322 versus Carboplatina / Paclitaxel / Bevacizumabe como Tratamento de Primeira Linha do Câncer Recorrente ou Avançado de Pulmão de Células Não-pequenas com Histologia Não-Escamosa**

**Registration Date:** March 3, 2011, 8 a.m.

**Last Update:** June 13, 2011, 8:46 p.m.

**Study Type:**

Intervention Study

**Scientific Title:**

PT-BR

Estudo de Fase II, Randomizado, Duplo-cego para avaliar Carboplatina / Paclitaxel / CT-322 versus Carboplatina / Paclitaxel / Bevacizumabe como Tratamento de Primeira Linha do Câncer Recorrente ou Avançado de Pulmão de Células Não-pequenas com Histologia Não-Escamosa

PT-BR

A Randomized Double-Blinded Phase II Study of Carboplatin/Paclitaxel/CT-322 versus Carboplatin/Paclitaxel/Bevacizumab as First-Line Treatment for Recurrent or Advanced Non-Small Cell Lung Cancer with Non-Squamous Histology

# Overview of a clinical trial

## Health Conditions

### Health Condition(s) or Problem(s):

Câncer Recorrente ou Avançado de Pulmão de Células Não-pequenas com Histologia Não-Escamosa PT-BR

Recurrent or Advanced Non-Small Cell Lung Cancer with Non-Squamous Histology PT-BR

### General Descriptors for Health Condition(s):

**C00-D48: II - Neoplasias [tumores]** PT-BR

**C00-D48: II - Neoplasms** EN

**C04: Neoplasias** PT-BR

**C04: Neoplasias** ES

**C04: Neoplasms** EN

# Overview of a clinical trial

## Interventions:

PT-BR

Grupo de comparação: 127 pacientes receberão paclitaxel (200 mg/m<sup>2</sup>) e carboplatina no dia 1 de um ciclo de 21 dias e CT-322 administrado em esquema cego (2 mg/kg) semanalmente. Grupo controle: 127 pacientes receberão paclitaxel (200 mg/m<sup>2</sup>) e carboplatina no dia 1 de um ciclo de 21 dias, e bevacizumabe (15 mg/kg) no dia 1 e placebo nos dias 8 e 15 de um ciclo de 21 dias, em esquema cego, para correspondência com o esquema de administração do CT-322. Todos os pacientes receberão paclitaxel e carboplatina por, no máximo, 6 ciclos de tratamento (cada ciclo de 21 dias), ou até que apresentem doença prograssiva

PT-BR

Comparison group: 127 patients receive paclitaxel (200 mg/m<sup>2</sup>) and carboplatin on day 1 of a cycle of 21 days and CT-322 administered in blinded (2 mg / kg) weekly. Control group: 127 patients receive paclitaxel (200 mg/m<sup>2</sup>) and carboplatin on day 1 of a 21-day cycle, and bevacizumab (15 mg / kg) on day 1 and placebo on days 8 and 15, a 21-day cycle, in blinded, to match the schedule of administration of CT-322. All patients will receive carboplatin and paclitaxel for a maximum of 6 treatment cycles (each cycle 21 days) or until they have documented progressive disease or unacceptable toxic signs develop, or withdraw consent. whichever occurs first.

# Overview of a clinical trial

## Inclusion Criteria:

PT-BR

Os pacientes deverão assinar um termo de consentimento antes da realização de qualquer procedimento relacionado ao estudo.

Pacientes com Câncer de Pulmão de Células Não-Pequenas confirmado histológica ou citologicamente, em estágio IIIB (derrame pleural maligno), estágio IV ou recorrente;

Doença mensurável segundo os critérios Critérios de Avaliação da Resposta em Tumores Sólidos, com pelo menos 1 lesão-alvo fora de qualquer campo prévio de radioterapia;

Status de performance do Grupo de Cooperação em Oncologia da Região

PT-BR

Subjects must sign an informed consent prior to any study-related procedures.

Histologically or cytologically confirmed, stage IIIB (malignant pleural effusion), stage IV or recurrent Non Small Cell Lung Cancer;

Measurable disease by Response Evaluation Criteria In Solid Tumors guidelines, with at least 1 target lesion outside any previous radiotherapy field; Eastern Cooperative Oncology Group performance status < 1;

Life expectancy of at least 3 months; Accessible for treatment and follow-up.

Subjects enrolled in this trial must be treated at the participating center(s).



# Overview of a clinical trial

## Exclusion Criteria:

PT-BR

Mulheres com potencial de engravidar que não desejarem ou não conseguirem utilizar um método contraceptivo aceitável durante todo o período de estudo e por até 6 semanas após a última dose do produto de investigação

Gestantes ou lactantes

Mulheres com um teste de gravidez positivo no momento da inclusão no estudo ou antes da administração do produto de investigação

Homens férteis e sexualmente ativos que não estejam em uso de um método contraceptivo eficaz caso as suas parceiras sejam consideradas Mulheres com potencial de engravidar.

PT-BR

Women of childbearing potential who are unwilling or unable to use an acceptable method to avoid pregnancy for the entire study period [and for up to 6 weeks after the last dose of investigational product]

Women who are pregnant or breastfeeding

Women with a positive pregnancy test on enrollment or prior to investigational product administration

Sexually active fertile men not using effective birth control if their partners are Women of childbearing potential.

Evidence of predominantly squamous-cell histology (mixed cell type tumors only)

Known central nervous system metastasis

# Overview of a clinical trial

## Primary Outcomes:

PT-BR

O objetivo primário é comparar a Sobrevida Livre de Progressão proporcionada pelo CT-322 com aquela proporcionada pelo bevacizumabe em combinação com a carboplatina e o paclitaxel, em pacientes virgens de quimioterapia e portadores de Câncer de Pulmão de Células Não-Pequenas não-escamoso recorrente ou avançado.

EN

The primary objective is to compare the Progressio Free Survival of CT-322 versus bevacizumab in combination with carboplatin and paclitaxel in chemonative subjects with recurrent or advanced non-squamous Non Small Cell Lung Cancer.

# Corpus construction: Pipeline

- Data download
- OpenXML Trials parsing
- Sentence splitting
- Sentence alignment
- Quality checking

# Data download

[HOME](#) / REGISTERED TRIALS

All below

<input checked="" type="checkbox"/>	<b>Title</b>	<b>Primary Id Number</b>	<b>RBR-88btzp</b>
	Evaluation of the Implantation of an Anticoagulation Clinic in the Assistance of Chagas and Non-Chagas Patients at the Hospital of Clinics of the Universidade Federal de Minas Gerais - UFMG	<b>Recruitment Status</b>	recruitment completed
		<b>Date of Registration</b>	May 10, 2011, 7:20 p.m.
<input checked="" type="checkbox"/>	<b>Title</b>	<b>Primary Id Number</b>	<b>RBR-36w269</b>
	A Randomized Double-Blind Phase 3 Trial Comparing Docetaxel Combined with Dasatinib to Docetaxel Combined with Placebo in Castration-Resistant Prostate Cancer	<b>Recruitment Status</b>	recruitment completed
		<b>Date of Registration</b>	June 13, 2011, 7:25 p.m.
<input checked="" type="checkbox"/>	<b>Title</b>	<b>Primary Id Number</b>	<b>RBR-48pb9h</b>
	A Randomized Double-Blinded Phase II Study of Carboplatin/Paclitaxel/CT-322 versus Carboplatin/Paclitaxel/Bevacizumab as First-Line Treatment for Recurrent or Advanced Non-Small Cell Lung Cancer with Non-Squamous Histology	<b>Recruitment Status</b>	recruiting
		<b>Date of Registration</b>	June 13, 2011, 8:46 p.m.

# Data download

	Date of Registration	
<input checked="" type="checkbox"/> <b>Title</b>	<b>Primary Id Number</b>	<a href="#">RBR-4y59tq</a>
<a href="#">Effect of the Informative Paper on patients and families in a Oncogenetic Ambulatory :evaluation of information and perception of coercion.</a>	<b>Recruitment Status</b>	not yet recruiting
	<b>Date of Registration</b>	June 30, 2011, 12:51 a.m.
<input checked="" type="checkbox"/> <b>Title</b>	<b>Primary Id Number</b>	<a href="#">RBR-4hb6f6</a>
<a href="#">Evaluation of electroacupuncture on the treatment of neck and sholder pain</a>	<b>Recruitment Status</b>	data analysis completed
	<b>Date of Registration</b>	July 2, 2011, 10:08 a.m.

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [23](#) [24](#) [25](#) [26](#) [27](#) [28](#) [29](#) [30](#) [31](#) [32](#) [33](#) [34](#) [35](#) [36](#) [37](#) [38](#) [39](#) [40](#) [41](#) [42](#) [43](#) [44](#) [45](#) [46](#) [47](#) [48](#) [49](#) [50](#) [51](#) [52](#) [53](#) [54](#) [55](#)  
[56](#) [57](#) [58](#) [59](#) [60](#) [61](#) [62](#) [63](#) [64](#) [65](#) [66](#) [67](#) [68](#) [69](#) [70](#) [71](#) [72](#) [73](#) [74](#) [75](#) [76](#) [77](#) [78](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [100](#) [101](#) [102](#) [103](#) [104](#) [105](#)  
[106](#) [107](#) [108](#) [109](#) [110](#) [111](#) [112](#) [113](#) [114](#) [115](#) [116](#) [117](#) [118](#) [119](#) [120](#) [121](#) [122](#) [123](#) [124](#) [125](#) [126](#) [127](#) [128](#) [129](#) [130](#) [131](#) [132](#) [133](#) [134](#) [135](#) [136](#) [137](#) [138](#) [139](#) [140](#) [141](#) [142](#) >

>>

(120 links as of January 4th)

# OpenXML parsing

```

-<trials version="1">
  -<trial language="en" status="published" date_registration="2011-06-13" created="2011-03-03" updated="2011-06-13">
    -<trial_identification>
      <trial_id>RBR-48pb9h</trial_id>
      <utrn_number>U1111-1119-7435</utrn_number>
      <reg_name>REBEC</reg_name>
    -<public_title>
      A Randomized Double-Blinded Phase II Study of Carboplatin/Paclitaxel/CT-322 versus Carboplatin/Paclitaxel/Bevacizumab as First-Line Treatm
      Cell Lung Cancer with Non-Squamous Histology
    </public_title>
    <acronym/>
    <acronym_expansion/>
    -<scientific_title>
      A Randomized Double-Blinded Phase II Study of Carboplatin/Paclitaxel/CT-322 versus Carboplatin/Paclitaxel/Bevacizumab as First-Line Treatm
      Cell Lung Cancer with Non-Squamous Histology
    </scientific_title>
    <scientific_acronym/>
    <scientific_acronym_expansion/>
  </trial_identification>
  -<sponsors_and_support>
    -<primary_sponsor country_code="BR" type="industry">
      <name>Bristol-Myers Squibb</name>
      -<address>
        Rua Carlos Gomes, 924 santo Amaro São Paulo - SP 04743-903
      </address>
      <state/>
      <city/>
    </primary_sponsor>
    -<source_support country_code="BR" type="industry">
      <name>Bristol-Myers Squibb</name>
      -<address>
        Rua Carlos Gomes, 924 santo Amaro São Paulo - SP 04743-903
      </address>
      <state/>
      <city/>
    </source_support>
  </sponsors_and_support>
  </trial>
</trials>

```



# OpenXML parsing

```

- <interventions>
  <i_code value="drug"/>
  - <keyword vocabulary="decs" version="" code="E02.183.750.500">
    <text>Antineoplastic Combined Chemotherapy Protocols</text>
    - <text_translation lang="es">
      Protocolos de Quimioterapia Combinada Antineoplásica
    </text_translation>
    - <text_translation lang="pt-br">
      Protocolos de Quimioterapia Combinada Antineoplásica
    </text_translation>
  </keyword>
  - <keyword vocabulary="icd-10" version="" code="Z51.1">
    <text>Chemotherapy session for neoplasm </text>
    <text_translation lang="es">Sesión de quimioterapia por tumor </text_translation>
    <text_translation lang="pt-br">Sessão de quimioterapia por neoplasia</text_translation>
  </keyword>
  - <freetext>
    Comparison group: 127 patients receive paclitaxel (200 mg/m2) and carboplatin on day 1 of a cycle of 21 days
    patients receive paclitaxel (200 mg/m2) and carboplatin on day 1 of a 21-day cycle, and bevacizumab (15 mg /
    match the schedule of administration of CT-322. All patients will receive carboplatin and paclitaxel for a maxim
    documented progressive disease or unacceptable toxic signs develop, or withdraw consent, whichever occurs f
    day cycle) or placebo and bevacizumab (bevacizumab on day 1 of cycle and placebo on days 8 and 15 of the cy
    or until unacceptable toxic signs develop, or withdraw consent, whichever occurs first. Assessments of efficacy
    death or the introduction of a subsequent therapy for Lung Cancer Non-Small Cell. Security will be evaluated v
    cycle 21 days). After the sixth cycle, patients enter the maintenance period and will only receive CT-322 (on day
    (bevacizumab on day 1 and placebo in cycle 8 and 15 cycle) according to the arm of the study until they have d
    withdraw consent, whichever occurs first.
  </freetext>
</interventions>

```

# OpenXML parsing

```

-<outcomes>
  -<primary_outcome value="The primary objective is to compare the Progressio Free S
    subjects with recurrent or advanced non-squamous Non Small Cell Lung Cancer.">
    <outcome_translation value="" lang="es"/>
    <outcome_translation value="O objetivo primário é comparar a Sobrevida Livre d
      combinação com a carboplatina e o paclitaxel, em pacientes virgens de quimioterapia
      " lang="pt-br"/>
  </primary_outcome>
  -<primary_outcome value="The primary efficacy analysis will be performed for all ran
    progression free survival events have occurred. The difference between the 2 treatment
    the primary analysis. In addition, a sensitivity analysis will be performed to test this diff
    Group Performance Status (0 vs 1) and disease stage (Stage IIIB or IV). Additional analy
    survival functions. The progression free survival hazard ratio of CT-322 to bevacizumab
    computed using unadjusted and adjusted Cox proportional hazards modeling. Cox mode
    an adjusted model which will include a pre-defined list of covariates (listed below) as we
    function for each treatment arm will be estimated using Kaplan-Meier methodology. In a
    free survival will be computed by treatment arm. The following are the prognostic facto
    Cooperative Oncology Group Performance Status and disease stage. ">
    <outcome_translation value="" lang="es"/>
    <outcome_translation value="Todos os pacientes randomizados serão submetidos
      realizada quando tiverem ocorrido 170 eventos de Sobrevida Livre de Progressão. A
      para manter o nível do alfa em 0,15. Esta será considerada a análise primária. Além
      rank" monocaudal para manter o nível do alfa em 0,15 estratificado pelo status de pe
      REGIÃO LESTE (0 vs 1) e o estágio da doença (Estágio IIIB ou IV). As análises adicio
      das funções de sobrevida. Serão computados a proporção de riscos da Sobrevida Livre
      associado e um intervalo de confiança de 95% bicaudal, usando-se modelos de riscos
      tratamento estratificado pelos fatores de estratificação acima mencionados e um mo
      estratificação acima mencionado, todos na condição de covariadas. A função de Sobr
      Kaplan-Meier. Além disso, um limite de confiança de 85% monocaudal e um intervalo
      tratamento. Estes são os fatores prognósticos a serem incluídos nos modelos de risco
      PERFORMANCE DO GRUPO DE COOPERAÇÃO EM ONCOLOGIA DA REGIÃO LEST
    </primary_outcome>
  
```

# OpenXML parsing

-<hc\_freetext>  
Câncer Recorrente ou Avançado de Pulmão de Células Não-pequenas com Histologia Não-Escamosa  
</hc\_freetext>

-<i\_freetext>  
Grupo de comparação: 127 pacientes receberão paclitaxel (200 mg/m<sup>2</sup>) e carboplatina no dia 1 de um ciclo de 21 dias, em esquema cego, para correspondência com o esquema de administração do CT-322. Todos os tratamentos (cada ciclo de 21 dias), ou até que apresentem doença progressiva documentada, ou desenvolvimento independentemente do que ocorrer primeiro. Os pacientes dos Braços A e B receberão CT-322 (nos dias 1, 8 e 15 do ciclo) e placebo nos dias 8 e 15 do ciclo, respectivamente, em esquema cego, até que apresentem doença inaceitável, ou retirem o consentimento, independentemente do que ocorrer primeiro. As avaliações de sobrevida ou a introdução de uma terapia subsequente para Câncer de Pulmão de Células Não-Pequenas. A duração do período de tratamento consiste de 6 ciclos (cada ciclo de 21 dias). Após o ciclo 6, os pacientes entrarão em um período de 21 dias) ou bevacizumabe e placebo (bevacizumabe no dia 1 do ciclo e placebo nos dias 8 e 15 do ciclo), ou até que desenvolvam manifestações tóxicas inaceitáveis, ou retirem o consentimento, independentemente do que ocorrer primeiro.  
</i\_freetext>

-<inclusion\_criteria>  
Os pacientes deverão assinar um termo de consentimento antes da realização de qualquer procedimento diagnóstico ou terapêutico. Câncer de Células Pequenas confirmado histológica ou citologicamente, em estágio IIIB (derrame pleural maligno), estágio IIIA ou IV da Resposta em Tumores Sólidos, com pelo menos 1 lesão-alvo fora de qualquer campo prévio de radioterapia, e com expectativa de vida < 1; Expectativa de vida de pelo menos 3 meses; Acessibilidade ao tratamento e seguimento. Os pacientes deverão ser capazes de compreender o estudo e participar do estudo. O participante(s). Disposição para fornecer uma amostra de sangue total para o estudo de proteínas e polímeros de cadeia longa e o Fator de Crescimento do Endotélio Vascular. Pacientes de ambos os sexos >18 anos de idade. Mulheres com potencial de engravidar adequado ao longo de todo o estudo e por um período de até 6 semanas após a última dose do produto (o potencial de engravidar inclui qualquer mulher que já tenha tido a sua menarca e não tenha sido submetida a ligadura tubária bilateral ou ooforectomia bilateral) e nem seja considerada pós-menopáusia. Define-se como mulheres com períodos menstruais irregulares e que estejam em uso de terapia de reposição hormonal, Internacionais/mL Mulheres que estejam em uso de contraceptivos orais, outros contraceptivos hormonais (pílulas injetáveis), ou aquelas em uso de produtos mecânicos como, por exemplo, dispositivo intrauterino, ou de qualquer método para evitar a gravidez, ou que estejam praticando abstinência sexual ou tenham parceiro estéril (submetidos, ou não submetidos, à contracepção). Mulheres com potencial de engravidar deverão apresentar um teste de gravidez no soro ou em urina (ou em unidades equivalentes de Gonadotrofina coriônica humana) nas 72 horas que antecederem a introdução  
</inclusion\_criteria>

# Open XML parsing

- Fields that have been considered:
  - (a) trial identifier
  - (b) public title
  - (c) scientific title
  - (d) interventions to be carried out
  - (e) inclusion criteria for taking part
  - (f) exclusion criteria for not participating
  - (g) primary outcome
  - (h) secondary outcome

Final documents based on the concatenation of the various fields  
**following the order in the OpenXML Trials file**

# Sentence splitting

- „Sentence Detector“ models for EN and PT



# Sentence alignment

## Geometric Mapping and Alignment (GMA)

by [Ali Argyle](#), Luke Shen, Svetlana Stenchikova, and [I. Dan Melamed](#)

- Default parameters of GMA
- List of stopwords
  - EN: <http://www.textfixer.com/tutorials/common-english-words.txt>
  - PT: <http://www.linguateca.pt/chave/stopwords/chave.MF300.txt>  
and English



# Quality checking

(Sample of 50 trials)

003/891

rebec train sample 17

Efeito da laserterapia no músculo masseter de crianças com paralisia cerebral **Grupo experimental: Será composto por 30 crianças com diagnóstico de paralisia cerebral, cujos cuidadores relatam dificuldade de higienização por diminuição de abertura bucal, travamento de utensílios usados para a alimentação e história de trama em tecidos orais.** null

— Source

**The experiental group will be composed of 30 children diagnosed with cerebral palsy whose caregivers to report difficulty in cleaning by decreased mouth opening, locking utensils used for food and story plot in oral tissues.**

— Translation

OK

Source>Target

Target>Source

Overlap

No alignment

# Results

- Clinical trials corpus: 1188 documents
  - EN: 23,843 sentences, 625,881 tokens
  - PT: 23,666 sentences, 665,325 tokens

# Results

- Manual validation of 50 trials
  - 67% of the sentences are correctly aligned
  - 28% of the sentences were not aligned
  - 5% of the sentences had some overlap
- In contrast, our results for the Scielo corpus had a >80% correct alignment

# Discussion

- Many of the wrong alignments were due to shifted sentences
  - Mainly due to fields being placed in different order due to multiple instances of the same type

## Primary Outcomes:

PT-BR

Para a hipótese I não há diferença quanto à longevidade clínica das restaurações de classe II de ART (Tratamento Restaurador Atraumático), em dentes decíduos, com e sem retenção adicional. Para a hipótese II não há diferença quanto à longevidade clínica das restaurações de classe II de ART em dentes permanentes com retenção adicional em comparação com as restaurações de classe II de resina compostas.

EN

For Hypothesis I there is no difference in the clinical performance of ART class II restorations in deciduous teeth with and without additional retentions. For Hypothesis II there is no difference in the clinical performance of ART class II restorations in permanent teeth with additional retentions in comparison with composite resin restorations.

PT-BR

Os critérios para essa avaliação serão do Tratamento Restaurador Atraumático (ART) e do Sistema de Avaliação de Saúde Pública dos Estados Unidos (USPHS) após seis meses e um ano e dois anos e três anos.

EN

The criteria for this evaluation will be the Atraumatic Restorative Treatment (ART) and the US Public Health Assessment System (USPHS) after six months and one year and two years and three years.

# Discussion

- Some few errors were due to wrong sentence splitting

Subject must be at least 18 years of age; males and females with a documented diagnosis of ulcerative colitis (UC) at least 4 months prior to entry into the study; subjects with moderately to severely active UC based on Mayo score criteria; subjects must have failed or be intolerant of at least one of the following treatments for UC: corticosteroids (oral ou intravenous), azathioprine or 6 mercaptopurine (6MP), anti TNF alpha therapy (infliximab ou adalimumab).

# Discussion

- Some errors were due to splitting of (very) long sentences

Diseases which cause damage in the intestinal mucosa, diseases that significantly increase the gastrointestinal transit as infectious enteritis, celiac disease, inflammatory bowel disease (Chron), drug-induced enteritis or radiation, diverticular disease of the colon; History of surgery: heart (whatever), renal (exercises kidney or renal agenesis), intestinal (partial or total removal of the esophagus, stomach, duodenum, jejunum, ileum, ascending colon, transverse colon, descending colon, sigmoid colon or rectum) , liver or pancreas; Volunteers smoking more than five cigarettes a day; different eating habits of the population standard, eg vegetarianism, veganism; History of alcohol consumption or use of drugs of abuse; Made use of antibiotics as regular medication (continuous use) within the 4 weeks preceding the valuation date and / or the start of the breath test; This examination Colonoscopy one month before the breath test H2 expired.

# Conclusions

- A novel comparable/parallel corpus of clinical trials for EN/PT
  - Reasonable size, easy to obtain and freely available
- However, further processing is necessary to improve the quality of the corpus.
- Experiments still pending to evaluate its suitability for MT.
- Available at:
  - <https://github.com/biomedical-translation-corpora/wmt-task>



# Thank you!

Looking forward to answering your questions!

Mariana Neves

Current email: [Mariana.Lara-Neves@bfr.bund.de](mailto:Mariana.Lara-Neves@bfr.bund.de)

Current affiliation: German Federal Institute for Risk Assessment