



FAKULTÄT
FÜR INFORMATIK

Faculty of Informatics

Bootstrapping a Comparable Corpus from Patent Family Members

Mihai Lupu

lupu@ifs.tuwien.ac.at

Vienna University of Technology

ESTeam AB

Outline

- Motivation
- Background
- Data
- Method
- Test case & Evaluation
 - Error analysis
- Conclusions

Motivation

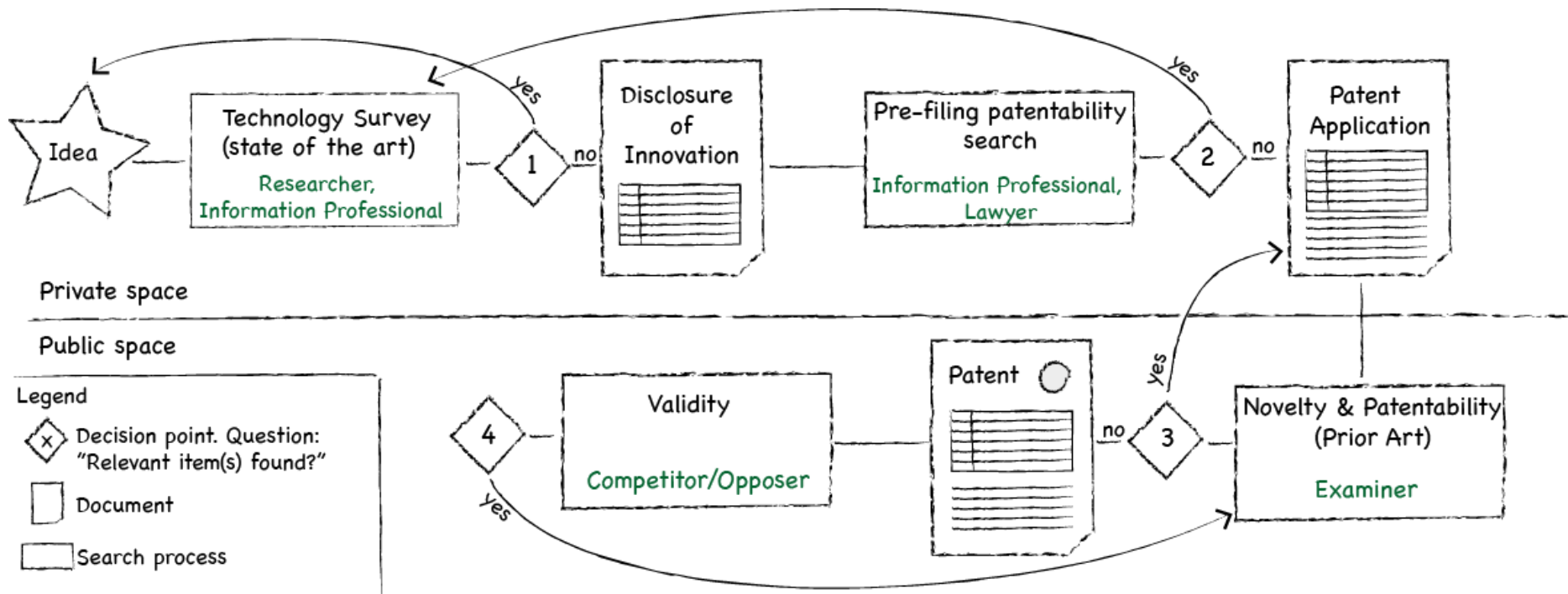
Cross-Lingual IR

- Patent search is [generally] independent of language
 - Often, it **must** be multi-lingual
 - Machine translation
 - Adapted to the patent genre
 - Needs training

Background - Patent

- WIPO definition:
 - A patent is an **exclusive right** granted for an invention, which is a **product or a process** that provides, in general, a **new way of doing** something, or offers a new technical solution to a problem. In order to be patentable, the invention must fulfill certain conditions:
 - It must be of **practical use**;
 - It must show an element of **novelty**: some new characteristic which is not known in the body of existing knowledge in its technical field. This body of existing knowledge is called "prior art".
 - It must show an **inventive step** which could not be deduced by a person with average knowledge of the technical field.
 - Its subject matter must be accepted as "**patentable**" under law.

Background – Patent search



Background – Patent documents



US 2003/0076301A1

(19) **United States**
 (12) **Patent Application Publication** (10) Pub. No.: US 2003/0076301 A1
 Tsuk et al. (43) Pub. Date: Apr. 24, 2003

(54) **METHOD AND APPARATUS FOR ACCELERATED SCROLLING** Publication Classification
 (75) Inventors: Robert W. Tsuk, Cupertino, CA (US); Jeffrey L. Robbin, Los Altos, CA (US)
 (51) Int. Cl.⁷ G09G 5/08
 (52) U.S. Cl. 345/159

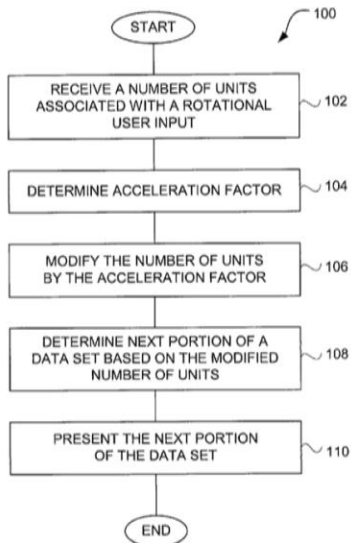
Correspondence Address:
 BEYER WEAVER & THOMAS LLP
 P.O. BOX 778
 BERKELEY, CA 94704-0778 (US)

(73) Assignee: Apple Computer, Inc.
 (21) Appl. No.: 10/256,716
 (22) Filed: Sep. 26, 2002

Related U.S. Application Data

(60) Provisional application No. 60/346,237, filed on Oct. 22, 2001. Provisional application No. 60/387,692, filed on Jun. 10, 2002. Provisional application No. 60/359,551, filed on Feb. 25, 2002.

(57) **ABSTRACT**
 Improved approaches for users with graphical user interfaces of computing devices are disclosed. A rotational user action supplied by a user via a user input device can provide accelerated scrolling. The accelerated nature of the scrolling enables users to scroll or traverse a lengthy data set (e.g., list of items) faster and with greater ease. The amount of acceleration provided can be performed in successive stages, and/or performed based on the speed of the rotational user action. In one embodiment, the rotational user action is transformed into linear action with respect to a graphical user interface. The resulting acceleration effect causes the linear action to be enhanced such that a lengthy data set is able to be rapidly traversed.



Description:

Scientific/technical text

Claims:

1. A method for scrolling through portions of a data set, said method comprising: receiving a number of units associated with a rotational user input; determining an acceleration factor pertaining to the rotational user input; modifying the number of units by the acceleration factor; determining a next portion of the data set based on the modified number of units; and presenting the next portion of the data set.
2. A method as recited in claim 1, wherein the data set pertains to a list of items, and the portions of the data set include one or more of the items.
3. A method as recited in claim 1, wherein the data set pertains to a media file, and the portions of the data set pertain to one or more sections of the media file.
4. A method as recited in claim 3, wherein the media file is an audio file.
5. A method as recited in claim 1, wherein the rotational user input is provided via a rotational input device. [...]

Background – Patent families

- Inventor desires protection in several jurisdictions (countries)
 - Submits patent application in several countries
 - Provides translations

The image displays four pages of patent documents for a patent family. The first page is the Chinese abstract (CN 20081006293.2), the second is the English specification (EP 1436), the third is the Korean abstract (KR 20081006293.2), and the fourth is the German abstract (DE 202 21 878 U1). The English page includes a detailed description of the 'Method and Apparatus for Accelerated Scrolling' and a flowchart (FIG. 1) illustrating the process of determining a scrolling factor based on user input and device characteristics.

They contain essentially the same text, in different languages

Listing 1: English description

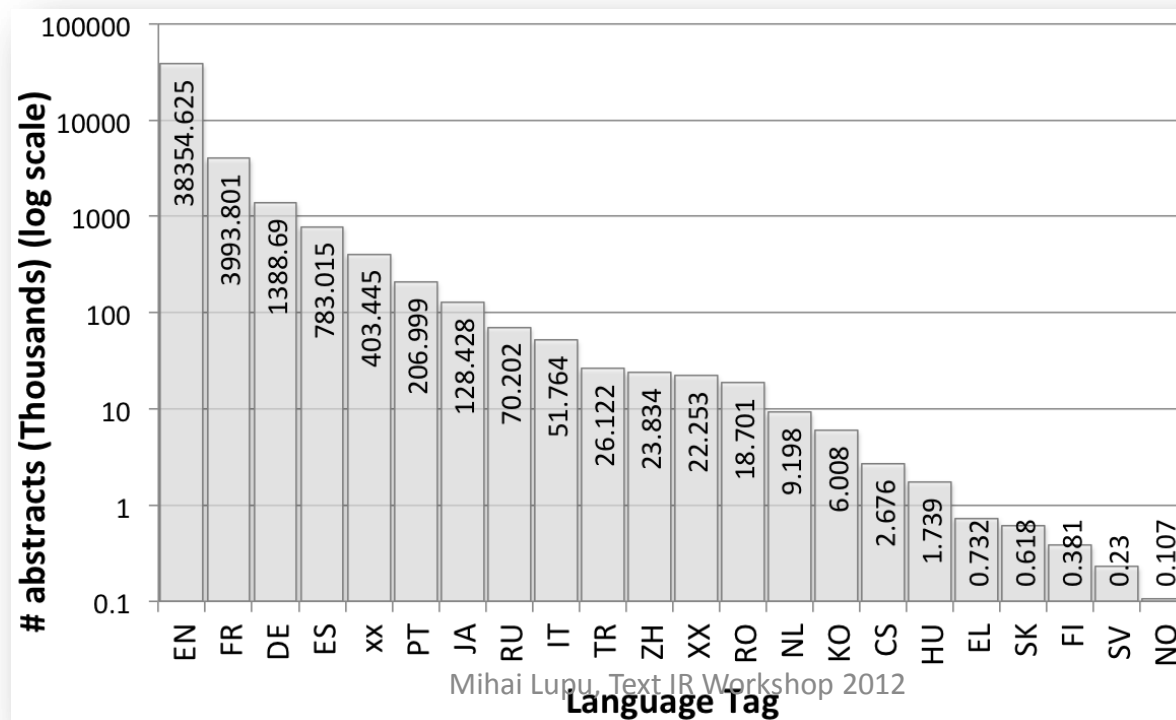
```
<description id="PDES17286951" lang="EN"
  load-source="patent-office">
  <heading>
    <b>Field of the invention</b>
  </heading>
  <p>The present invention relates to
    rotors for electromagnetic speed
    reducers, based on eddy currents (
    Foucault currents), such as those
    used in self-propelled vehicles,
    power test benches for internal
    combustion engines and other
    applications, built with ferritic
    nodular cast iron.</p>
  <heading>
    <b>Prior art</b>
  </heading>
  <p>One of the main problems concerning
    electromagnetic speed reducer rotors
    consists in their being subjected to
    strong heating because they have to
    convert the vehicle kinetic energy
    into heat under the effect of eddy
    currents. This means that the design
    of such rotors requires that their
    metallic mass be quite heavy, as a
    result of which the operation thereof
    leads to a high consumption of
    mechanical energy, with a negative
    effect on vehicle fuel consumption.</
  p>
</description>
```

Listing 2: Spanish description

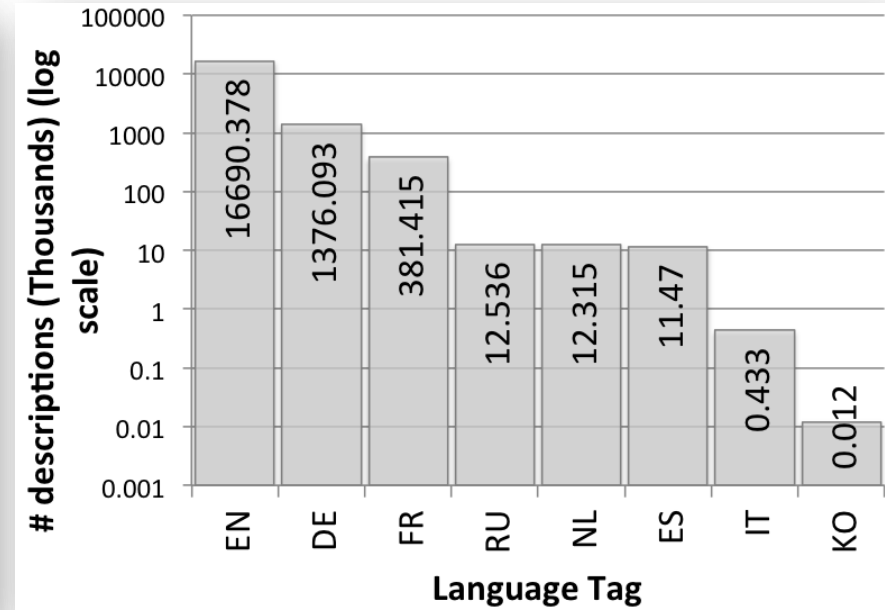
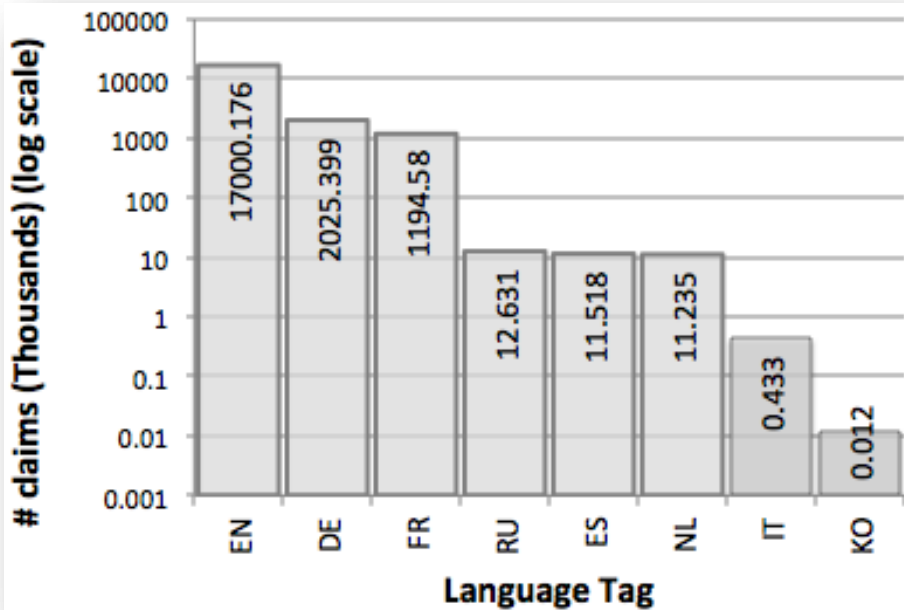
```
<description mxw-id="PDES139212" ref-ucid="
  WO-1994001592-A1" lang="ES" load-source
  ="patent-office">
  <!-- EPO <DP n="3"/>-->
  <p num="p0001"> D E S C R I P C I Ó N</p>
  <p num="p0002">ROTORES INDUCIDOS DE
    RALENTIZADORES ELECTROMAGNÉTICOS</p>
  <p num="p0003">FABRICADOS CON FUNDICIONES
    NODULARES FERRITICAS</p>
  <p num="p0004">Campo de la técnica</p>
  <p num="p0005">La presente invención se
    refiere a rotores de ralentizadores
    electromagnéticos basados en
    corrientes de Foucault, de los que se
    emplean en vehículos autoproñ
    pulsados, bancos de ensayo de
    potencia de motores de explosión y
    otras aplicaciones, fabricados
    mediante el empleo de fundiciones
    nodulares ferriticas.</p>
  <p num="p0006">Estado de la técnica</p>
  <p num="p0007">Uno de los principales
    problemas de los rotores de
    ralentizadores electromagnéticos
    consiste en que se encuentran
    sometidos a fuertes calentamientos
    debido a que deben convertir la energ
    ía cinética del vehículo en calor por
    efecto de las corrientes de Foucault
    . Ello provoca que el diseño de
    dichos rotores exija que su masa metá
    lica
```

Data

- “Alexandria” patent collection
 - Fairview Research (IFI Claims)
 - ~72million patent documents



Data



- More data available each year
 - 250k common applications between the five major patent offices (US, European, Japanese, Chinese, Korean)

Data

- Motivation – in numbers

	Abstracts			Claims		
	EN	FR	DE	EN	FR	DE
FR	3584	-	436	886	-	886
DE	994	436	-	886	886	-
ES	165	14	0	0	0	0
JA	121	121	0	0	0	0
RU	35	1	0	0	0	0
IT	10	0	0	0	0	0

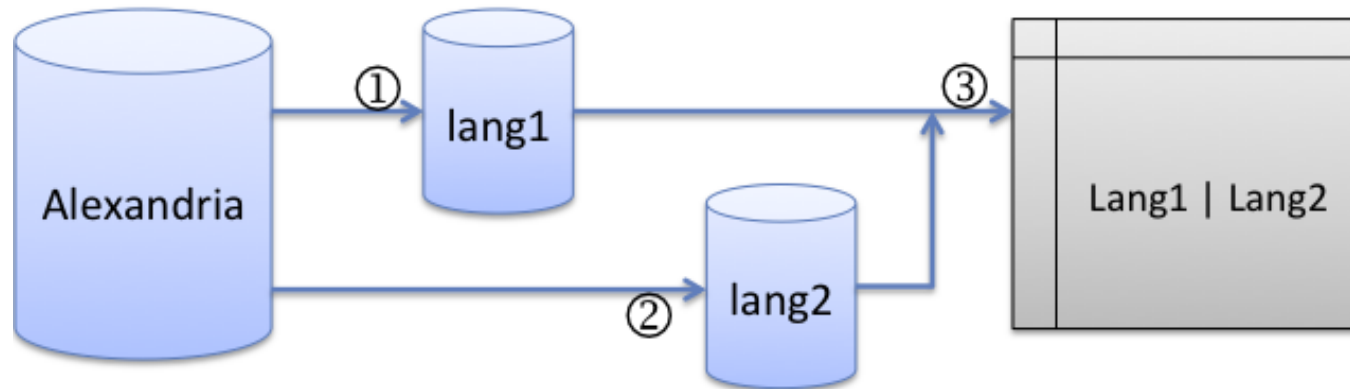
Table I

PARALLEL SECTIONS WITHIN THE CORPUS (THOUSANDS)

Method

- Idea
 - Based on family identifiers, select sections that belong to the same family but have different language tags
- In practice:
 - **Phase 1** – create a temporary candidate match table
 - **Phase 2** – match at paragraph level and clean up

Method – Phase 1

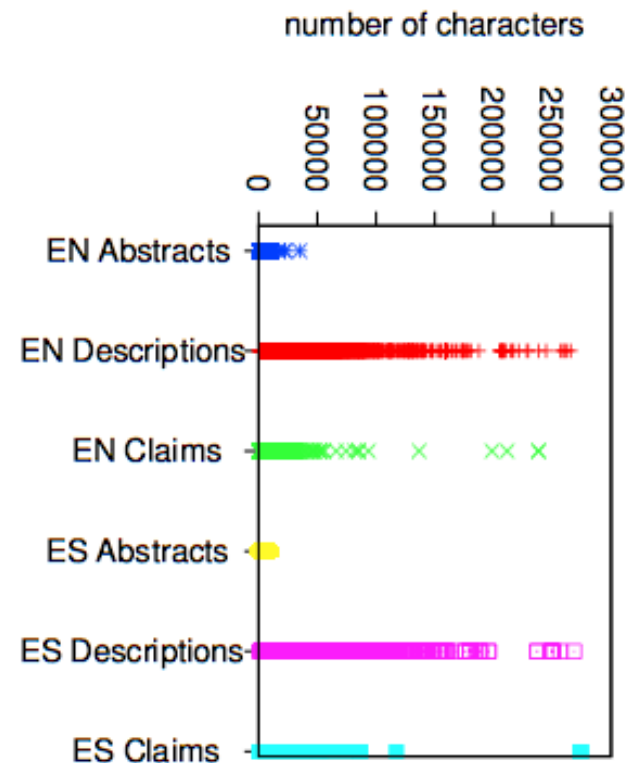


- ① Extract all sections with language tag *lang1* and populate a table with family IDs
- ② Repeat for *lang2*
- ③ Join on family identifier

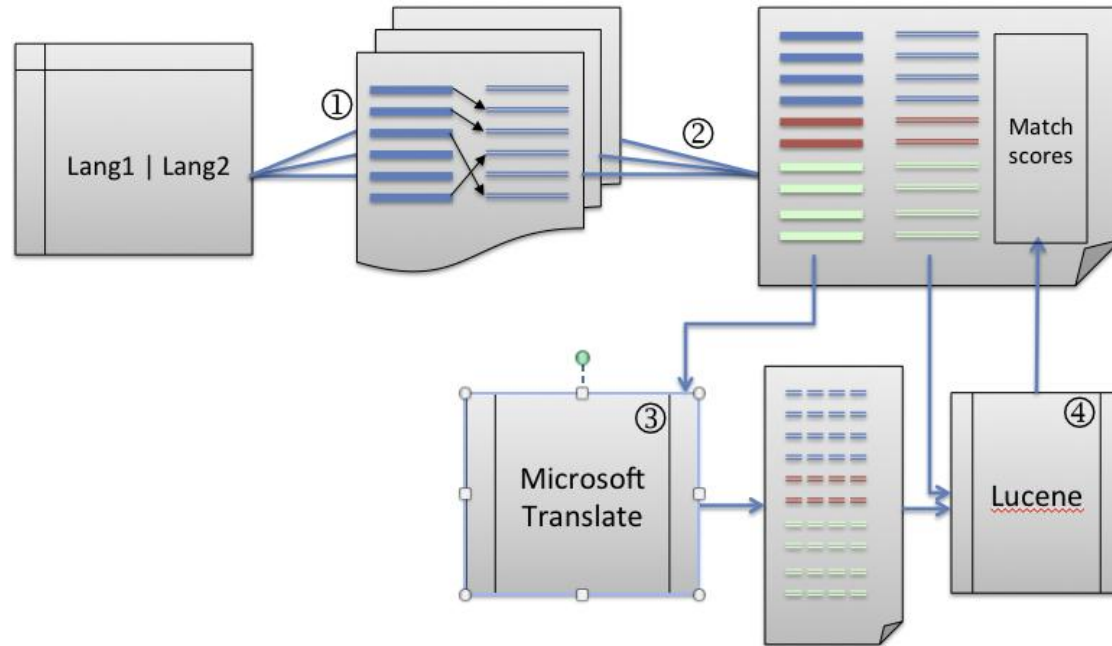
Note: due to the data architecture, ① and ② already involve a join

Method – Phase 1

- Results – Unusable in practice
 - Sections are too large
 - Descriptions/claims may change from patent office to patent office
- Needs
 - Split at paragraph level
 - Match & Clean



Method – Phase 2



- ① Match paragraphs from paired sections iff their size difference $< 20\%$
- ② Eliminate duplicate paragraphs
- ③ Translate from *lang1* to *lang2* using a generic MT
- ④ Index and match *lang2-translated* to *lang2* paragraphs

Test Case & Evaluation

- English – Spanish

- Numbers of sections, and pairs after Phase 1 (in thousands)


	Abstracts	Description	Claims
English	31788	10107	10417
Spanish	783	11	12
Pairs	2423	10	10

- Observations

- # English sections smaller than in the original data
 - Eliminated the result of previous MT
- # pairs > # Spanish abstracts
 - Existence, within same family, of multiple English abstracts

Test Case & Evaluation

- Selected 30 claim and 15 description pairs
 - ~1 million characters (limited to 2 million by the translation engine)
 - Abstracts not interesting because they are 1 paragraph only
 - Resulting paragraphs : 13,813

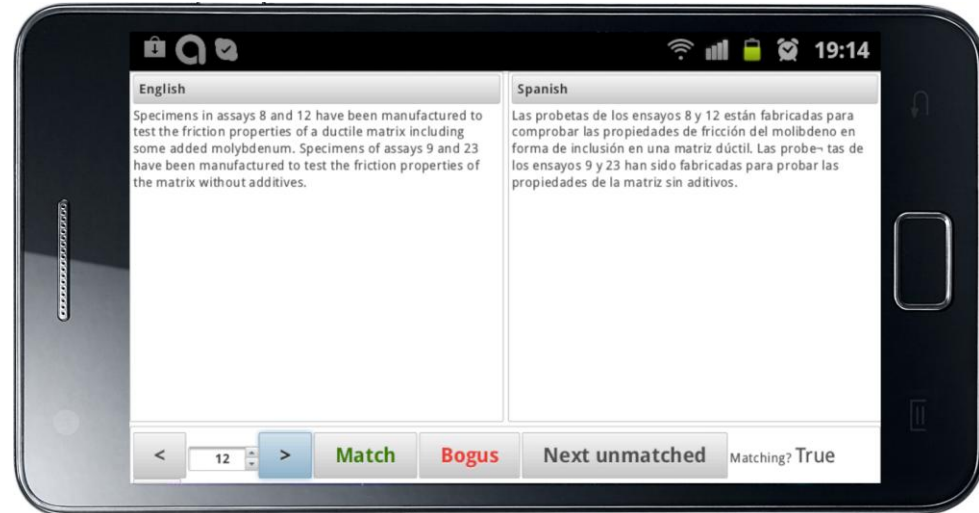
	Descriptions	Claims	Total	
English	656	520	1176	 Translated → Indexed
Spanish	811	325	1136	
Pairs	8919	4496	13813	

Test Case & Evaluation

- For each translated paragraph
 - *MoreLikeThis* to retrieve the most similar Spanish paragraphs
 - No filtering on documents
 - Inneficient, but necessary
 - A patent at one patent office may be split in two at another
 - Candidate filter:
 - Drop paragraphs which were the result of translation
 - Return a match only if within the top 10 most similar
- Result: 242 pairs (143 from descriptions, 99 from claims)

Test Case & Evaluation

- Manual evaluation
 - ~20seconds/pair
 - Patent domain specific
 - Stricter on claims



<http://ifs.tuwien.ac.at/~lupu/MultilingualPatentSections.zip>

	Abstracts	Description	Claims
Matching	115	69	184
Not Matching	28	30	58
Total	143	99	242

Error analysis

- 3 types:
 - Small paragraphs
 - Similar, yet with important differences
 - Erroneous paragraph segmentation in original data

Errors analysis – small paragraphs

Row	English	Spanish
1	Identification of T Cell Epitopes	(vi) DATOS DE SOLICITUD DE PRIORIDAD:
2	1.2 Prediction of T Cell Epitopes	(vi) DATOS DE SOLICITUD DE PRIORIDAD:
3	The model that is assumed for (t), or source model.	El modelo comprende los siguientes componentes:
4	Transgenic plants according to claim 45, wherein the plant is of the <u>Carica genus</u> .	53. Plantas transgénicas de la reivindicación 52, donde la planta es de la especie <u>Carica papaya</u> .
5	Transgenic plants according to claim 38, wherein the plant is a <u>dicotyledonous plant</u> .	53. Plantas transgénicas de la reivindicación 52, donde la planta es de la especie <u>Carica papaya</u> .

- Not enough data for similarity function
- 1 noun phrased changed

Errors analysis – similar, yet not

Row	English	Spanish
1	28. The process of , wherein the size of the largest particle of carbon black in the rubber is less than about 3 microns, and wherein the rubber has a ratio of tan at 0ř C. to tan at 80ř C. that exceeds about 2.0.	26. El hule cargado con negro de humo según la reivindicación 25, en donde el hule cargado con negro de humo tiene una proporción tan a 0řC/tan a 80řC que excede aproximadamente 1.5.
2	The recombinant DNA molecule according to claim 19, wherein the gene that codes for the enzyme that synthesizes organic acids' is a gene that codes for the enzyme <u>citrate synthase</u> .	26. La molécula de ADN recombinante de la reivindicación 19, donde el gen que codifica para la enzima que sintetiza ácidos orgánicos es un gen que codifica para la enzima <u>Malato Deshidrogenasa</u> .
3	A procedure for the production of a modified acrylic sheet with high impact resistance, in accordance with claim 24, characterized furthermore because the <u>demolding agent is added in quantities of 0.003% to 0.021% in weight with respect to the prepolymer, approximately.</u>	21.- Un procedimiento para la obtención de una 1 mina acrílica modificada de alta resistencia al impacto, de conformidad con al reivindicación 21, caracterizado además porque el prepolímero frío tiene un peso molecular en número aproximado de
4	3. <u>The method</u> according to , wherein said transmission of information occurs by use of a transmitter generating said binary sequences for spread spectrum applications by multiplying said Golay complementary sequences modulated by said amplitude values A, representing digital input, thereby multiplying a quantity of information bits per symbol interval by $m=\log A$	5. - <u>Un aparato</u> , según reivindicación 2, donde el generador de secuencias binarias para aplicaciones de espectro ensanchado está caracterizado por la posibilidad de multiplicar secuencias complementarias Golay moduladoras por A valores de amplitud que representan la información digital de entrada de modo que permite multiplicar por $z_n=\log A$, la cantidad de bits de información por intervalo de símbolo.
5	<u>Method for spread spectrum digital communication by Golay complementary sequence modulation</u> according to claim 1, which allows the transmission of information through a communication channel, comprising the generation of binary Golay sequences with low cross-correlation, which encodes the entry data, which in turn are amplitude modulated by means of amplitudes	2.- <u>El aparato</u> , según la reivindicación 1, que permite transmitir información a través de un canal de comunicaciones, que comprende la generación de secuencias binarias Golay de baja correlación cruzada, que codifican los datos de entrada, modulados a su vez en amplitud mediante A amplitudes, y que mediante modulación transmite dicha información al medio de transmisión.

Errors analysis – erroneous data

Row	English	Spanish
1	A device for producing bypasses under pressure in fluid piping systems according to , characterized in that it comprises at the upper portion of the radial conduit () a neck () or coupling with an inner threaded area () for coupling of the cutter () and an outer threaded area () for coupling of a cover () or a perforation tool.	REDES DE CONDUCCIÓN DE FLUIDOS, de conformidad con la reivindicación 1 , caracterizado porque comprende en la parte superior de la conducción radial (1) un cuello (8) o acoplamiento con una zona roscada interior (10) de acoplamiento de la fresa (3) y una zona (9) roscada exterior de acoplamiento de una tapa (2) o un útil de perforación. 4.- DISPOSITIVO PARA EFECTUAR DERIVACIONES BAJO PRESIÓN EN
2	. A device for producing bypasses under pressure in fluid piping systems according to , characterized in that in an embodiment alternative, it comprises a retention clip, catch or the like housed in the transverse hole of the male connector () of the cutter (), under the cover ().	21.- DISPOSITIVO PARA EFECTUAR DERIVACIONES BAJO PRESIÓN EN REDES DE CONDUCCIÓN DE FLUIDOS, de conformidad con la reivindicación 19, caracterizado porque el pasador (62) de relación del macho (61) y la embocadura (64) del eje (22) comprende una rosca o medio de bloqueo. 22.- DISPOSITIVO PARA EFECTUAR DERIVACIONES BAJO PRESIÓN EN
3	The most significant representatives in this family of TPEC are characterized by assuming that $\sigma = I$, where σ is a variance common to the sensors of the same type. The different linear solutions are characterized by the postulated as indicated below. Minimum Norm TPEC (MN) for which This solution requires that the configuration of generators has minimum energy () compatible with the likelihood. This method produces spatially widespread estimates that do not localize discrete sources correctly. Variants of this methodology were presented in the .	TCEP Mínima Norma (MN) en que $\sigma = j 1$. Esta solución exige que la configuración de generadores tenga mínima energía (Wang J.Z., Williamson S.J. and Kaufman L. Magnetic source images determined by lead-field analysis: the unique minimum-norm least squares estimation. IEEE Trans. Biomed. Eng., 39, 7, 665-667, 1992) compatible con la verosimilitud. Este método hace estimaciones dispersas que no localizan correctamente las fuentes discretas. Variantes de esta metodología fueron presentadas en la patente US 5,228,443.

Conclusions

- Cross-language retrieval important for patent search
 - Needs MT – needs more data
- A simple method to create large amounts of comparable data at paragraph level
 - An evaluation tool for quick result validation
- Errors more evident in claims data
- The result = input to a sentence alignment algorithm for MT training

Acknowledgements

