Finding the main themes in a Spanish document [*]

Adolfo Guzmán[1]
Centro de Investigación en Computación, Instituto Politécnico Nacional, México
aguzman@pollux.cenac.ipn.mx

SUMMARY. The computer can easily carry out many operations on systematic collections of data when these are numbers:

- What is this data about? What are its main topics?
- Make a summary. Obtain a summary of May sales of a given store.
- Compare. Compare May sales in stores A and B.
- Find similarities and discrepancies. How are sales of stores A and B similar?
- Find averages. Find the sales in the South of Mexico, in Fall 1997.
- Find tendencies. Extrapolate.

On the other hand, when data appears in documents in Spanish, organized in sections, paragraphs and sentences, it is not possible for the computer to carry out the above operations. Since much of human knowledge is in texts written in natural language, it is convenient to discover methods to carry out those operations. For that, the computer must **understand** or comprehend the text.

This paper shows how to analyze a document containing natural language sentences, in order to recognize its **main topics** or themes.

1. INTRODUCTION AND OBJECTIVES.

In the Center for Computing Research of the National Polytechnic Institute, the *Laboratory of Natural Language and Text Processing* works on intelligent text processing, needed to carry out the operations given in the summary. One of these operations, finding the main topics contained in a document written in Spanish, has been solved.

1.1 The importance of analyzing written Spanish.

Intelligent text analysis will allow the computer to *understand* documents written in natural language, for instance, to summarize them, to find tendencies, to compare two documents (with respect to a given theme) and to answer non trivial questions:

Having read this text	**Answer this question**
Frogs live in water	*Do frogs get wet?*
Benito Juárez is buried in San Fernando Cemetery.	*Where is the left big toe of Benito Juárez buried?*

[*] published in *Journal Expert Systems with Applications,* Vol. 14, No. 1/2, Jan/Feb 1988, pages 139-148
[1] This work was carried out by the autor at SoftwarePro International, owner of all rights on CLASITEX.

Another advantage: Natural language would be a new method to give orders to the computer: in Spanish, instead of using FORTRAN, C or Java.

Another advantage: natural language will become a new way for the computer to learn: it will only have to read books, the very books used by people. To teach it Physics, one would no longer need to write procedures or methods (and the corresponding objects), or data structures about concepts and formulas of that science.

1.2 Difference between word and concept.

One word can represent several concepts or meanings. Concepts, on the other hand (by definition), are unambiguous:

STAR (word)

- astronomic-star (concept), sidereal body;
- artist-star (concept), famous person;
- star-decoration (concept), adornment used by military people.

Wittgenstein said: "If words are ambiguous and represent several concepts, let us get rid of words and let us use only the concepts." Of course, such mandate can not be carried out literally. What we can do is to replace in a text, the words by the concepts such word represents. To select the appropriate concept (of the several possible concepts being denoted by the word), we need to *disambiguate.* This uses information about the context (of the document, or about the topics covered in the sentence). This is termed *semantic analysis*, since it pertains to the meaning of each word, sentence, paragraph, etc.

1.3 Concepts form a tree.

It is interesting to observe that concepts form families or hierarchies, as we can learn by browsing a Thesaurus. The relation most often used in these trees is "subset", where the "children" denote subsets, or specializations, or more particular or more specific concepts that the "parent." Let us take a concept denoting an object (noun, subject), for instance:

WEARABLE GARMENT (concept)
    SHOE
        sandal
        moccasin
        boot
    SHIRT
        T-shirt
        Long-sleeve shirt
    PANTS

In this example, the concept “moccasin” is a subset of the concept “shoe”, meaning that all moccasins are shoes, but not all shoes are moccasins. We have shown only a tiny fraction of the tree of concepts.

Now, let us observe another part of this tree, containing processes (actions, verbs). Let us concentrate in changes of position:

MOVE
- *CHANGE OF (X, Y) POSITION (of coordinates in Earth)*
  - Float
  - Swim
  - Walk
    - *run*
    - *jump*
  - Fly
    - *soar*
- *ROTATE (change orientation)*
- *VIBRATE*

CHANGE OF SIZE
- *INFLATE-EXPAND*
- *CONTRACT-GET SMALLER*

CHANGE OF HEALTH
- *GET WELL*
- *GET SICK*
- *DIE*

CHANGE OF ECONOMIC STATUS
- *GET POOR*
- *GET RICH*

1.4 The number of concepts is finite.

In particular, the number of common knowledge (or common sense, see next section) concepts is finite [Ref. 1]. Thus, it is possible to form the concepts tree, or some parts of it. That is the basis of the algorithm and program, named CLASITEX, that finds the themes in a document.

2. COMMON KNOWLEDGE.

It is the knowledge common to all people. It is the knowledge that an specialist has when his specialty knowledge is removed. Thus, it is the knowledge that has a child who is, say, 9 years old, or somebody who has not learnt an specialty [Ref. 1]. It is also called, for obvious reasons, common sense.

2.1 The tree of common knowledge.

As we saw in §1.3, it is possible to organize knowledge, including common knowledge, into a tree. It is estimated that this tree has between one and ten million concepts (nodes). Project CYC [Ref. 1] was devoted to the huge task of building this tree, by hand, introducing concept after concept (plus the relations among them). It lasted ten years, but it did not reach its objective. In spite of this, its goal is valid, and such failure should not deter other researchers; sooner or later the tree will be built.

This tree contains the concepts (objects or subjects, transformations, relations, and so on) that people (such as a 9 years old child) have acquired through their experience on Earth and their relation with other human beings. It does not contain specialized concepts, from Chemistry or Rheology, say. *The concepts in the tree are not known, before hand, to the computer*. But they are needed to understand natural language texts. It makes sense, thus, to try to introduce them to the computer.

The common sense tree has, in addition to "subset", other relations (such as "part of", "formed by", and others) between two concepts. The relations themselves form a tree and, as such, are part of the knowledge tree. As an example we show RELATIVE OF.

RELATIVE OF
- *FATHER OF*
- *SON OF*
  - First son of
  - Preferred son of
- *BROTHER OF*

3. FINDING THE TOPICS OF A DOCUMENT, FIRST ATTEMPT.

To begin, we can count the words in the paper. Those appearing more often denote, without doubt, important themes. For instance, if the word "revolution" appears 17 times, the paper talks about revolution.

3.1 Problem: many words refer to the same topic.

For instance, a document only talks about **revolution** twice, but it mentions Zapata, Francisco Villa, Flores Magón, ...

For instance, first inning, pitcher, El Toro Valenzuela, Red Sox, all speak of (suggest the topic) base ball.

This suggests to build the concept tree, and to replace the words in the document by the respective concepts. Once this is done, let us find the most popular concepts: those are the principal topics of the paper. When counting concepts, we must count not only the "immediate"

concepts denoted by a word (for instance, the word "moccasin" represents the concept "moccasin-shoe"), but also concepts higher up in the tree (towards its root), which in the example shall be "shoe", "wearable garment", etc. The method seeks to mark concepts and to count them over the tree. At the end of the analysis of the document, some parts of the tree will contain more marks (counters with higher counts, blacker regions), and those parts of the tree are the topics of the document.

Problems in the proposed method:

- The tree has to be built by hand (This is precisely how we do it, manually).

- Replacing a word by the concept it denotes is not easy, since it is necessary to disambiguate (Cf. §1.2). If we did not do it, a word would increment the counter of each of the concepts it represents. (This is solved by CLASITEX by skipping the disambiguation, which, as we will see below, does not appreciably increases the imprecision).

3.2 Summary of proposed procedure.

Find the concepts that appear more often – those are the topics. The following observations apply:

1. Count concepts, not words.
   Refinement: Count concepts over the concept tree.
   Count concepts and their generalizations (more general concepts).
   - When "moccasin" is found, count also shoe, wearable garment, ... The current version of CLASITEX does not count generalizations; nevertheless, its results are acceptable. ♣
   - Some parts of the tree will become blacker than others. The darker regions show the **topics** of the document or paper.

   Accuracy increases with length of document.

2. Each word "votes" for (increases the counter of, blackens the node in the concept tree) the concepts it denotes.
   When we find **channel** we add 1 to the counter of concept "water-channel", and 1 too to the counter of concept "communications-channel."

3. *Most votes will be cast wrong.* Nevertheless, in the long run, correct votes will prevail.

   A) t is possible, too, to disambiguate in a second step: words that voted more than once, shall withdraw all their votes except one (they keep the "correct" sense), once the principal themes are found at the end of first step. ♣[2] These votes withdrawn should not change (except in rare occasions) the main themes already found in the first step.

[2] With this symbol we mark suggestions and extensions for future work, or possible improvements to CLASITEX.

B) In practice, we have seen that this second step is not necessary: the very accumulation of (counters of) topics, both correct and incorrect, distinguishes the correct topics. It is as if, instead of getting rid of the noise in a signal, we accumulate n such signals, causing the noise to diminish like square root of n.

3.3 There are concepts denoted by a sequence of more than one Spanish words.

For instance,

COCINA (one word) – denotes the concept “cocina-kitchen”, a room where food is produced.

NUEVO MÉXICO (two words). Denotes the concept “Nuevo Mexico-New Mexico”, an state belonging to U. S.

Benito Juárez (two words). Denotes a distinguished Mexican president.

Los Tres Mosqueteros (three words). Denotes a book written by Alexander Dumas.

Instituto Politécnico Nacional (three words). Denotes the National Polytechnic Institute, a famous Mexican technical university.

Adolfo Guzmán Arenas (three words). Denotes the author of this paper.

Universidad Nacional Autónoma de México (four words). Denotes UNAM, a Mexican University.

This implies that, when scanning the text, we should look not only for single words, but also for couples, triplets, ..., of words, finding if each of these sequences denotes (or not) some concept. CLASITEX now uses sequences of up to fourd words. It could be improved. ♣

3.4 The tree used in CLASITEX.

CLASITEX uses a tree that employs, instead of the “subset” relation between its nodes (Cf. §1.3), the relation “suggests the topic”, or “votes for the topic...”. For instance, “Benito Juárez” suggests the topics “Mexico”, “Oaxaca”, and “Mexican History”, even when Benito Juárez is a person, México is a country, and Oaxaca is a state.

3.4.1 Meaningless words.

This simplification brings another consequence. One word such as “tres” (which represents the concept “3”, an odd number), or such as “entonces” (“therefore”), does not suggest any topic, and does not vote for any topic. We think of these words (somewhat inappropriately) as having “weak sense”, or being “meaningless”, signifying that they do not contribute to ascertain the topic of a given document.

4. PROCEDURE USED IN CLASITEX.

We describe the algorithm used by CLASITEX.

We travel with a pointer the Spanish document (residing in a file), from left to right.

1. (BEGIN) Observe the sequence of four words pointed at by the pointer. Do they denote some concept(s)?
   Yes: increase by 1 the counters of each of these concepts. Go to step (5).
   No: Go to next step (2).
2. Observe the sequence or chain of three words pointed at by the pointer. Do they denote some concept(s)?
   Yes: increase by 1 the counters of each of these concepts. Go to step (5).
   No: Go to next step (3).
3. Do same for the sequence of two words. Go to step (5) or next step (4).
4. Observe the word pointed at. Does it denote some concept(s)?
   Yes: increase by 1 the counters of each of these concepts. Go to step (5).
   No: Is it a word denoting no concept? That is, is it a meaningless word, according to §3.4.1?
      Yes: ignore it. Go to step (5) next.
      No: Print it as "I do not know what this word means, or if it is meaningless" (One has to add these meanings, later, by hand, to the CLASITEX tree). Go to step (5) next.
5. Move the pointer to the right of the word(s) already analyzed. And repeat the operation or iteration: go to step (1) (BEGIN).When all text has been analyzed (we can no longer move the pointer to the right), report the more popular topics as the themes of the document.

In practice, we have seen that the concepts histogram contains a well marked separation between popular and rare topics. We will see some examples below.

4.1 Examples of concepts represented by one word.

CLASITEX uses directory ARBOL1 ("tree1") to store those concepts represented by one word (to be more exact, the concepts "suggested" or "voted for" one word). Each concept is a file, and inside the file we find the words (single words in case of ARBOL1; sequences of two words in case of ARBOL2, of tree words in case of ARBOL3, etc.). A partial listing of the ARBOL1 directory follows:

accidente	encuesta	musica
acertijo	enfermedad	narcotrafico
acto-juridico	enloquecer	nayarit
africa	entregar	nino
agencia-noticia	envase-recipiente	noche
agricultura	error	noruega
aguascalientes	escandalo	noticia-internacional
aislado	escoger	oaxaca
alemania	escritura	oficina

amenaza
amor-pasion
animal
anunciar-promulgar
apicultura
arma-de-fuego
asesorar
astronomia
atraso
auto-parte
autom-marca
autor-libro
autoritarismo
bahamas
bajacalifornia
banarse
banco-de-mexico
bebida-alcoholica
beisbol
belgica
bolsa-valores-mex
buceo
budismo
cambio
campeche
campo-rural
cantar
carretera
casa-parte
casamiento
cereal
chiapas
chihuahua
chile
ciencia
cirugia
ciudad
ciudad-mexico
cnsp
colima
comer
comercio
compositor-musica
compositor-musica-mex
computadora
concept1
congreso
constitucion
consumo
convalescencia
corea
correr
corrupcion
ctm
escuela-colegio
espana
estado-de-mexico
excelsior
experto
exponer-exhibir
familia
famoso
ferrocarril
fiesta
filosofia
finanzas
flor
francia
fruta
futbol-soccer
geografia
gerencia
gobie-actual
gobie-municipio
gobierno
guanajuato
guerra
guerrero
guerrero-edo
habitacion
hacer-fabricar
hecho-violento
herr-carpintero
herr-fierro
herr-fontanero
herr-jardinero
hidalgo
hipoteca
hombre
honestidad
hospital
hueso-humano
iglesia
imss
indus-petrolera
industria
inglaterra
inmigrante
institucion
inteligencia-seguridad
intervencion-meter
inver-dinero
inver-extranjera
iraq
irlanda
italia
jalisco
jardin
pacto-arreglo
pais
pan-bizcocho
pan-partido
papa-catolico
pasado
paz
pedir
pelear
periodico
perro
persona-vieja
platillo-comida
poder-judicial
poder-legislativo
poeta
policia
politecnico
politica
postre
pps-partido
prd-partido
premio
prenada
preocupacion
pri
pri-partido
problema-crisis
profesion
protestantismo
psicologia
pt-partido
queretaro
quintanaroo
radiodifusion
reino
relig-catolica
relig-cristiana
religion
renunciar
reproducir
restaurante
ritmo-cadencia
roma
rusia
salud
sanluispotosi
secr-comercio-mex
secr-hacienda-mex
secr-tesoro-eeuu
secuencia
sexual
sinaloa
sobrevivir

cuerpo-humano	leguminosa	sociedad
cultura	lenguaje-calo	softwarepro
delincuencia	lentejuela	soga-hilo
democracia	ley	sonreir
deporte	libro-revista	substancia-quimica
desfile	llorar	suburbio
desgarrar-romper	lluvia	sufrimiento
deuda-prestamo	lugar-en-ciudad	tabasco
devaluacion	magisterio	tamaulipas
dinero	maletin-veliz	taurino
dirigente	mar	tecnologia-ingenieria
discriminacion	marcha-protesta	telefono
discurso	marxismo	television
discutir	matematicas	tienda
disgusto	medicamento	tlaxcala
divorcio	mexico	trabajo
dormir	michoacan	trafico
dulce-caramelo	miedo	triunfar
durango	militar	unidad-medida
ecologia	miseria-pobreza	utensilio-cocina
economia	misterio	venezuela
eeuu	mito-creencia	veracruz
el-financiero	morelos	vestimenta
eleccion-votar	mueble	viaje
empleado	muerte	viejo-antiguo
empleo-puesto	mujer	yucatan
empresa	mundo	zacatecas

It can be observed that (due to use of UNIX Santa Cruz Operation) we did not use accents, nor the letter ñ. ♣

4.1.1 Examples of files representing a concept.

In CLASITEX, a node of the tree (a concept) is represented by a file containing single, double, triple or quadruple words (we are not handling "meaningful phrases" of more than 4 words [§3.3], for the time being ♣[3]), these being the words (or sequences) that "vote for" or "refer to the topic" or "refer to" the concept that the file denotes (that is, to the *name* of the file). For instance, the file **acto-jurídico** (jurisprudential act, a concept), contains the following single words:

apelacion. audiencia. condena. condenada. condenadas. condenado. condenados. culpable. delictuosa. delictuosas. delictuoso. delictuosos. Delito. embargo. juicio. procesado. procesados. procesamiento. procesar. proceso. sentencia. sentenciado. sobresellar.

Another node of the common sense tree, CLASITEX version, is the concept **platillo-comida** (food-dish), represented by a file named platillo-comida containing the following single *words:*

albondiga. albondigas. alimentar. alimenticia. alimenticias. alimenticio. alimenticios. alimento. arroz. birria. caldo. ceviche. comer. comida. degustar. enchiladas. flautas. freido. frito. guisado. guiso. menudo. mole. mondongo. paladear. pambazo. pambazos. pancita. picante. picosa. picoso.

[3] With this symbol we mark suggestions and extensions for future work, or possible improvements to CLASITEX.

platillo. platillos. plato. pozole. pulpo. quesadilla. quesadillas. saborear. sabrosa. sabroso. servilleta. servilletas. sope. sopes. taco. tacos. tampiquena. taqueando. taquear. taquiza. tortilla. tortillas. tortillera. tortillero. tostada. tostadas. vianda.

Here we can observe that the current version of CLASITEX lacks a dismemberer to decompose a word in its root and suffixes (and prefixes), and for this reason we have to file words such as alimentar, alimenticias, alimentos, etc., when with only the root "aliment-" would be enough. ♣
Next example denotes the concept **platillo-comida**, and shows the contents of its file. The contents are pairs of words that "vote" for the concept. This file lies in the directory ARBOL2 (tree2), which contains the concepts (files) or nodes that, in turn, contain (are represented or voted for by) couples of words.

agua fresca. aguas frescas. carne asada. pollo rostizado. pollos rostizados.

The concept **casamiento** (marriage), and its double words voting for it:
media naranja. otra mitad. recien casada. recien casado. recien casados.

The concept **utensilio-de-cocina** (kitchen-tool), and its double words voting for it:
bano maria. plato hondo. plato sopero. plato tendido. vaso refresquero.

The concept **platillo-comida** (food-dish), and its triple words voting for it:
chilpachole de jaiba. sopa de medula. sopa de pescado. sopa de tortilla.

The concept **casamiento** (marriage), and the sequences of three words suggesting it as a topic:
unir sus destinos. vestida de blanco.

The concept **computadora** (computer), and the 3-word sequences voting for it:
arquitectura de computadoras. base de datos. bases de datos. convertidor analogico digital. convertidor digital analogico. convertidores analogico digital. convertidores analogico digitales. convertidores digital analogico. convertidores digital analogicos. editor de texto. editores de texto. estructura de computadoras. hoja de calculo. hojas de calculo. programa de computo. programa de graficacion. programas de computo. sistema de informacion. sistemas de informacion.

The concept **miedo-pavor** (fear-panic), and the 3-word sequences suggesting it:
lleno de pavor. lleno de temor. presa del panico. presas del panico. preso de panico. preso de temor. preso del panico. presos del panico.

4.2 Weak sense or meaningless words, voting for no concept.

These words are ignored, and they do not contribute to any concept counter. They are stored in a file that means "these words shall be ignored, since they denote no concept. Their meaning is weak".

a. abajo.	abalance.	abalanzaban.	abalanzado.	abalanzados.	abalanzamos.
abalanzando.	abalanzar.	abalanzaran.	abalanzaras.	abalanzare.	abalanzariamos.
abalanzarse.	abalanzas.	abalanzaste.	abalanzo.	abaldonado.	abaldonar.
abalizada.	abalizadas.	abalizado.	abalizados.	abalizar.	abandona.
abandonaba.	abandonabamos.	abandonaban.	abandonada.	abandonadas.	abandonado.
abandonados.	abandonamiento.		abandonamos.	abandonando.	abandonar.
abandonaras.	abandonare.	abandonaria.	abandonas.	abandonaste.	abandono.
abate.	abatia.	abatiamos.	abatida.	abatidas.	abatido.
abatidos.	abatimiento.	abatimientos.	abatir.	abatiras.	abatire.
abatiremos.	abierta.	abiertas.	abierto.	abiertos.	abria.
abriendo.	abril.	abriles.	abrimos.	abrio.	abrir.

abriremos.	abriria.	absurda.	absurdamente.	absurdas.	absurdo.
absurdos.	acababan.	acabada.	acabadas.	acabado.	acabados.
acabamos.	acaban.	acabando.	acabar.	acabara.	acabaria.
acabarian.	acabarias.	acabaron.	acabaste.	acabe.	acabemos.
acabes.	acabo.	acatada. acatadas.		acatado.	acatados.
acatamiento.	acatamientos.	acatando.	acatar.	accesible.	acceso.
acepta.	aceptaban.	aceptacion.	aceptada.	aceptamos.	aceptan.
aceptar.	aceptaran.	aceptasen.	acepto.	acerca.	acercaba.
acercada.	acercadas.	acercado.	acercados.	acercamiento.	acercamos.
acercando.	acercandose.	acercar.	acercara.	acercarse.	acerquemos.

(Many more words that begin with a... and continue up to z).

4.2.1 Tree of weak sense concepts, represented by a single word.

This tree (a directory) contains files (concepts), which in turn contain single meaningless words. Contrasting with the file of §4.2, we use here a directory, in order to allow a possible transfer of these concepts from "meaningless" to "with meaning", that is, to arbol1. ♣

abajo	colores	feo-horror	llevar-a-cabo	rechazo
abandonado	constituir	frio-caliente	lograr	renovar
afirmar	dia-mes	futuro	marcar	repeticion
afrenta	diferente	hablar	narrar	total
ahora	dificultad	igualar	nunca-siempre	unir
eliminar	ilimitado	obedecer		
brilloso	enunciar	impedir	obtener	
caminar	felicidad	ir-acudir	precipitar	

One of these files, **caminar** (to walk, a "weak" concept), contains the following single words:

andaban. andado.	andamos.	andando.	andanza.	andanzas.	andar.	anduve.
anduvieron.	anduviste.	caminaba.	caminamos.	caminan.		
caminando.	caminante.	caminar.	caminas.	camine.		
caminemos. camino.	recorrer.	recorreran.	recorri.	recorria.	recorrido.	
recorriendo.	recorrimos.	recorriste.	trasladaba.	trasladaban.	trasladada.	
trasladado.	trasladados.	trasladando.	trasladar.	trasladare.	trasladaren.	
trasladarian.	trasladarse.	traslade.	traslado.			

The file **felicidad** (happiness, a concept now considered with weak meaning) contains the following single words:

agradable.	agradablemente.	agradables.	agrado.	agrados.	alegria.	alegrias.
felices.	felicidad.	felicitacion.	felicitar.	felicito.	feliz.	
satisfaccion.	satisfacciones.	sentimiento.				

4.2.2 Pairs of words representing no meaning.

desde abajo.	hasta abajo.	luego entonces.	no obstante.	por abajo.	por consiguiente. por debajo.
sin barreras.	sin embargo.	sin limite.	sin limites.	sin obstaculo.	sin obstaculos.

5. RESULTS.

5.1 OPERAN AL PAPA (Surgery to the Pope, text analyzed by CLASITEX).

Operaron al papa; largo, el periodo de rehabilitacion.
Recibio una protesis que reemplaza la cabeza del femur.
Se resbalo, insisten.

Ciudad del Vaticano, 29 de abril. (AP, AFP, EFE y ANSA). El papa Juan Pablo II fue sometido hoy a una intervencion quirurgica de "artroprotesis" tras haberse fracturado una pierna al caerse en el bano, y los medicos dijeron que tendra varias semanas de recuperacion y luego un largo periodo de rehabilitacion.

Una declaracion de la Policlinica Gemelli en Roma dijo que la intervencion quirurgica, que segun las tecnicas modernas se considera de rutina, "estuvo muy bien." El Papa estaba nuevamente consciente y fue llevado de regreso a su habitacion del decimo piso.

El Pontifice, que tiene 73 anos, recibio una protesis para reemplazar el hueso en la zona en que el femur se articula con la cadera. La operacion se inicio alrededor de las 13:30 horas (11:30 GMT).

El papa se fracturo la parte superior del femur derecho y se disloco la cadera al salir de la ducha, mojado y descalzo, y resbalar en el piso de su habitacion, el jueves por la noche. Fue atendido primero por su medico privado, Renato Buzzonetti, y durmio algunas horas con analgesicos.

Pero esta manana fue trasladado al hospital Gemelli, la sexta vez que lo visita como paciente.

Ante la polemica sobre por que fue ingresado el Papa 11 horas despues de caer, el Vaticano senalo que el medico personal de Juan Pablo II estimo que no era necesario trasladarlo de manera urgente al hospital romano.

El vocero del Vaticano, Joaquin Navarro, dijo que la caida no fue provocada por perdida del conocimiento ni por ninguna enfermedad.

Agrego que a Juan Pablo II se le ha colocado una protesis de cierta aleacion metalica que se ha introducido en el resto del femur y cuya implantacion ha sido perfecta gracias a la robustez del hueso propio de un hombre habituado a practicar deportes.

Asimismo, se le ha aplicado inmovilizacion interna mediante una pieza metalica, que ha hecho innecesaria la colocacion de una escayola externa, preciso Navarro.

Antes de la intervencion y cuando ya acusaba el dolor de la fractura, el Papa exclamo, segun dijo Navarro: "Tal vez faltaba esto para el Ano de la Familia".

Hace dos dias, durante la audiencia publica de los miercoles, el Papa hablo precisamente del valor redentor del sufrimiento humano.

El Papa dijo a los medicos que lo recibieron en el hospital: "Tienen que admirar mi fidelidad a la Catolica", es decir, la Universidad, de la cual depende la Policlinica Gemelli.

El director del hospital, Emilio Tresalti, describio al Papa como "un hombre muy fuerte" cuyo estado general es "excelente".

Tresalti agrego: les digo solamente que mi madre, que tiene la misma edad de Juan Pablo II, fue operada de la misma cuestion, por el mismo equipo, el ano pasado, y camina normalmente.

El boletin medico entregado a la prensa dijo que la intervencion quirurgica a la que fue sometido durante dos horas el Papa Juan Pablo II fue normal, habiendose mantenido "los parametros bioquimicos y funcionales dentro de los limites normales".

Despues de la anestesia, el Papa "volvio en si de modo tranquilo y normal. Sus condiciones son optimas", agrego el boletin, que indico ademas que Juan Pablo II permanecera internado "dos o tres semanas", salvo complicaciones.

Durante la intervencion perdio sangre pero esta misma sangre le fue introducida otra vez en el sistema circulatorio, no siendo necesaria ninguna transfusion externa, preciso el portavoz vaticano.

A los pocos meses del atentado de 1981, ocasion en la que le fueron practicadas varias transfusiones, el Papa padecio una enfermedad provocada por un virus probablemente transmitido por la sangre externa recibida.

El profesor Gianfranco Fineschi, jefe del equipo que opero al Pontifice, aseguro que el Papa se recuperara para hacer una vida normal, aunque no podra volver a esquiar "si podra realizar paseos por la montana y nadar, que es la mejor terapia", manifesto.

La protesis "funcionara" aunque "la cadera del Santo Padre no sera ya como Dios la ha hecho, sino como la ha hecho un bioingeniero", comento Fineschi.

La protesis que se ha colocado al Papa, indico, es de larguisima duracion y muy costosa.

En ese momento, el Pontifice se encontraba acompanado del secretario del Estado Vaticano, cardenal Angel Sonado; el sustituto de la Secretaria, Giambattista Re; su secretario privado, el sacerdote polaco Stanislao Dziwisz, los doctores Fineschi y Buzzonetti, y dos religiosas que otras veces lo han atendido.

Los medicos dijeron que, luego de las dos o tres semanas en cama, el Papa tendra una rehabilitacion de algunos meses, tal vez con ayuda de un baston.

Pero luego podra dejar el baston, agrego un cirujano.

El Director del hospital dijo que habia conversado con el Papa y no mostro signos de tension. Estaba solo preocupado por haber debido renunciar a su programado viaje a Sicilia durante este fin de semana.

El papa celebro misa -sentado-, esta manana, en su apartamento privado, antes de abandonar el Vaticano para ser intervenido en la Policlinica Gemelli.

Navarro dijo que el Papa cancelaria los viajes y reuniones previstos para las proximas semanas.

El Papa se estaba preparando para un viaje a Sicilia, donde pronunciaria una alocucion contra la mafia y tenia una nutrida agenda para el mes proximo.

Estaba previsto que presidiera la clausura de un sinodo de obispos africanos, encabezara una conferencia de cardenales sobre los preparativos para la celebracion del ano 2000 y partiera el 13 de mayo para una visita de tres dias a Belgica. Tambien tenia programada una reunion con el canciller aleman Helmut Kohl.

El 18 de mayo el Papa cumplira 74 anos.

En junio, segun su agenda, tenia previsto entrevistarse con el presidente Bill Clinton, en el Vaticano.

Desde su cama, Juan Pablo continuara guiando a la Iglesia catolica, sin delegar la herencia de Pedro, asistido por el cardenal Sodano.

Aun convaleciente, podra firmar, como lo hizo en otras ocasiones, decretos urgentes.

Es perfectamente consciente y capaz de hacerlo, subrayo Navarro, quien sonriendo, anadio que no hay necesidad, ni estan previstas "regencias".

"Aqui, entre nosotros, no existe el problema de algun maletin negro con los codigos secretos", anadio Navarro.

El papa recibio los buenos augurios de personalidades de todas partes, el nuevo primer ministro de Italia, Silvio Berlusconi; el presidente italiano, Oscar Luigi Scalfaro; el rey de Espana Juan Carlos y los monjes budistas del Dalai Lama, entre otros.

Scalfaro senalo en su mensaje al Papa que "mientras, en el misterio amoroso de la Providencia, otro sufrimiento le es solicitado como oferta para la humanidad tan confusa y doliente, le envio mi augurio de rapida recuperacion, que es el augurio de todo el pueblo italiano."

El practicamente ex premier italiano, Carlo Azeglio Ciampi, los presidentes del Senado, Carlo Scognamiglio de la Camara de Diputados, Irene Pivetti, enviaron tambien telegramas en los cuales manifiestan su dolor por el accidente de Juan Pablo II y la esperanza de un restablecimiento rapido.

Al hospital empezaron a llegar numerosas monjas para rezar por la salud del jefe maximo de la Iglesia catolica.

Una de las primeras personas que llego a la policlinica fue la madre Elizabeth Patrizi, de la orden fundada por el martir de Auschwitz, con un gran ramo de flores.

Desde que fue elegido pontifice en 1978, Juan Pablo II ha pasado 106 dias en la Policlinica Gemelli, adonde ha sido intervenido seis veces.

La primera fue el 13 de mayo de 1981, a causa del atentado en la plaza de San Pedro.

La segunda, pocas semanas despues de salir por el atentado, a causa de una enfermedad viral.

En julio de 1992, volvio al hospital para ser operado de un tumor benigno del intestino.

En los meses sucesivos volvio para varios controles.

En diciembre de 1993 debio ser internado por una luxacion del hombro derecho causada por una caida durante una audiencia.

5.1.1 Results for “Operan al Papa”.

In file “Cuen-xxx” (count-xxx), CLASITEX leaves the counts of the concepts:

papa-catolico *(catholic Pope)*: 49 cirugia *(surgery)*: 28 familia *(family)*: 26
enfermedad *(sickness)*: 17 accidente *(accident)*: 16
relig-catolica: 15 convalescencia: 11 hospital: 10
anunciar-promulgar: 8 medicamento: 8 cuerpo-humano: 7 dirigente: 7 lugar-en-ciudad: 7 hecho-violento: 6 hueso-humano: 6 roma: 6 habitacion: 5 sufrimiento: 5
banarse: 4 casa-parte: 3 deporte: 3 economia: 3 empleo-puesto: 3
imss: 3 intervencion-meter: 3 acto-juridico: 2 alemania: 2
bebida-alcoholica: 2 budismo: 2 ciudad-del-vaticano: 2 dormir: 2
empresa: 2 escuela-colegio: 2 espana: 2 exponer-exhibir: 2
fiesta: 2 gobie-actual: 2 hombre: 2 ley: 2 mueble: 2
noche: 2 religion: 2 sexual: 2 viaje: 2 africa: 1
agencia-noticia: 1 aleacion-metalica: 1 amor-pasion: 1 belgica: 1
computadora: 1 discutir: 1 eeuu: 1 eleccion-votar: 1 escritura: 1
flor: 1 gobierno: 1 jardin: 1 libro-revista: 1 maletin-veliz: 1
misterio: 1 muerte: 1 noticia-internacional: 1 paz: 1

periodico: 1 pierna-humana: 1 poder-legislativo: 1 renunciar: 1
salud: 1 secuencia: 1 sonreir: 1 tecnologia-ingenieria: 1

As we see, the first two or three topics have a salient count, and *they are* the themes of the document. The remainder have much smaller counts; these counters may be correct, or may be due to incorrect votes when a word votes for the several concepts it denotes. Observe that *family* with 26 counts was (incorrectly) found as a main thopic, the reason is that "papa *(Pope)*" is written the same as "papá *(father)*", which votes for *family*. The mistake will be solved when we introduce accents (remarked in §4.1). Also observe that ciudad-del-vaticano (Vatican City) received only two votes, because the current version of CLASITEX lacks the vote propagation algorithm (§3.2, remark 1).

In file "Res-xxx", CLASITEX writes in each line, first the chosen concept, and then the word (or words) selecting it. This file is used to understand votes from words towards concepts, allowing improvements to CLASITEX.

accidente: accidente. accidente: disloco. accidente: fractura.
accidente: fracturado. accidente: fracturo. accidente: hospital.
accidente: hospital. accidente: hospital. accidente: hospital.
accidente: hospital. accidente: hospital. accidente: hospital.
accidente: luxacion. accidente: resbalar. accidente: resbalo.
accidente: terapia. acto-juridico: audiencia. acto-juridico: audiencia.
africa: africanos. agencia-noticia: afp.
aleacion-metalica: aleacion metalica. alemania: canciller aleman.
alemania: helmut kohl. amor-pasion: amoroso. anunciar-promulgar: alocucion.
anunciar-promulgar: boletin .anunciar-promulgar: comento.
anunciar-promulgar: conferencia. anunciar-promulgar: declaracion.
anunciar-promulgar: manifesto. anunciar-promulgar: portavoz.
anunciar-promulgar: vocero. banarse: bano. banarse: descalzo.
banarse: ducha. banarse: mojado.
bebida-alcoholica: presidente. bebida-alcoholica: presidente.
belgica: belgica. budismo: dalai lama. budismo: monjes budistas.
casa-parte: bano. casa-parte: piso. casa-parte: piso.
cirugia: anestesia. cirugia: artroprotesis. cirugia: cirujano.
cirugia: escayola. cirugia: implantacion. cirugia: inmovilizacion.
cirugia: intervencion quirurgica. cirugia: intervencion quirurgica. cirugia: intervencion quirurgica.
cirugia: intervencion. cirugia: intervencion. cirugia: operacion.
cirugia: operada. (More words follow, I have truncated the list)

5.2 Una mujer española. (A Spanish woman) Text analyzed by CLASITEX.

Selecciones del Reader´s Digest. Junio 1994.
Una mujer espanola mas guapa que fea, llamada Maria Antonia Perez Blanco, ha sido la protagonista del mas reciente escandalo del sexo y del poder en el Reino Unido. La historia de esta increible aventura amorosa que ha ocupado las principales paginas de todos los medios informativos de Inglaterra, ha eclipsado las noticias de la propia corte de su graciosa majestad Isabel II.

No deja de ser apasionante y curioso el desarrollo de un relato que, como dice la letra de un aplaudido "bolero", "es la historia de un amor como no hay otro igual".

La verdadera historia de Maria Antonia Perez se origina cuando llega a londres de nina, despues del divorcio de sus padres; transcurrio el tiempo y la nina se convirtio en una joven dotada de grandes encantos fisicos y decidio que

era mejor la "buena vida" que tener que trabajar. Y de pronto se convirtio en la amante de un hombre de negocios iraqui, que la introdujo en los circulos de la alta sociedad inglesa. Al morir su amante, Maria Antonia adopta el nombre de "Bienvenida", nombre con el cual seguira ya su trayectoria en la vida. Su amante no le dejo herencia alguna, pero Bienvenida Perez sabia ya el camino que habia que seguir, y lo siguio.

Despues de trabajar poco tiempo como secretaria en una compania inmobiliaria, se convirtio en disenadora de modas, inventandose un titulo en "diseno y confeccion" expedido supuestamente en Dallas, Texas, EU. Y dan comienzo los "milagros": como por arte de magia sus disenos y confecciones comienzan a ser seguidos por las damas de la aristocracia britanica que la invitaban a fiestas, reuniones y cenas. Y fue en una de ellas donde conocio a sir Antony Buck, un alto personaje adinerado y diputado conservador, con quien se caso vestida de blanco a las tres semanas exactas de haberlo conocido.

Y es a partir de ese matrimonio, cuando se convertira en lady Bienvenida Buck y cuando da comienzo el gran escandalo.

Lady Buck no se conforma con tener un esposo, sino que aspira a tener tambien un "amante" cosa que, por medio de sus encantos fisicos, logra, cuando en otra reunion conoce a sir Peter Harding, un militar profesional, de muy alto rango y que es, nada menos, que ministro de la Defensa Nacional.

Pasa el tiempo, ya tenemos a Lady Buck convertida en esposa y con un "amante." Entonces a Bienvenida Perez se le ocurre otra "idea", que tambien le reportara cuantiosos beneficios economicos y hara llegar el escandalo a su punto maximo: discretamente se pone en contacto con un conocido hombre de "relaciones publicas", el cual a cambio de una comision de 20%, negocia con un periodico sensacionalista y por una cantidad de 175,000 libras esterlinas, la historia de Bienvenida Perez, la espanola esposa de un diputado con escano en la Camara de los Comunes y, al mismo tiempo, amante del jefe del ejercito britanico.

El periodico acepta el pago a cambio de la historia, pero exige logicamente una condicion: esta sera un "beso en la boca" a su amante sir Peter Harding.

Asi queda acordado. Y pocos dias despues, en un restaurante londinense, cuando Bienvenida aproxima sus rojos labios al amante, el cual le responde calurosamente, aparece un fotografo que recoge el momento estelar.

Momento estelar que aparecera en primera plana del periodico sensacionalista y cuyos efecto seran fulminantes: sir Peter Harding se ve obligado a dimitir; Bienvenida Perez, lady Buck, cobra las 175,000 libras esterlinas. La vida militar de sir Peter Harding totalmente arruinada; sir Antony Buck tramita el divorcio y lo obtiene; la esposa de sir Peter Harding, madre de tres hijos, hace lo mismo; los servicios secretos ingleses investigan ahora si Harding revelo a la espanola algunos "datos" de caracter militar y si el amante iraqui le comunico algun secreto sobre la guerra del golfo Persico. Y entre tanto, Bienvenida Perez cierra con broche de oro este escandalo contrayendo matrimonio con el ciudadano, de origen ruso, Nikola Sokolov, asegurando ella, que es "por amor", y recuerda la celebre cancion que dice: "La espanola cuando besa, es que besa de verdad". Pero el beso que dio a sir Peter Harding fue el "beso de Judas".

5.2.1 Results for “Una mujer española”.

File Cuen-una-mujer-espanola contains the following concepts histogram:

inglaterra *(England)*: 27 sexual *(sexual)*: 27 amor-pasion *(love-passion)*: 21
escandalo *(scandal)*: 10 periodicos *(newspapers)*: 6 espana: 4 iraq: 4
eeuu: 3 fiestas: 3 ahora: 2 congreso: 2 gobierno: 2
institucion: 2 poder-legislativo: 2 reino: 2 casamiento: 1
cuerpo-humano: 1 dias: 1 dirigente: 1 elecciones: 1
familia: 1 futbol-soccer: 1 inteligencia-seguridad: 1 muerte: 1
musica: 1 rusia: 1 sociedad: 1

5.3 México en el Mundial (Mexico in the World Tournament). Text analyzed by CLASITEX.

ARTICULO. Mexico: por su propia superacion. (Selecciones, junio 1994).
el Campeonato Mundial de los Estados Unidos 1994 significa la oportunidad de superacion para el seleccionado mexicano de futbol que dirige Miguel Mejia Baron. Un Mundial esperanzador para el futbol mexicano que asi regresa al concierto del futbol internacional despues de aquel castigo de dos anos que nos privo de asistir a la cita de Italia 1990.

En su paso por el Mundial, a Mexico le ha tocado en el Grupo E, el llamado de la muerte donde participara frente a los seleccionados de Italia, Eire y Noruega. La ventaja de esta situacion es que los tres europeos practican un mismo estilo de juego y bajo esa base, Mexico tiene la ventaja de no tener que entenderse por el momento con otros estilos diferentes. En su proceso, Mejia Baron observo a cerca de 60 jugadores y saco sus debidas conclusiones para afrontar el reto mundialista.

El equipo mexicano cuenta con grandes individualidades que en conjunto pueden dar un extraordinario resultado. Todo comienza atras con la presencia de Jorge Campos, el arquero-libero que juega al filo de la navaja, pero que con su inteligencia sale a flote. El Capi Ramirez Perales que con prestancia y buen futbol sale tocando de atras y por el lateral izquierdo la presencia de Ramon Ramirez, convertido en figura importante del esquema nacional, en la media cancha es importante la contencion de Ignacio Ambriz y el trabajo creativo que puedan aportar Benjamin Galindo y Alberto Garcia Aspe, adelante opciones brillantes con Hugo Sanchez, maduro y en gran momento, la velocidad de Luis Garcia, el arranque de Zague y el olfato goleador de Carlos Hermosillo. Ellos forman la columna de un equipo que tiene ante si _tambien en cada posicion_ elementos que han respondido a la confianza del tecnico nacional.

CUEN-MEXICO-EN-EL-MUNDIAL

futbol-soccer *(soccer)*: 14		deportes *(sports)*: 8		mexico *(México)*: 6	geografia *(geography)*: 3
mundo: 3	dias-meses: 2	escoger: 2	italia: 2	eeuu: 1	enfermedades: 1
hechos-violentos: 1		inteligencia-seguridad: 1		irlanda: 1	muerte: 1
musica: 1	noruega: 1	textiles-ropa: 1		trabajo: 1	trafico: 1

5.4 Conclusions and comments.

We have restricted our analysis of Spanish texts to articles appearing in magazines and newspapers of general circulation, not specialized. This limits the number of specialized concepts needed.

In short articles (México en el Mundial), CLASITEX offers less assurance in its themes: histograms contain fewer concepts, and the difference between main and secondary topics (and noise) is less marked.

The proposed approach looks promising, but many more experiments need to be carried in order to reach a definite conclusion. In particular, this approach requires the manual construction (§3.1) of the concept tree.

Since a mapping from words to nodes in a concept tree is not restricted to Spanish, the approach used in this paper applies to texts written in other languages, too (obviously, it requires the tree of concepts to be in the appropriate language). The approach, as it stands, will fail if one tries to apply it to the more demanding chores of summary formation, finding trends, documents comparison, answer of non-trivial questions, and other *language understanding* tasks mentioned

in the summary. To succeed here, a deeper approach is needed, using parsers, a tree with many more relations, disambiguation of references (“him”, “its”), and other tools, following the general lines of [1]. Such approach is being followed by the *Laboratory of Natural Language and Text Processing*, at C. I. C.-Politécnico.

The current implementation is made using Shell commands of Unix, and utilities such as awk, sed, etc. This allows easy changes to the program, but produces a slow execution (approximately as fast as a person reads the text). When a stable version is reached, we will rewrite the program in C, storing the tree in memory structures, accessed with indexes. We will use hashing if the tree needs to be stored in secondary memory.

Some improvements to CLASITEX appear marked with ♣ in the text. Since *all* words have *some* meaning, the meaningless words of §4.2 and of the directory of §4.2.1 shall be *given* the meaning they represent, by creating the corresponding files in arbol1, arbol2, arbol3 or arbol4. Sequences longer than 4 words (§3.3) need to be considered. The main improvement consists in increasing to a much bigger extent the concept tree. (In other words, most improvements will come not from clever programming, but by painfully tree construction).

5.5 Bibliography.

1. Douglas B. Lenat and R. V. Guha. Building large knowledge-based systems. Addison Wesley. 1989.