CAPÍTULO 3

# DESCRIPCIÓN DE LA TÉCNICA *BLINDLIGHT*

*En el capítulo anterior se analizaron algunas de las técnicas aplicables a tareas de PLN como clasificación, categorización, recuperación de información o extracción de resúmenes poniendo énfasis en los métodos estadísticos y en aquellos aspectos "transversales"a las distintas técnicas (p.ej. la representación vectorial de los documentos o el concepto de similitud interdocumental). En este capítulo se describirá blindLight, la técnica propuesta como prueba empírica de la tesis del autor. Se trata de un método de PLN de inspiración biológica, sencillo, robusto, y aplicable a múltiples idiomas. La idea subyacente es simple: obtener para cada documento un "genoma" único e independiente de cualquier colección que incluya a dicho documento. Este genoma estará constituido por n-gramas, secuencias de unos pocos caracteres extraídos del texto y ponderados estadísticamente para indicar su distinto grado de significatividad. Como se verá, blindLight está relacionada con el modelo vectorial y, al igual que éste, puede ser aplicada a la clasificación, categorización y recuperación de documentos aunque presenta importantes diferencias respecto al mismo. Por otro lado, la idea de un "genoma" documental es más que una metáfora puesto que dicho genoma puede ser "activado" resumiendo el documento del cual fue "extraído". A lo largo de este capítulo se describirá la técnica blindLight comparándola con otras técnicas aplicables a tareas similares y se apoyará el argumento del autor acerca de la capacidad de esta técnica para representar la semántica subyacente a los textos con independencia del idioma en que estén escritos.*

## 1 *blindLight*, una técnica bio-inspirada

La aplicación de técnicas bioinformáticas al tratamiento del lenguaje natural como propuso el autor para la Web Cooperativa no es una idea nueva. La alineación de múltiples secuencias de texto se ha aplicado, por ejemplo, a la generación de texto (Barzilay y Lee 2002), al aprendizaje automático de técnicas de paráfrasis (Barzilay y Lee 2003) o a la inducción de gramáticas (Kruijff 2002).

Tampoco es nueva la idea de "extraer" algún tipo de "ADN" a partir de texto libre. Por ejemplo, la mayor parte de las tecnologías pendientes de patente y presuntamente

desarrolladas por la empresa *Meaningful Machines*[1] emplean de un modo u otro la idea de "ADN del lenguaje":

> *existe un número finito de ideas discretas […] que Fluent Machines llama el 'ADN' del significado y que son universales y expresables en cualquier idioma.* (Abir *et al.* 2002, p. 216)

Resulta claro, en especial tras estudiar las solicitudes de patente (Abir 2003a, 2003b, 2003c y 2004), que su sistema se basa en el uso de *n*-gramas de caracteres y textos paralelos para establecer la asociación entre *n*-gramas de distintos idiomas:

> *el sistema […] construye bases de datos multilingües que contienen n-gramas de ADN de diversas longitudes […] y conecta traducciones de n-gramas en el lenguaje objetivo produciendo texto traducido.* (Abir *et al.* 2002, p. 217)

Puede parecer entonces que *blindLight* no es una técnica excesivamente novedosa al pretender utilizar técnicas bioinformáticas (básicamente alineación de secuencias) para comparar cadenas de pseudo-ADN constituidas, a su vez, por *n*-gramas de caracteres. No obstante, como se irá mostrando a lo largo de este trabajo, *blindLight* supone la aportación de una serie de ideas nuevas y originales:

1. Para construir el "genoma" de los documentos se emplea una medida estadística de la **"significatividad"** de los distintos *n*-gramas que va un paso más allá de la clásica frecuencia de aparición.
2. El "genoma" de un documento puede "actuar" sobre el texto del documento, a modo de "ARN transferente", transformándolo en resúmenes y frases clave.
3. No se emplean técnicas de alineación de secuencias para las comparaciones de "genomas" sino que éstos pueden combinarse constituyendo "híbridos" que serán comparados con los originales a fin de determinar la similitud entre los mismos.

Una versión preliminar de estas ideas, en particular de las dos primeras, se puede encontrar en (Gayo Avello, Álvarez Gutiérrez y Gayo Avello 2004a y 2004b). En el siguiente apartado se describirá con detalle cómo se construyen y comparan tales "genomas documentales" y en posteriores capítulos el modo en que dichos "genomas" transforman el texto plano original generando resúmenes automáticos. Baste para concluir la definición de "ADN de un documento" tal y como se entiende en *blindLight*:

> *El ADN de un documento es un conjunto de genes donde cada gen está formado por un n-grama de caracteres y su correspondiente significatividad dentro del documento de origen.*

## 2 Fundamentos teóricos de *blindLight*

En este apartado se describirá con detalle la nueva técnica que satisface la afirmación con que el autor comenzaba su tesis:

> *Se puede obtener para los distintos n-gramas, $g_i$, de un texto escrito en cualquier idioma una medida de su significatividad, $s_i$, distinta de la frecuencia relativa de aparición de los mismos en el texto, $f_i$, pero calculable a partir de la misma. Esta métrica de la significatividad intradocumental de los n-gramas permite asociar a cada documento, $d_i$, un único vector, $v_i$, susceptible de comparación con cualquier otro vector obtenido del mismo modo aun cuando sus respectivas longitudes puedan diferir.*

[1] `http://www.meaningfulmachines.com/`

*blindLight*, al igual que otras técnicas basadas en vectores de *n*-gramas, representa cada documento como un vector de pesos. Sin embargo, estos vectores difieren en varios aspectos de los utilizados en modelos vectoriales "clásicos". En primer lugar, no se considera a los documentos vectores en un espacio *T*-dimensional sino que dos vectores cualesquiera tendrán, muy probablemente, distintas dimensiones; o lo que es lo mismo, en esta propuesta no existe un auténtico "espacio vectorial" y no se recurre a medidas de asociación análogas a operaciones vectoriales. Por otro lado, los pesos empleados en cada vector no son las frecuencias relativas de aparición[1] de los *n*-gramas sino la "significatividad" de cada *n*-grama dentro del documento. Dicha significatividad, como se verá más adelante, se obtiene a partir de las frecuencias relativas.

Calcular una medida de la relación entre los elementos de un *n*-grama y, así, la relevancia del *n*-grama en su conjunto, su significatividad[2], no es un problema reciente. No obstante, tan sólo se citarán aquí dos trabajos representativos, el hecho por Dunning y el llevado a cabo por Ferreira da Silva y Pereira Lopes. Ted Dunning (1993) describió un método basado en el test de razón de verosimilitud (*likelihood ratio test*) para detectar palabras clave y terminología. No obstante, empleando su técnica sólo se podían detectar bigramas de palabras (por ejemplo, `likelihood ratio` o `ratio test`, nunca `likelihood ratio test`). Fueron Joaquim Ferreira da Silva y Gabriel Pereira Lopes (1999) quienes presentaron un método para generalizar una serie de estadísticos[3] a *n*-gramas de palabras de longitud arbitraria a fin de extraer frases clave. Además de esto, introdujeron una nueva medida (véase la ecuación 2), la Probabilidad Condicional Simétrica (*Symmetrical Conditional Probability*), que según sus autores supera a las anteriores, incluyendo los resultados alcanzados por Dunning.

*blindLight* aplica estos estadísticos no a *n*-gramas de palabras sino de caracteres, midiendo de este modo la relación entre los caracteres constituyentes de cada *n*-grama y, por tanto, la significatividad de éste dentro de un único documento. Así, para cada *n*-grama se calcula su significatividad constituyendo cada par `(n-grama, significatividad)` un componente[4] del vector correspondiente a un documento dado. Por tanto, los vectores de los documentos no se construyen en relación a un *corpus* ponderando los términos en función de la frecuencia de aparición en la colección de documentos.

La técnica no obliga a emplear un estadístico en particular para la ponderación de los *n*-gramas y debería estudiarse para cada aplicación cuál resulta el más adecuado. No obstante, es preciso señalar que, mientras no se indique lo contrario, los resultados descritos en este trabajo se han obtenido empleando la información mutua (véase la ecuación 3) y en

---

[1] Normalizadas sobre la base de un *corpus* y/o la longitud del documento.

[2] A lo largo de este trabajo se utilizará el término "significatividad" para referirse al valor real asignado a cada *n*-grama de caracteres dentro de un documento. Se usa éste término en lugar del habitual "peso" para distinguir el modo en que se obtienen ambos valores en *blindLight* y en el modelo vectorial, respectivamente. En este último el cálculo de los pesos involucra no sólo la frecuencia de aparición de los términos en el propio documento (*tf*) sino también su distribución en los distintos documentos de la colección (*idf*) siendo de hecho más importante este último valor. Por el contrario, en *blindLight* el valor asignado a cada *n*-grama se deriva únicamente a partir del propio documento siendo innecesario recurrir a la colección de documentos.

[3] Información mutua (véase ecuación 3), $\phi^2$, *log likelihood* (Dunning 1993) y Dice. Las ecuaciones para la generalización a *n*-gramas del resto de estadísticos se encuentran en la página 148.

[4] Un "gen" del "ADN documental".

algunos casos, como en el sistema de recuperación de información participante en *CLEF*[1] *2004* (Peters *et al.* 2005), la probabilidad condicional simétrica (véase la ecuación 2).

$$Avp = \frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot p(w_{i+1}...w_n) \quad (1)$$

$$SCP\_f((w_1...w_n)) = \frac{p(w_1...w_n)^2}{Avp} \quad (2)$$

$$SI\_f((w_1...w_n)) = \log(\frac{p(w_1...w_n)}{Avp}) \quad (3)$$

**Fig. 27 Cálculo de la probabilidad condicional simétrica (*SCP_f*) y la información mutua (*SI_f*) para *n*-gramas de caracteres.**

*$(w_1...w_n)$* es un *n*-grama, *$(w_1...w_i)$* y *$(w_{i+1}...w_n)$* son fragmentos consecutivos del mismo (p.ej. para el *n*-grama *'info'* se tendría <*'i'*, *'nfo'*>, <*'in'*, *'fo'*> y <*'inf'*, *'o'*>). *$p((w_1...w_n))$* es la probabilidad del *n*-grama *$(w_1...w_n)$* en el texto, *$p((w_1...w_i))$* es la probabilidad de que un *n*-grama comience con los caracteres *$(w_1...w_i)$* y *$p(w_{i+1}...w_n)$* de que termine en *$(w_{i+1}...w_n)$*.

A continuación se muestra el modo en que se calcula la significatividad de 4-gramas de caracteres utilizando una de las historias más cortas jamás escritas: *"El Dinosaurio"* de Augusto Monterroso.

```
Cuando despertó, el dinosaurio todavía estaba allí.
```

**Fig. 28 *"El Dinosaurio"* de Augusto Monterroso.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Cuan | uand | ando | ndo_ | do_d | o_de | **_des** | desp |
| espe | sper | pert | ertó | rtó_ | tó_e | ó_el | **_el_** |
| el_d | l_di | **_din** | dino | inos | **nosa** | osau | saur |
| auri | urio | rio_ | io_t | o_to | **_tod** | **toda** | odav |
| daví | **avía** | vía_ | ía_e | a_es | **_est** | **esta** | stab |
| **taba** | aba_ | **ba_a** | a_al | **_all** | allí | | |

**Fig. 29 4-gramas del texto anterior** (se han sustituido los espacios en blanco por guiones bajos)**.**

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **_→** | **6** | _a→ | 1 | _al→ | 1 | _d→ | 2 | _de→ | 1 | _di→ | 1 | _e→ | 2 | _el→ | 1 |
| _es→ | 1 | _t→ | 1 | _to→ | 1 | a→ | 7 | a_→ | 2 | a_a→ | 1 | a_e→ | 1 | ab→ | 1 |
| aba→ | 1 | al→ | 1 | all→ | 1 | an→ | 1 | and→ | 1 | au→ | 1 | aur→ | 1 | av→ | 1 |
| aví→ | 1 | b→ | 1 | ba→ | 1 | ba_→ | 1 | C→ | 1 | Cu→ | 1 | Cua→ | 1 | d→ | 4 |
| da→ | 1 | dav→ | 1 | de→ | 1 | des→ | 1 | di→ | 1 | din→ | 1 | do→ | 1 | do_→ | 1 |
| e→ | 4 | el→ | 1 | el_→ | 1 | er→ | 1 | ert→ | 1 | es→ | 2 | esp→ | 1 | est→ | 1 |
| i→ | 2 | í→ | 1 | ía→ | 1 | ía_→ | 1 | in→ | 1 | ino→ | 1 | io→ | 1 | io_→ | 1 |
| l→ | 1 | l_→ | 1 | l_d→ | 1 | lí→ | 0 | ll→ | 0 | lli→ | 0 | n→ | 2 | nd→ | 1 |
| ndo→ | 1 | no→ | 1 | nos→ | 1 | o→ | 4 | ó→ | 1 | o_→ | 2 | ó_→ | 1 | o_d→ | 1 |
| ó_e→ | 1 | o_t→ | 1 | od→ | 1 | oda→ | 1 | os→ | 1 | osa→ | 1 | p→ | 1 | pe→ | 1 |
| per→ | 1 | r→ | 2 | ri→ | 1 | rio→ | 1 | rt→ | 1 | rtó→ | 1 | s→ | 3 | sa→ | 1 |
| sau→ | 1 | sp→ | 1 | spe→ | 1 | st→ | 1 | sta→ | 1 | t→ | 3 | ta→ | 1 | tab→ | 1 |
| to→ | 1 | tó→ | 1 | tó_→ | 1 | tod→ | 1 | u→ | 2 | ua→ | 1 | uan→ | 1 | ur→ | 1 |
| uri→ | 1 | v→ | 1 | ví→ | 1 | vía→ | 1 | | | | | | | | |

**Fig. 30 Fragmentos iniciales de los 4-gramas anteriores junto con sus frecuencias absolutas.**

[1] *Cross Language Evaluation Forum* es un foro internacional que tiene como objetivos el desarrollo de una infraestructura para la evaluación de sistemas de recuperación de información que operen sobre idiomas europeos así como la creación de conjuntos de prueba reutilizables <http://www.clef-campaign.org>.

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| →_ | 6 | →_a | 1 | →_al | 1 | →_d | 2 | →_de | 1 | →_di | 1 | →_e | 2 | →_el | 1 |
| →_es | 1 | →_t | 1 | →_to | 1 | **→a** | **6** | →a_ | 2 | →a_a | 1 | →a_e | 1 | →ab | 1 |
| →aba | 1 | →al | 1 | →all | 1 | →an | 1 | →and | 1 | →au | 1 | →aur | 1 | →av | 1 |
| →aví | 1 | →b | 1 | →ba | 1 | →ba_ | 1 | →C | 0 | →Cu | 0 | →Cua | 0 | →d | 4 |
| →da | 1 | →dav | 1 | →de | 1 | →des | 1 | →di | 1 | →din | 1 | →do | 1 | →do_ | 1 |
| →e | 4 | →el | 1 | →el_ | 1 | →er | 1 | →ert | 1 | →es | 2 | →esp | 1 | →est | 1 |
| →i | 2 | →í | 2 | →ía | 1 | →ía_ | 1 | →in | 1 | →ino | 1 | →io | 1 | →io_ | 1 |
| →l | 3 | →l_ | 1 | →l_d | 1 | →lí | 1 | →ll | 1 | →llí | 1 | →n | 2 | →nd | 1 |
| →ndo | 1 | →no | 1 | →nos | 1 | →o | 4 | →ó | 1 | →o_ | 2 | →ó_ | 1 | →o_d | 1 |
| →ó_e | 1 | →o_t | 1 | →od | 1 | →oda | 1 | →os | 1 | →osa | 1 | →p | 1 | →pe | 1 |
| →per | 1 | →r | 2 | →ri | 1 | →rio | 1 | →rt | 1 | →rtó | 1 | →s | 3 | →sa | 1 |
| →sau | 1 | →sp | 1 | →spe | 1 | →st | 1 | →sta | 1 | →t | 3 | →ta | 1 | →tab | 1 |
| →to | 1 | →tó | 1 | →tó_ | 1 | →tod | 1 | →u | 1 | →ua | 0 | →uan | 1 | →ur | 1 |
| →uri | 1 | →v | 1 | →ví | 1 | →vía | 1 | | | | | | | | |

**Fig. 31 Fragmentos finales de los 4-gramas anteriores junto con sus frecuencias absolutas.**

$$p(\text{Cuan}) = \frac{1}{46} \left| \begin{array}{lll} p(\text{C}\rightarrow) = \frac{1}{138} & p(\text{Cu}\rightarrow) = \frac{1}{138} & p(\text{Cua}\rightarrow) = \frac{1}{138} \\ \times & \times & \times \\ p(\rightarrow\text{uan}) = \frac{1}{138} & p(\rightarrow\text{an}) = \frac{1}{138} & p(\rightarrow\text{n}) = \frac{2}{138} \end{array} \right.$$

$$\frac{1}{138^2} + \frac{1}{138^2} + \frac{2}{138^2} = \frac{4}{138^2}$$

$$Avp = \frac{1}{3} \cdot \frac{4}{138^2}$$

$$SI\_f(\text{Cuan}) = \log \frac{\frac{1}{46}}{\frac{1}{3} \cdot \frac{4}{138^2}} = 2{,}492$$

**Fig. 32 Cálculo de la significatividad del 4-grama `Cuan` empleando información mútua (*SI_f*).**

$$p(\text{ando}) = \frac{1}{46} \left| \begin{array}{lll} p(\text{a}\rightarrow) = \frac{7}{138} & p(\text{an}\rightarrow) = \frac{1}{138} & p(\text{and}\rightarrow) = \frac{1}{138} \\ \times & \times & \times \\ p(\rightarrow\text{ndo}) = \frac{1}{138} & p(\rightarrow\text{do}) = \frac{1}{138} & p(\rightarrow\text{o}) = \frac{4}{138} \end{array} \right.$$

$$\frac{7}{138^2} + \frac{1}{138^2} + \frac{4}{138^2} = \frac{12}{138^2}$$

$$Avp = \frac{1}{3} \cdot \frac{12}{138^2}$$

$$SI\_f(\text{ando}) = \log \frac{\frac{1}{46}}{\frac{1}{3} \cdot \frac{12}{138^2}} = 2{,}015$$

**Fig. 33 Cálculo de la significatividad del 4-grama `ando` empleando información mútua (*SI_f*).**

Por otro lado, *blindLight* no utiliza operaciones vectoriales para comparar vectores de documentos. En cambio, para determinar la similitud entre dos vectores dados se obtiene un nuevo vector mediante la intersección de los dos anteriores y se compara la significatividad total del vector resultante con la de los vectores originales.

Esta operación es, en cierto modo, similar al coeficiente de solapamiento propuesto por Salton (1968, citado por Jardine y van Rijsbergen 1971) con dos salvedades. En primer lugar, no se interpreta la intersección entre vectores con pesos reales como un producto escalar sino como la suma de los pesos de un vector que tiene como componentes los *n*-gramas presentes en ambos documentos y como pesos los mínimos que aparecen en cada

vector original. En segundo lugar, *blindLight* no utiliza un único coeficiente sino dos al comparar el vector intersección resultante con cada vector documento de manera independiente.

El motivo de todo esto se explicará más adelante ya que antes de continuar es preferible expresar lo anterior en forma de ecuaciones. Sean $Q$ y $T$ dos vectores *blindLight* de dimensiones $m$ y $n$:

$$Q = \{(k_{1Q}, w_{1Q}) \quad (k_{2Q}, w_{2Q}) \quad \ldots \quad (k_{mQ}, w_{mQ})\} \tag{4}$$

$$T = \{(k_{1T}, w_{1T}) \quad (k_{2T}, w_{2T}) \quad \ldots \quad (k_{nT}, w_{nT})\} \tag{5}$$

$k_{ij}$ es el $n$-grama i-ésimo en el documento $j$ y $w_{ij}$ es la significatividad de dicho $n$-grama calculada aplicando cualquiera de los estadísticos generalizados a $n$-gramas (Ferreira da Silva y Pereira Lopes 1999).

Es posible definir entonces la significatividad total para los vectores $Q$ y $T$, $S_Q$ y $S_T$ respectivamente, como:

$$S_Q = \sum_{i=1}^{m} w_{iQ} \tag{6}$$

$$S_T = \sum_{i=1}^{n} w_{iT} \tag{7}$$

El operador de intersección antes mencionado, Ω, se define de la forma siguiente:

$$Q\Omega T = \left\{ (k_x, w_x) \middle/ \begin{array}{l} (k_x = k_{iQ} = k_{jT}) \wedge (w_x = \min(w_{iQ}, w_{jT})), \\ (k_{iQ}, w_{iQ}) \in Q, 0 \le i < m, \\ (k_{jT}, w_{jT}) \in T, 0 \le j < n \end{array} \right\} \tag{8}$$

De manera similar a como muestran las ecuaciones 6 y 7 se puede calcular la significatividad total para el vector resultante de la intersección de los vectores a comparar:

$$S_{Q\Omega T} = \sum w_{iQ\Omega T} \tag{9}$$

Finalmente, se definen dos medidas de asociación, una para comparar $Q$ con $T$, Π (Pi mayúscula), y otra para comparar $T$ con $Q$, P (Ro mayúscula):

$$\Pi = S_{Q\Omega T} / S_Q \tag{10}$$

$$\mathrm{P} = S_{Q\Omega T} / S_T \tag{11}$$

Si se supone que $T$ es un documento y $Q$ una consulta pueden asimilarse las medidas Π y P con la precisión y la **exhaustividad** (*recall*), respectivamente. Recordemos que la precisión es la proporción de documentos retornados por un sistema que son realmente relevantes mientras que la exhaustividad es la proporción de documentos relevantes en la colección que aparecen entre los resultados. En este sentido, Π revelaría en

qué medida la consulta queda "satisfecha" por la intersección entre ésta y el documento resultante mientras que P señalaría lo propio entre la intersección y el documento.

De este modo, un valor de Π igual a la unidad implicaría que la necesidad de información formulada en la consulta queda totalmente satisfecha con el documento mientras que un valor de P unitario indicaría que la consulta define por completo al documento. Naturalmente, puesto que la significatividad total de consulta y documento depende en gran medida del número de *n*-gramas de cada uno y, en consecuencia, de la longitud del texto, los valores de Π y P difícilmente tomarán valores próximos a la unidad de manera simultánea[1] por lo que será necesario combinar ambas medidas de modo que sea posible obtener un único valor que indique la similitud entre documentos de distinto tamaño (para más detalles sobre este tipo de medidas véase la página 141). A continuación se presenta un ejemplo ilustrativo de estos conceptos empleando de nuevo el texto de Monterroso junto con su traducción al portugués.

Cuando despertó, el dinosaurio todavía estaba allí. (*Q*)
Quando acordou, o dinossauro ainda estava lá. (*T*)

**Fig. 34 *"El Dinosaurio"* de Augusto Monterroso, original en español y traducción al portugués.**

| **Vector Q (45 elementos)** | | **Vector T (39 elementos)** | | **QΩT (10 elementos)** | |
|---|---|---|---|---|---|
| Cuan | 2,492 | va_l | 2,545 | saur | 2,244 |
| l_di | 2,392 | rdou | 2,323 | inos | 2,177 |
| stab | 2,392 | stav | 2,323 | uand | 2,119 |
| … | | … | | _est | 2,091 |
| saur | 2,313 | saur | 2,244 | dino | 2,022 |
| desp | 2,313 | noss | 2,177 | _din | 2,022 |
| … | | … | | esta | 2,012 |
| ndo_ | 2,137 | a_lá | 2,022 | ndo_ | 1,981 |
| nosa | 2,137 | o_ac | 2,022 | a_es | 1,943 |
| … | | … | | ando | 1,876 |
| ando | 2,015 | auro | 1,908 | | |
| avía | 1,945 | ando | 1,876 | | |
| _all | 1,915 | do_a | 1,767 | **Π: 0,209 P: 0,253** | |

**Fig. 35 Vectores *blindLight* para los documentos mostrados en Fig. 34.**

Los vectores se han calculado siguiendo el proceso descrito desde Fig. 29 a Fig. 33, además, se han truncado para mostrar 10 elementos (a excepción del vector intersección que aparece completo). Los espacios en blanco han sido reemplazados por guiones bajos.

Es preciso señalar que los vectores *blindLight* no están normalizados, esto es, su módulo no es unitario (véase Fig. 35) sino que los valores de significatividad obtenidos en cada documento son utilizados directamente como pesos de los *n*-gramas. La razón es sencilla, puesto que el operador intersección antes definido produce un nuevo vector que tiene como pesos los mínimos de los vectores intersecados, la normalización de los mismos tendría como consecuencia que el documento más largo de los dos, aquel, por tanto, con más *n*-gramas diferentes y, en consecuencia, menores pesos sería siempre el más influyente en la comparación lo cual es contrario al objetivo de la normalización, esto es, garantizar que el tamaño de los documentos no es un factor determinante en las comparaciones.

[1] Salvo que se comparen documentos de tamaño similar como en los siguientes ejemplos.

Por otro lado, si todos los vectores fuesen unitarios no tendría sentido alguno hablar de dos medidas de comparación asimétricas que, en opinión del autor, dotan de gran flexibilidad a la nueva técnica ya que combinando linealmente Π y P es posible construir nuevas medidas de asociación. En el capítulo dedicado a recuperación de información se tratará más este asunto, baste mostrar aquí una de las más simples[1], la denominada *PiRo*:

$$\frac{\Pi + P}{2} \qquad (\mathbf{12})$$

Esta medida de asociación es, por supuesto, simétrica[2] y proporciona valores entre 0 y 1 (parecido nulo y documentos idénticos, respectivamente). Pruebas experimentales manifiestan una correlación elevada entre esta medida de asociación y la función del coseno; podría ser interesante verificar si es o no monótona con respecto a esa y otras funciones de asociación pero, no siendo central para este trabajo, aún no se ha llevado a cabo dicho estudio. Por otro lado, y adelantándose al capítulo dedicado a trabajo futuro, la utilización de dos medidas asimétricas como base para construir medidas de asociación ofrece la interesante posibilidad de utilizar programación genética para obtener nuevas medidas de un modo similar al seguido por Fan, Gordon y Pathak (2004a y 2004b).

Además, existe una serie de trabajos relativos a nuevas métricas de la similitud entre *ítems* de información que resultan muy interesantes de cara a una futura integración con la técnica del autor:

- Bennett *et al.* (1998) describen cómo se puede elaborar una medida universal de la distancia "cognitiva" entre dos objetos cualesquiera (representables mediante cadenas de bits) basándose en la complejidad de Kolmogorov (función *K*). Puesto que dicha función *K* no es computable Chen, Kwong y Li (1999) la aproximan a partir de los resultados obtenidos con un algoritmo de compresión[3] para cadenas de ADN sugiriendo la posibilidad de emplear esta técnica para la comparación de genomas sin necesidad de recurrir a su alineación. Posteriormente, Li *et al.* (2004) continuan los trabajos anteriores demostrando su aplicación a la clasificación de organismos biológicos y lenguajes naturales. Varré, Delahaye y Rivals (1999) llevaron a cabo un trabajo muy similar[4].
- Rubner, Tomasi y Guibas (2000) describen la *Earth Mover's Distance* (*EMD*) basada en el coste mínimo necesario para transformar una distribución de valores en otra. Algunas de las ventajas de la *EMD* son su capacidad para trabajar sobre representaciones de longitud variable y para soportar coincidencias parciales. Por el momento, esta distancia se ha utilizado con imágenes, no con texto.
- Muthukrishnan y Sahinalp (2000) estudian el problema de los "vecinos más próximos a una secuencia" (*sequence nearest neighbors*), es decir, la forma de encontrar en una colección *D* de secuencias aquella cuya distancia de edición a otra secuencia

[1] Como se describirá en el capítulo dedicado a recuperación de información esta medida es válida para comparar documentos. Sin embargo, no es demasiado adecuada para comparar consultas con documentos debido a las diferencias de tamaño; en este último caso se hace necesario "escalar" de algún modo los valores de P a fin de permitir una comparación razonable con Π.

[2] Es decir, $PiRo(d_1, d_2) = PiRo(d_2, d_1)$.

[3] `http://www.cs.cityu.edu.hk/~cssamk/gencomp/GenCompress1.htm`

[4] `http://www.lifl.fr/~varre/TD/tdsoft.html`

*Q* sea mínima. En su trabajo describen un método eficiente para obtener una solución aproximada para dicha búsqueda.

En resumen, *blindLight* es una técnica que utiliza vectores de pesos asignados a *n*-gramas de caracteres para representar documentos sin requerir ningún *corpus* para determinar dichos pesos. Para ello, emplea una medida de la significatividad de los *n*-gramas que puede ser calculada a partir de su frecuencia relativa de uso en un único documento. Los vectores así obtenidos pueden ser comparados no recurriendo a operaciones vectoriales sino empleando un nuevo vector "híbrido" resultado de la intersección de los vectores a comparar. De este modo, al relacionar la significatividad total de este nuevo vector con la de los vectores originales es posible obtener dos medidas asimétricas combinables linealmente para construir nuevas funciones de asociación entre documentos o entre documentos y consultas.

Como se verá en los siguientes capítulos el hecho de que la información para describir un documento se obtenga únicamente del propio documento sin recurrir a lo colección que lo incluye ni a un *corpus* externo no resulta en perjuicio de los resultados sino que estos son comparables[1] a los de otras técnicas que sí utilizan información de la colección para construir los vectores documentales.

## 3 Diferencias entre *blindLight* y otras técnicas PLN

Así pues, *blindLight* permite obtener para documentos escritos en cualquier lenguaje natural un "genoma" representado mediante un vector de *n*-gramas de caracteres. Cada *n*-grama (cada "gen") tiene asociado un peso que indica su significatividad en el texto original, significatividad calculada empleando un estadístico generalizado a *n*-gramas (Ferreira da Silva y Pereira Lopes 1999).

En este sentido *blindLight* diferiría muy poco de diversas implementaciones del modelo vectorial que han empleado como términos *n*-gramas de caracteres extraídos de los documentos. No obstante, a diferencia de tales implementaciones, la nueva técnica aquí descrita no emplea medidas de similitud "tradicionales" como por ejemplo la función del coseno. En su lugar, plantea la posibilidad de "hibridar" los vectores correspondientes a dos documentos obteniendo un tercer vector "artificial" y, mediante el mismo, obtener dos medidas asimétricas, Π y P, susceptibles de ser combinadas para constituir distintas medidas de similitud documental.

De este modo, resulta muy simple aplicar *blindLight* (al igual que el modelo vectorial) a tareas de clasificación, categorización o recuperación de información pero, además de éstas, también es posible emplear el "genoma" extraído de un documento para obtener resúmenes automáticos.

En cuanto método de PLN que emplea *n*-gramas de caracteres pueden encontrarse ciertas similitudes entre *blindLight* y técnicas como *Acquaintance* (Damashek 1995), *Highlights* (Cohen 1995) o la descrita por Neto *et al.* (2000). Sin embargo, tales similitudes son superficiales. Así, en el caso de *Acquaintance* se utilizan como pesos las frecuencias relativas de los *n*-gramas y no es aplicable a la obtención de resúmenes automáticos. *Highlights*, sólo permite obtener términos clave (no resúmenes) y se basa en la utilización de un "contexto" para cada documento al contrario que *blindLight* que extrae resúmenes empleando únicamente el documento a resumir. Por lo que respecta a la propuesta de Neto *et al.*

[1] A excepción, de momento, del sistema de recuperación de información como se verá en el capítulo 6.

únicamente obtiene resúmenes automáticos y lo hace siguiendo un enfoque análogo a la recuperación de información extendido a la recuperación de sentencias.

En cuanto a las propuestas de Eli Abir (2003a, 2003b, 2003c y 2004) (Abir *et al.* 2002) existe una aparente semejanza debido a la metáfora del "genoma documental" y la utilización de *n*-gramas de caracteres. No obstante, Abir no utiliza la misma técnica de ponderación que se emplea en *blindLight*, no parece aprovechar la posibilidad de "combinar" los genomas de distintos textos y no está claro si sus técnicas permiten obtener resúmenes automáticos ni en qué modo lo harían.

En definitiva, *blindLight*, al igual que otras técnicas, utiliza vectores de *n*-gramas de caracteres para representar documentos. Sin embargo, a diferencia de ellas emplea estos estadísticos generalizados para obtener los pesos de tales *n*-gramas y parece ser la única propuesta que plantea calcular un vector único para cada documento con independencia de la colección en la que se incluya. Además, no utiliza ninguna de las medidas de similitud habituales en los modelos vectoriales sino una basada en el uso de vectores obtenidos por combinación de otros y que en modo alguno son centroides. Por otro lado, el "genoma documental" va más lejos que la simple metáfora al poder combinarse con el texto original para producir resúmenes automáticos.

## 4 Semántica subyacente a los vectores *blindLight*

Antes de continuar es necesario dar soporte a la siguiente afirmación formulada por el autor en su tesis y sobre la que se apoyan todas las aplicaciones posteriores:

> …*Puesto que tales vectores almacenan ciertos aspectos de la semántica subyacente a los textos originales, el mayor o menor grado de similitud entre los mismos constituye un indicador de su nivel de relación conceptual*…

Para ello se debe aclarar en primer lugar el modo en que se utiliza aquí el término "semántica" y la forma más sencilla de hacerlo es señalando qué es lo que NO se afirma acerca de la nueva técnica propuesta por el autor:

- Al hablar de semántica no se pretende establecer ningún tipo de relación con la rama de la lingüística del mismo nombre. Ya se dijo con anterioridad que *blindLight* pretende, por el contrario, ser una técnica válida para escenarios en los que el conocimiento acerca de un idioma sea nulo.
- Tampoco se trata de establecer paralelismos con problemas de PLN tales como desambigüación o construcción automática de tesauros. Es necesario insistir en que la técnica no tiene como objetivo extraer ningún tipo de conocimiento estructurado a partir de los textos procesados.
- De igual modo *blindLight* tampoco tiene como objetivo facilitar la construcción automática de ontologías ni comparte rasgo alguno con la Web Semántica.

Así pues, ¿en qué sentido emplea el autor el término "semántica"? De un modo semejante al que lo hacen Susan Dumais *et al.* (1988, p. 282):

> *Damos por supuesto que bajo los datos de uso de palabras existe algún tipo de estructura semántica "latente" parcialmente oculta por la variabilidad en la elección de esas palabras. Utilizamos técnicas estadísticas para estimar dicha estructura latente y eliminar el "ruido".*

De manera similar, la presunción básica de este trabajo es que al descomponer el texto en *n*-gramas de caracteres y estimar una medida de la significatividad de dichos

*n*-gramas en un documento se dispone de una información no sólo valiosa sino fácilmente comparable con la de otros documentos a fin de determinar el grado de similitud a un nivel que va más allá de las simples palabras y que puede calificarse de "conceptual" o "semántico".

Como apoyo parcial a esto hay que señalar que se han desarrollado distintas iniciativas basadas en el uso de *n*-gramas para llevar a cabo tareas en las que habitualmente se requeriría el juicio de un humano. Así, Kishore Papineni *et al.* (2002) han desarrollado un método, *BLEU*, basado en *n*-gramas de palabras para evaluar la calidad de traducciones automáticas comparándolas con distintas traducciones de referencia producidas por humanos. Por otro lado, Chin-Yew Lin y Eduard Hovy (2003) demuestran que la utilización de *n*-gramas[1], nuevamente de palabras, permite evaluar con gran precisión la calidad de un resumen automático. Este trabajo evolucionó posteriormente hasta la implementación de una herramienta para realizar dicha evaluación: *ROUGE* (Lin 2004a).

En lo que resta de este apartado se va a describir un pequeño experimento que trata de sustentar la afirmación con que se comenzaba. Dicho experimento ha involucrado el uso de *blindLight* para clasificación de documentos. Los detalles sobre la implementación de dicha técnica se darán en capítulos posteriores y allí se proporcionarán más resultados. Los que se muestran aquí pretenden únicamente resultar suficientemente elocuentes acerca de la capacidad de *blindLight* para "modelar" la semántica subyacente a un texto.

## 4.1 Clasificación automática de (mini)corpora paralelos

Una de las utilidades más inmediatas de *blindLight* es la clasificación automática de documentos para lo cual se utiliza la medida de asociación *PiRo* mostrada en la página 66 (véase ecuación 12) y un algoritmo similar en ciertos aspectos al propuesto por R.A. Jarvis y Edward Patrick (1973). Si *blindLight* conservase realmente la semántica de los documentos entonces las clasificaciones de *corpora* paralelos deberían ser muy similares con independencia del idioma empleado en cada *corpus* y, además, plausibles según criterios de clasificación humanos.

En el experimento que se describe a continuación se puso a prueba semejante hipótesis. Para ello se buscaron en primer lugar documentos disponibles en formato electrónico y para los cuales existiesen traducciones en, al menos, los siguientes idiomas[2]: castellano (`ES`), finés (`FI`), francés (`FR`), hebreo (`HE`), holandés (`NL`), inglés (`EN`) y japonés (`JA`). Los documentos finalmente seleccionados fueron los siete siguientes:

- `CreativeCommons`: Licencia *Creative Commons* en su versión Atribución-No Comercial-Compartir en Igualdad.
- `Genesis`: Génesis 1:1-3:24.
- `GNU-GPL`: Licencia *GPL*.

[1] En particular unigramas y bigramas.

[2] Esos siete idiomas cubren las siguientes familias lingüísticas: indoeuropea (castellano, francés, holandés e inglés), urálica (finés), afroasiática (hebreo) y japonesa (japonés). Por lo que respecta a los idiomas indoeuropeos, el francés y el castellano son lenguas romances mientras que el holandés y el inglés son germánicas. *A priori*, sería de esperar encontrar similitudes entre las clasificaciones obtenidas para francés y castellano o para holandés e inglés debido a la mayor relación existente entre dichos lenguajes. En caso de encontrarse similitudes entre clasificaciones de documentos procedentes de familias distintas podría apoyarse con cierta confianza la capacidad de *blindLight* para conservar rasgos semánticos subyacentes al texto e independientes de características de un idioma o familia de idiomas en particular.

- `GoogleTermsOfService`: Condiciones de Servicio de *Google* en su versión de septiembre de 2004.
- `MSNTermsOfService`: Condiciones de Uso de *MSN* en su versión de abril de 2003.
- `UN-ConventionSaleGoods`: Convención de Viena sobre Compraventa Internacional de Mercaderías.
- `UN-HumanRights`: Declaración Universal de los Derechos Humanos.

Los documentos originales en inglés tienen longitudes bastante diferentes variando entre las 1.662 palabras de `GoogleTermsOfService` a las 11.182 de `MSNTermsOfService`. A fin de evitar que el tamaño pudiese afectar a los resultados algunos documentos fueron truncados para aproximar su longitud (en bytes) a la de aquellos más pequeños. De este modo el *corpus* final en inglés tuvo las siguientes características:

- `CreativeCommons`: Truncado a partir del apartado 5, tamaño final: 1.723 palabras.
- `Genesis`: Completo, tamaño final: 2.204 palabras.
- `GNU-GPL`: Truncado a partir del segundo párrafo del apartado 7, tamaño final: 1.855 palabras.
- `GoogleTermsOfService`: Completo, tamaño final: 1.662 palabras.
- `MSNTermsOfService`: Truncado a partir de la lista de acciones no permitidas, tamaño final: 1.670 palabras.
- `UN-ConventionSaleGoods`: Truncado a partir del artículo 20, tamaño final: 1.816 palabras.
- `UN-HumanRights`: Completo, tamaño final: 1.746 palabras.

Para construir el resto de *corpora* se utilizaron como "puntos de corte" los mismos que se usaron en los documentos en inglés a fin de garantizar que los textos finales fuesen realmente paralelos. Por otro lado, a fin de aplicar la técnica *blindLight* del mismo modo en todos los idiomas los textos en japonés y hebreo fueron transliterados[1]. Dado que el transliterador de Kanji a Romaji utilizado[2] producía la salida íntegramente en minúsculas todos los textos del resto de idiomas fueron también convertidos a minúsculas.

Así pues, una vez truncados los documentos necesarios en cada idioma, transliterados los documentos en hebreo y japonés y convertidos todos los textos a minúsculas se disponía de siete *corpora* paralelos garantizando además que todos los documentos tenían el mismo tamaño aproximado en un idioma "patrón". De este modo, las posibles diferencias de longitud en el resto de idiomas deberían ser mínimas y sólo podrían ser atribuíbles a efectos de la traducción.

[1] En teoría empleando *Unicode* se podría aplicar la técnica de manera directa sobre los textos en hebreo y japonés. No obstante, dada la distinta naturaleza de los sistemas de escritura se optó por la transliteración (esto es, su representación empleando el alfabeto latino) a fin de usar *n*-gramas del mismo tamaño en todos los idiomas.

[2] `http://www.j-talk.com/nihongo/`

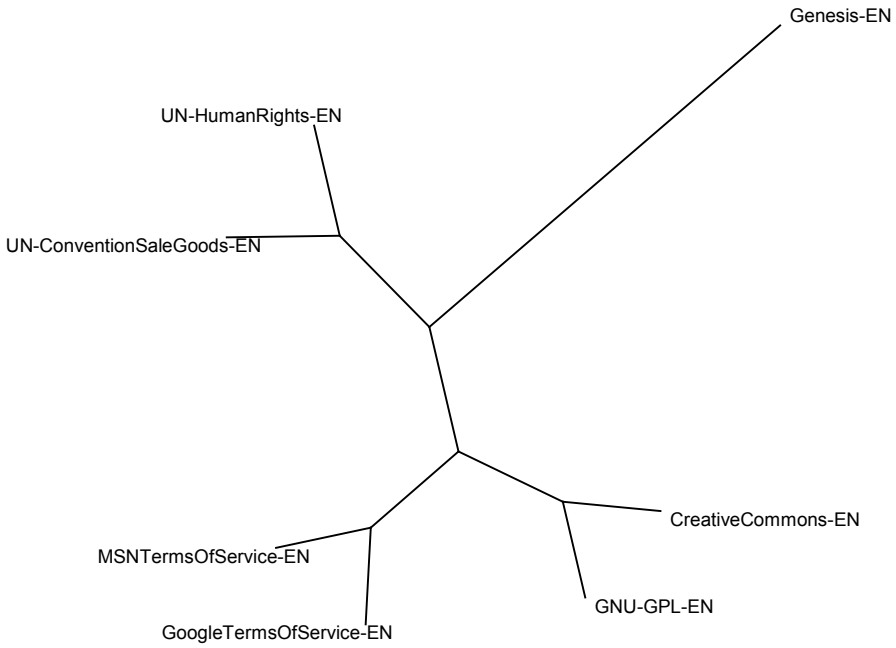



**Fig. 36 Clasificación *blindLight* del corpus de documentos escritos en inglés.**

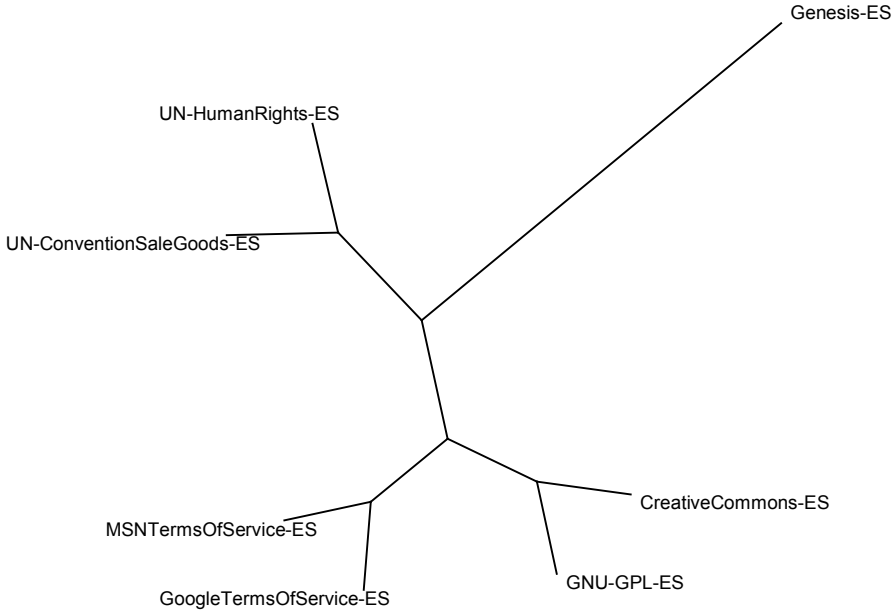



**Fig. 37 Clasificación *blindLight* del corpus de documentos escritos en castellano.**

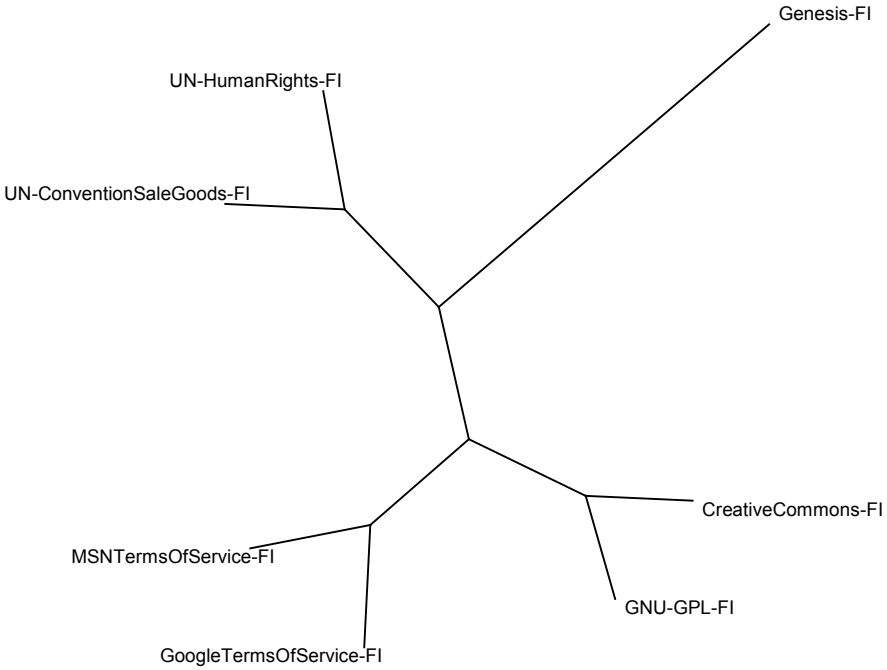



**Fig. 38 Clasificación *blindLight* del corpus de documentos escritos en finés.**

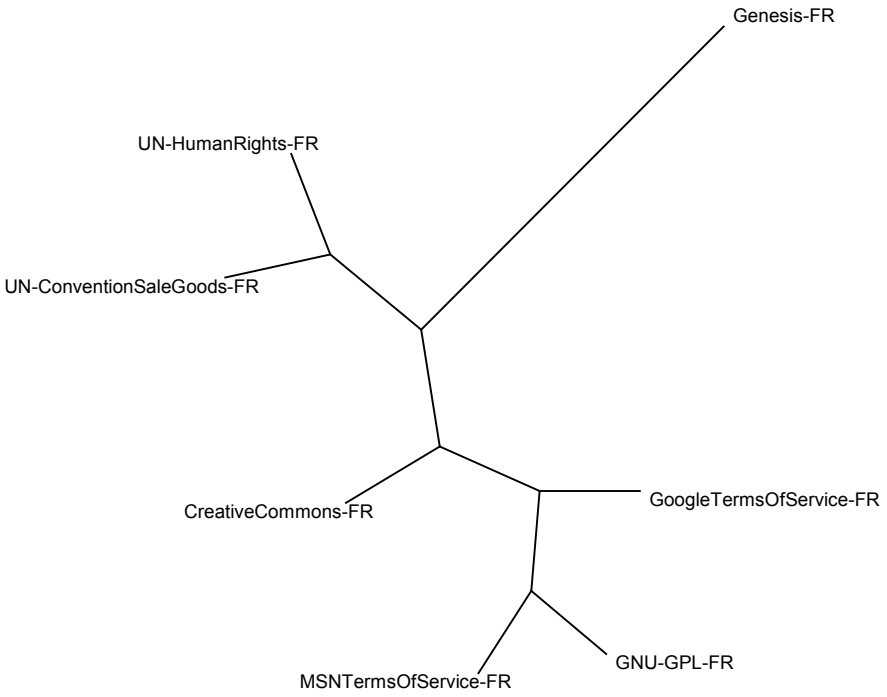



**Fig. 39 Clasificación *blindLight* del corpus de documentos escritos en francés.**

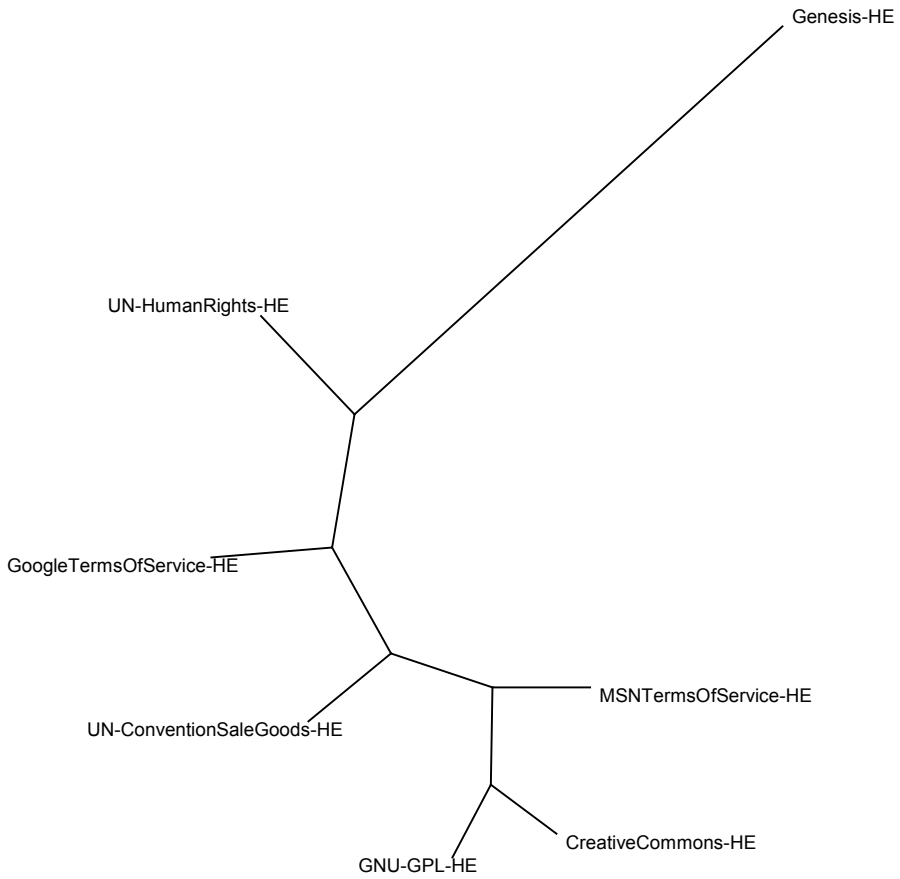



**Fig. 40 Clasificación *blindLight* del corpus de documentos escritos en hebreo.**

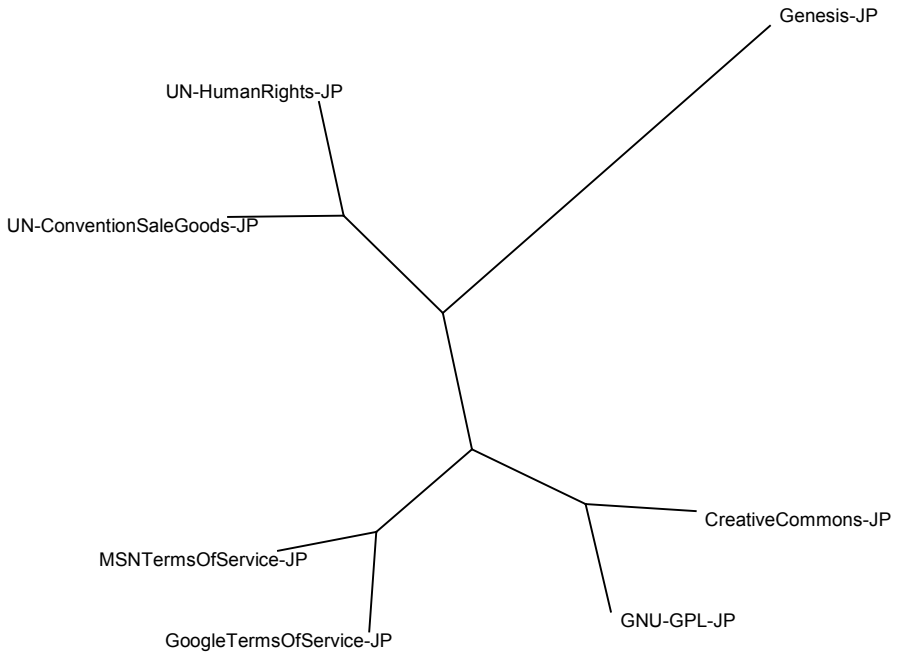



**Fig. 41 Clasificación *blindLight* del corpus de documentos escritos en japonés.**

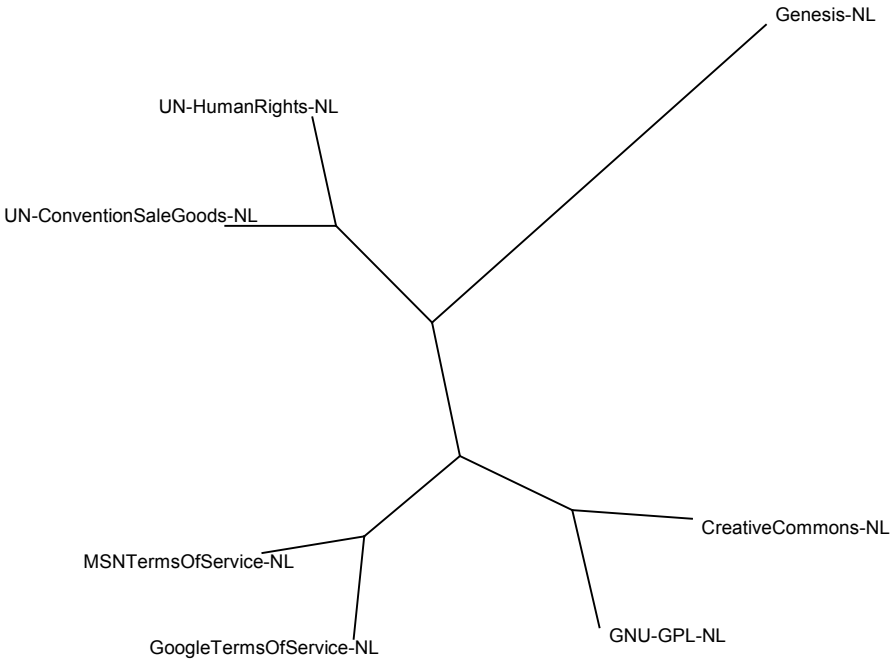
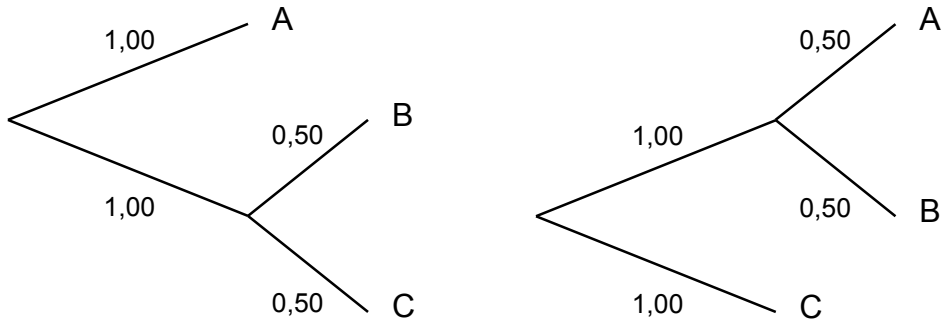



**Fig. 42 Clasificación *blindLight* del corpus de documentos escritos en holandés.**

Como se puede observar en las figuras Fig. 36 a Fig. 42 que muestran las distintas clasificaciones éstas son topológicamente equivalentes para todos los *corpora* a excepción del francés y del hebreo. Sin embargo, un simple parecido no es suficiente y es necesario evaluar numéricamente la similitud (o su ausencia) entre los distintos árboles obtenidos. Para ello se han empleado dos medidas de comparación entre árboles, la primera es la distancia *branch score* descrita por Mary Kuhner y Joseph Felsenstein (1994, p. 461) y la segunda, propuesta por el autor, está basada en el coeficiente de correlación de Spearman.




**Árbol 1** **Árbol 2**

**Fig. 43 La distancia branch score entre estos árboles es 0,5 y el coeficiente de correlación -0,25.**

A fin de explicar el cálculo de ambas medidas se emplearán los árboles mostrados en la Fig. 43. Para el cálculo de la distancia *branch score* hay que determinar, en primer lugar, todas las posibles particiones que se pueden establecer en cada árbol teniendo en cuenta que no existe ningún orden pre-establecido entre las distintas ramas.

Así, para los dos árboles del ejemplo es posible obtener las particiones `{A|B,C}`, `{A,B|C}` y `{A,C|B}` (véase Fig. 44). Naturalmente habrá casos más complejos donde no

todas las particiones posibles en un árbol sean posibles en el otro. Seguidamente, se determina la longitud de la rama de cada partición teniendo en cuenta que si una partición no existiese en un árbol se le asignaría longitud cero. En este caso particular las particiones tendrían las longitudes que se muestran en Fig. 45.

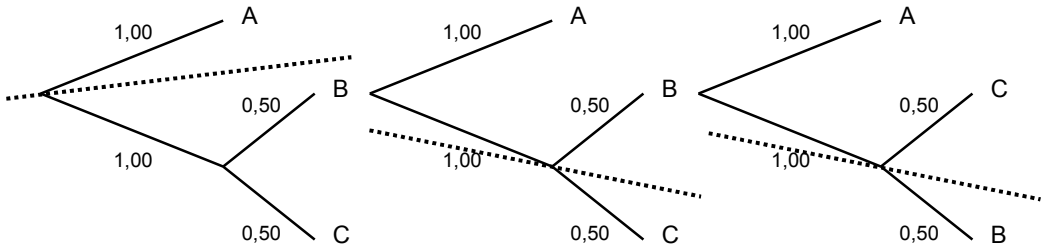



**Fig. 44 Particiones del primer árbol {A | B,C}, {A,B | C} y {A,C | B} con distancias 1.00, 0.50 y 0.50 respectivamente.**

Posteriormente, se calcula el cuadrado de las diferencias de las longitudes de cada partición y se suman. Así, en este ejemplo la distancia *branch score* sería de 0,50. Esta distancia es 0,00 para árboles idénticos y aumenta a medida que los árboles difieren entre sí. Sin embargo, depende del tamaño de los árboles comparados de tal manera que no es posible comparar distancias para parejas de árboles diferentes (Kuhner y Felsenstein, p. 461). Por esa razón el autor ha optado por utilizar una *branch score* "normalizada" al dividir la raíz cuadrada de la puntuación obtenida entre la suma de las longitudes de todas las ramas de los dos árboles comparados. En el ejemplo anterior la puntuación normalizada sería de 0,2357.

| | Árbol 1 | Árbol 2 |
|---|---|---|
| **{A \| B, C}** | 1,00 | 0,50 |
| **{A, B \| C}** | 0,50 | 1,00 |
| **{A, C \| B}** | 0,50 | 0,50 |

**Fig. 45 Longitudes de las particiones en cada árbol.**

La medida propuesta por el autor, por su parte, se basa en el coeficiente de Spearman de correlación entre listas ordenadas de elementos. La idea es simple, dados dos árboles que contienen los mismos nodos es posible definir en ambos dos listas idénticas de posibles parejas. En el ejemplo utilizado hasta el momento la lista sería `{AB,AC,BC}`. Para cada árbol se puede obtener la distancia que hay que "recorrer" para ir de un elemento de cada pareja al otro (véase Fig. 46).

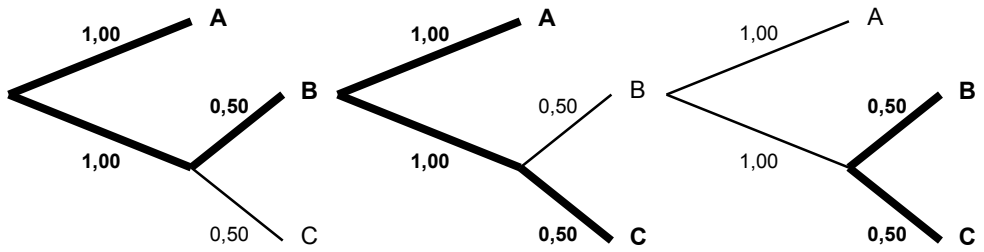



**Fig. 46 Distancias "sobre" el árbol entre A y B (2.50), A y C (2.50) y B y C (1.00).**

Una vez obtenidas las distancias para cada pareja se asigna a cada una un *ranking* en cada árbol y se procede a la determinación del correspondiente coeficiente de Spearman (véase Fig. 47). En el caso del ejemplo el coeficiente obtenido es -0,125 lo cual significa que apenas hay correlación entre ambos árboles.

| | Distancia "sobre" A1 | *Ranking* A1 | Distancia "sobre" A2 | *Ranking* A2 | Diferencia *rankings* | Cuadrado diferencias |
|---|---|---|---|---|---|---|
| **AB** | 2,50 | 2,5 | 1,00 | 1 | 1,5 | 2,25 |
| **AC** | 2,50 | 2,5 | 2,50 | 2,5 | 0 | 0 |
| **BC** | 1,00 | 1 | 2,50 | 2,5 | 1,5 | 2,25 |
| | | | | Suma cuadrados diferencias | | 4,50 |
| | | | | Número de parejas (n) | | 3 |

Coeficiente de correlación $1-\frac{6\sum_{i=1}^{n}(r_i^1-r_i^2)^2}{n^3-n}=1-\frac{6\cdot 4{,}50}{3^3-3}=-0{,}125$

**Fig. 47 Cálculo del coeficiente de correlación de Spearman para lo árboles 1 y 2. Obsérvese que, en caso de empate, el *ranking* asignado es la media de los *rankings* teóricos.**

La principal ventaja de esta medida de comparación es que sus valores varían entre -1 y 1, significando estos valores extremos una correlación perfecta (negativa o positiva) y los valores próximos a cero la ausencia de correlación. Por otro lado, puesto que las distancias entre los elementos de la pareja se miden "sobre" el árbol no se pierde totalmente la información topológica.

Las similitudes entre las clasificaciones para los distintos *corpora* determinadas mediante ambas medidas son las que se muestran en Fig. 48. Como se puede ver, existen varias parejas de clasificaciones en las que tanto la medida *branch score* como el coeficiente de Spearman coinciden en señalar una enorme similitud (p.ej. las formadas por inglés, español, finés y holandés). No obstante, también existen otras en las que sólo una de las medidas señala una similitud más o menos importante; sin embargo, en tales casos siempre interviene alguno de los siguientes idiomas: francés, hebreo o japonés.

| | BRANCH SCORE | | SPEARMAN | | |
|---|---|---|---|---|---|
| | *Branch Score* | *Branch Score* normalizada | Coef. Correlación | Fiabilidad | Correlación |
| **EN-ES** | 0,26 | 0,015 | 0,98 | 99% | Casi perfecta |
| **EN-FI** | 0,43 | 0,019 | 0,97 | 99% | Casi perfecta |
| **EN-FR** | 6,63 | 0,076 | 0,93 | 99% | Muy fuerte |
| EN-HE | 114,44 | 0,184 | 0,75 | 99% | Fuerte |
| EN-JP | 2,63 | 0,053 | 0,71 | 99% | Fuerte |
| **EN-NL** | 0,19 | 0,012 | 0,97 | 99% | Casi perfecta |
| **ES-FI** | 0,50 | 0,022 | 0,90 | 99% | Muy fuerte |
| **ES-FR** | 5,65 | 0,073 | 0,90 | 99% | Muy fuerte |
| ES-HE | 116,67 | 0,191 | 0,77 | 99% | Fuerte |
| ES-JP | 1,60 | 0,044 | 0,75 | 99% | Fuerte |
| **ES-NL** | 0,40 | 0,019 | 1,00 | 99% | Perfecta |
| **FI-FR** | 6,33 | 0,076 | 0,92 | 99% | Muy fuerte |
| FI-HE | 116,27 | 0,188 | 0,68 | 99% | Fuerte |
| FI-JP | 1,44 | 0,041 | 0,61 | 99% | Fuerte |
| **FI-NL** | 0,21 | 0,014 | 0,90 | 99% | Muy fuerte |
| FR-HE | 122,14 | 0,194 | 0,69 | 99% | Fuerte |
| FR-JP | 6,27 | 0,086 | 0,61 | 99% | Fuerte |
| **FR-NL** | 6,58 | 0,076 | 0,91 | 99% | Muy fuerte |
| HE-JP | 125,29 | 0,210 | 0,41 | Rechazar | Inexistente |
| HE-NL | 115,56 | 0,185 | 0,76 | 99% | Fuerte |
| JP-NL | 2,04 | 0,047 | 0,75 | 99% | Fuerte |

**Fig. 48 Medidas de la similitud entre las clasificaciones obtenidas para los distintos *corpora*.**

Se muestran con sombreado gris aquellas clasificaciones con mayor similitud de acuerdo a la distancia *branch score* normalizada y en negrita aquellas con mayor grado de similitud de acuerdo al coeficiente de correlación de Spearman.

Así, los resultados obtenidos con los documentos en japonés son muy similares a los obtenidos con el resto de idiomas de acuerdo con la medida *branch score* y no tan parecidos aplicando Spearman. En cambio, los resultados para el francés no son topológicamente parecidos (*branch score*) pero sí resultan muy similares mediante Spearman. Y por lo que respecta al hebreo, es el idioma cuyos resultados parecen estar menos relacionados con los obtenidos para otras lenguas.

En el caso del hebreo y japonés la principal razón para la menor correlación (a pesar de una obvia semejanza topológica en el caso del japonés) es, muy probablemente, el uso por parte de ambos de un sistema de escritura muy diferente al empleado por el resto de idiomas comparados obligando a una transliteración de los textos.

Así, el hebreo utiliza un alfabeto de 22 letras. Salvo en unos pocos casos especiales[1] puede decirse que el alfabeto permite representar tan sólo consonantes. Para indicar los sonidos vocálicos se utiliza un sistema de marcas (fundamentalmente puntos) situados en las proximidades de cada letra. Este sistema, conocido como *niqqud* o *nikkud*, también permite señalar qué sílaba de la palabra lleva el acento. La utilización del *niqqud* es opcional y la mayor parte de los textos escritos en hebreo, en particular los utilizados en este experimento, utilizan únicamente consonantes. Este hecho puede tener una gran influencia al comparar los resultados obtenidos para el hebreo con los de otras lenguas. Mientras que los textos empleados para el resto de idiomas recogen mucha información acerca de la vocalización de cada texto[2] en el caso de los documentos escritos en hebreo esa información simplemente no existe.

Por otro lado, dos de los documentos utilizados no se correspondían con la fidelidad necesaria al resto de traducciones utilizadas. El primer caso es el de la versión hebrea de la *"Convención de las Naciones Unidas sobre los contratos de compraventa internacional de mercaderías"* para la que no se pudo localizar ninguna versión del documento que incluyese el preámbulo (115 palabras en la versión inglesa). El segundo es el documento con las condiciones de uso de *Google* en su versión para Israel. Los dos últimos párrafos del apartado *"Renuncia a garantías"* (תעודת אחריות) no existen. El apartado de *"Limitación de responsabilidad"* (האחריות מגבלו) está vacío. En el apartado *"Solicitud de eliminación de vínculos o materiales en caché"* (בקשה להסרת קישורים או חומרי מאוחסנים) tan sólo aparece el primer párrafo pero no se citan los principios que emplea *Google* para decidir sobre las solicitudes de eliminación. El apartado *"Condiciones varias"* (תנאים שונים) también está vacío. En definitiva, 899 palabras de las 1662 del documento original en inglés no aparecen en la traducción.

En resumen, la versión hebrea del documento de la Convención de Viena sobre compraventa de mercaderías carece de, aproximadamente, un 6% del texto original y el correspondiente a las Condiciones de Servicio de Google de un 54%. Este hecho, unido a la ausencia de información vocálica, sin duda tiene un impacto en la taxonomía finalmente obtenida y en su falta de concordancia con el resto de idiomas estudiados.

Por lo que se refiere a los documentos escritos en japonés, hasta donde ha podido comprobar el autor, las traducciones son razonablemente fieles. Por otro lado, hay que

[1] Ciertas letras son mudas o vocales dependiendo de su combinación con sonidos vocálicos o al aparecer al final de palabra.

[2] El español escrito, por ejemplo, recoge de forma casi total la vocalización y acentuación de las palabras. Otros idiomas recogen mucha información "fonética" pero no toda la necesaria. Pensemos en el inglés con sus *bough* /baʊ/, *bought* /bɔːt/, *cough* /kɒf/, *though* /ðəʊ/ o *through* /θruː/.

señalar que ciertas características del japonés escrito y de su pronunciación resultan difíciles de manejar en su transliteración y, por tanto, han podido influir en los resultados finales.

En primer lugar, el japonés escrito puede emplear hasta cuatro tipos distintos de caracteres: *kanji*, *hiragana*, *katakana* y *rōmaji*. Los primeros son caracteres chinos adaptados a la escritura de sustantivos, adjetivos, verbos y nombres propios japoneses. *Hiragana* y *katakana* son silabarios utilizándose el segundo en particular para transcribir palabras y nombres no japoneses. Por último, el *rōmaji* emplea caracteres latinos y uno de sus posibles usos es la transliteración de los anteriores sistemas de escritura.

La transliteración de los silabarios *hiragana* y *katakana* no supone mayor problema. Sin embargo, la del *kanji* resulta más compleja puesto que un único carácter puede tener varias "lecturas" en función de la palabra de la que forma parte y, por tanto, del contexto (véase Fig. 49). El transliterador utilizado emplea el léxico del proyecto *EDICT*[1] de Jim Breen para segmentar el texto y proporciona, por tanto, una única romanización que, presumiblemente, es la más adecuada para cada contexto. Sin embargo, el autor no es en absoluto un experto en japonés por lo que no se puede asegurar que los documentos transliterados de manera automática estén totalmente libres de errores que podrían haber influido en los resultados finales.

| 米 | Hiragana Katakana | Pronunciación | Tipo pronunciación |
|---|---|---|---|
| | ベイ | bei | on'yomi |
| | マイ | mai | |
| | メエトル | meetoru | |
| | こめ | kome | kun'yomi |
| | よね | yone | |
| | は | ha | nanori |
| | べ | be | |
| | まべ | mabe | |
| | め | me | |
| | よ | yo | |
| | よな | yona | |
| | よの | yono | |
| | よま | yoma | |

| 国 | Hiragana Katakana | Pronunciación | Tipo pronunciación |
|---|---|---|---|
| | コク | koku | on'yomi |
| | くに | kuni | kun'yomi |
| | くな | kuna | nanori |
| | こ | ko | |

**Fig. 49 Distintas lecturas de los caracteres de la "palabra" japonesa 米国 (べいこく, *beikoku*, EEUU).**

Los caracteres japoneses poseen tres tipos distintos de pronunciaciones: *on'yomi* (pronunciación china), *kun'yomi* (pronunciación japonesa) y *nanori*. La primera es la pronunciación aproximada del carácter chino original. *Kun'yomi* es la pronunciación de una palabra japonesa equivalente al carácter chino importado. *Nanori* es la forma en que puede pronunciarse el *kanji* cuando se utiliza dentro de un nombre propio. Un transliterador de japonés debe determinar en qué contexto se está utilizando un carácter a fin de proporcionar su pronunciación correcta; esto habitualmente se hace empleando diccionarios que recogen la versión *kanji* y el correspondiente *hiragana/katakana* para cada entrada.

[1] http://www.csse.monash.edu.au/~jwb/j_edict.html

Por lo que respecta al francés se ha realizado un análisis *post mortem* de los documentos a fin de encontrar causas para las discrepancias, en particular con los idiomas indoeuropeos. Tras ese análisis se ha encontrado lo siguiente:

- Condiciones de Uso de *Google*: Unas pocas partes del texto original en inglés no se incluyen en la versión francesa, no obstante se trata de una traducción razonablemente fiel.
- Creative Commons: los puntos e) y f) del apartado 4 (*Restricciones*) no aparecen en la versión francesa de la licencia (232 palabras en el original en inglés). El apartado 5 (*Renuncia de responsabilidades*) presenta una redacción totalmente diferente (83 palabras en el original). De este modo, un 13% del texto original no aparece en la versión en francés y un 5% ha sido redactado de forma totalmente diferente adquiriendo un mayor peso dentro del documento traducido (11%). En resumen, la licencia Creative Commons en francés no es una traducción fiel de la licencia en inglés sino una adaptación.
- El resto de documentos (Condiciones de Uso de *MSN*, *GNU General Public License*, Génesis, Convención de Viena y Declaración Universal de Derechos Humanos) son traducciones fieles.

Las diferencias entre la clasificación de los documentos franceses y las obtenidas para el resto de idiomas tal vez puedan atribuirse a las diferencias perceptibles que se han descrito entre la licencia *CCPL* en su versión francesa.

No obstante, a pesar de lo señalado para los documentos en francés, hebreo y japonés, puede concluirse que al clasificar un conjunto de *corpora* paralelos utilizando *blindLight* se obtienen de manera sistemática clasificaciones idénticas o muy similares con independencia de la familia lingüística y la longitud de los documentos[1]. Dichas clasificaciones, además, son plausibles según criterios humanos puesto que tienden a agrupar documentos de contenido semejante y mantener separados textos de temática muy distinta en todos los idiomas.

El autor considera que tales resultados permiten sustentar lo que se afirmaba al comienzo del apartado, a saber, que al aplicar la técnica aquí descrita sobre texto natural se obtienen vectores que conservan ciertos aspectos de la semántica latente en los documentos originales permitiendo una comparación a un nivel conceptual.

En los siguientes capítulos se presentará la aplicación de la técnica a la clasificación y categorización de documentos, a la recuperación de información y a la obtención de resúmenes automáticos concluyendo de ese modo la demostración de la tesis del autor.

[1] Tan sólo en inglés los documentos tienen aproximadamente la misma longitud en *bytes*, en el resto de idiomas los documentos difieren en tamaño.