

# Práctica 1

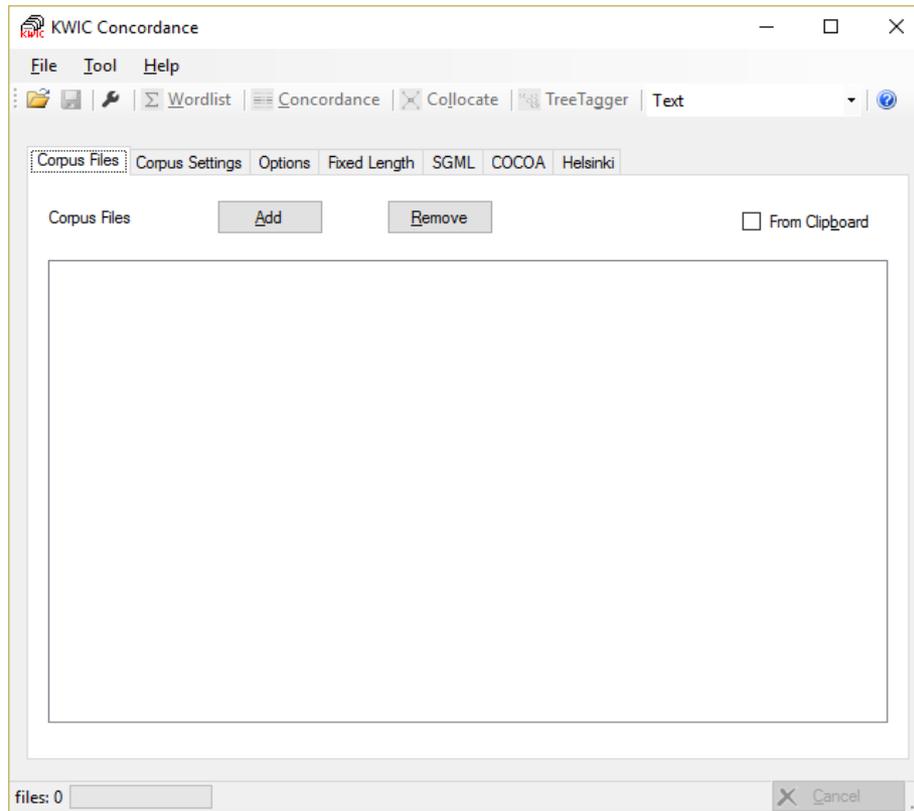
Herramientas de análisis del corpus  
Extracción de terminología

# KWIC Concordance for Windows

# KWIC Concordance for Windows

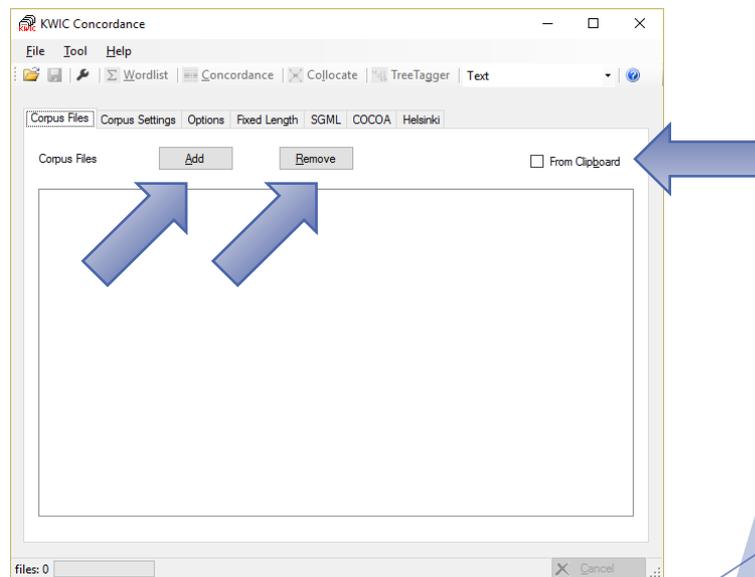
- ▶ Herramienta software lingüística
  - ▶ Listas de frecuencia de palabras
  - ▶ Concordancias
  - ▶ Colocaciones
  - ▶ N-gramas
  - ▶ Etiquetado gramatical (*part-of-speech*, POS) y lemas
- ▶ Sistema operativo Windows, licencia *freeware*
- ▶ Versión 5.3 (septiembre de 2016)
- ▶ [http://www.chs.nihon-u.ac.jp/eng\\_dpt/tukamoto/kwic\\_e.html](http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html)

# Pantalla principal de KWIC Concordance



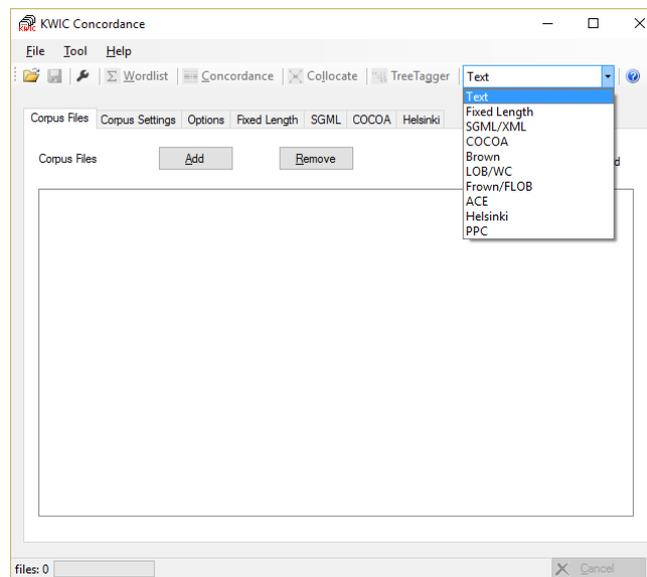
# Añadir/eliminar ficheros del corpus

- ▶ La herramienta permite trabajar con un único fichero o con varios simultáneamente
- ▶ Pestaña “Corpus Files”:
  - ▶ Permite añadir o eliminar los ficheros
  - ▶ Permite utilizar el portapapeles de Windows
  - ▶ Usar codificación Unicode mejor que ASCII



# Formato del fichero de entrada

- ▶ La herramienta es compatible con numerosos formatos
- ▶ Pestañas específicas para las opciones de los formatos “Fixed Length”, SGML, COCOA y Helsinki
- ▶ Información detallada: [http://www.chs.nihon-u.ac.jp/eng\\_dpt/tukamoto/input\\_e.html](http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/input_e.html)

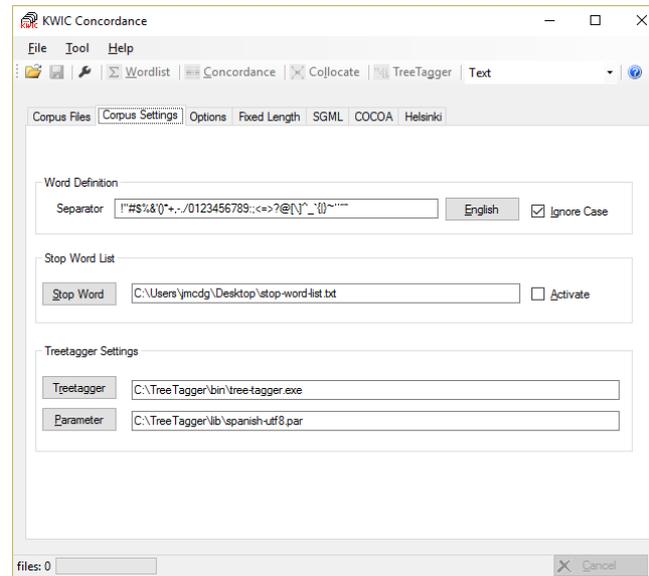


# Resumen de formatos

- ▶ *Text*: Fichero de texto plano sin ninguna referencia
- ▶ *Fixed Length*: Fichero con referencia en posición fija
- ▶ *SGML*: Fichero SGML (lenguaje de marcado generalizado estándar) con etiquetas del tipo <A> ... </A>, <B> ... </B>
- ▶ *COCOA*: Fichero con etiquetas del tipo <A >, <B b>
- ▶ *Brown*: Fichero de tipo *Brown Corpus (Fixed Length)*
- ▶ *LOB/WC*: Fichero de tipo Lancaster-Oslo-Bergen/Wellington Corpus (*Fixed Length*)
- ▶ *Frown/FLOB*: Fichero de tipo Freiburg-Brown Corpus/Freiburg-LOB Corpus (*Fixed Length*)
- ▶ *ACE*: Fichero de tipo Australian Corpus of English (*SGML*)
- ▶ *Helsinki*: Fichero de tipo Helsinki Corpus
- ▶ *PPC*: Fichero de tipo Penn-Helsinki Parsed Corpus

# Configuración

- ▶ Pestaña “Corpus Settings”:
  - ▶ Definición de caracteres que NO son constituyentes de palabras
  - ▶ Uso de listas de *Stop Words*
  - ▶ Configuración del programa TreeTagger para etiquetado gramatical (POS) y lemas

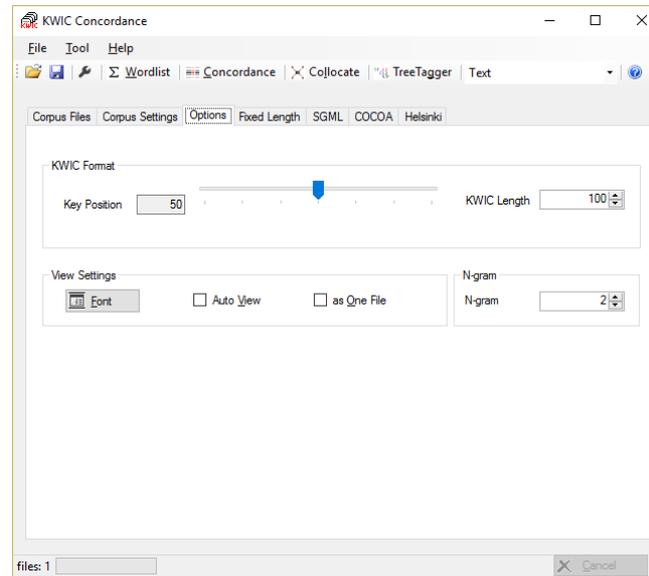


# Configuración de TreeTagger

- ▶ Página web de la herramienta: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>
- ▶ Pasos para la configuración:
  - ▶ Descargar la versión correspondiente (en nuestro caso Windows)
  - ▶ Extraer el fichero comprimido
  - ▶ Descargar el parámetro para TreeTagger del idioma deseado (*<idioma>-utf8.par*)
    - ▶ Moverlo a la carpeta *TreeTagger/lib*
  - ▶ Conectar TreeTagger con KWIC Concordance en la pestaña “Corpus Settings”
    - ▶ *TreeTagger*: ruta del fichero *tree-tagger.exe*
    - ▶ *Parameter*: ruta del fichero *<idioma>-utf8.par*

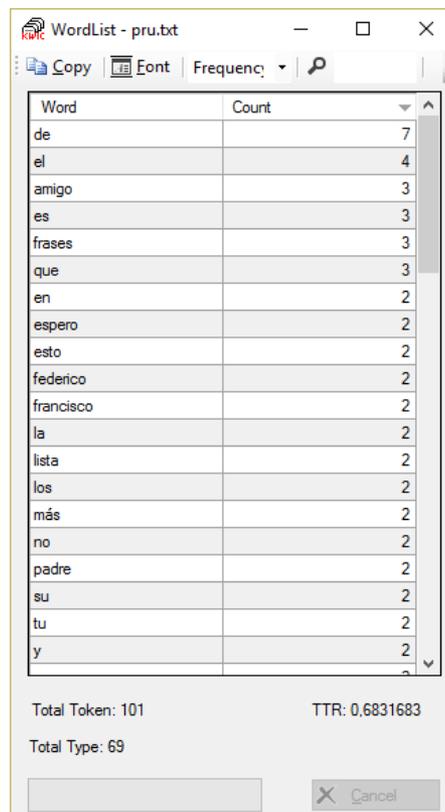
# Opciones

- ▶ Pestaña “Options”:
  - ▶ Formato de KWIC
    - ▶ Posición de la clave
    - ▶ Longitud del contexto
  - ▶ Opciones de visualización
    - ▶ Fuente
    - ▶ Ver fichero original desde la ventana de resultados de la concordancia
    - ▶ Utilizar un corpus con varios ficheros como uno solo
  - ▶ Número de elementos del n-grama (2-5)



# Lista de frecuencia de palabras

- ▶ Este comando devuelve la lista de palabras que conforman el corpus
- ▶ Puede ordenarse de acuerdo a diversos criterios:
  - ▶ Orden alfabético (directo e inverso)
  - ▶ Orden alfabético desde el final de la palabra
  - ▶ Frecuencia de aparición



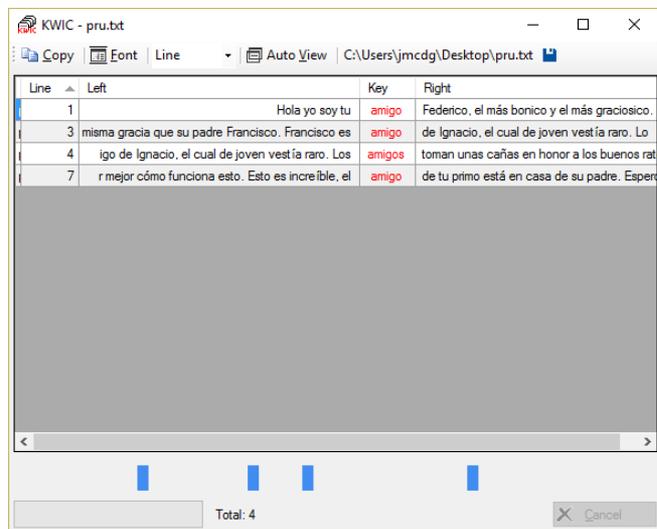
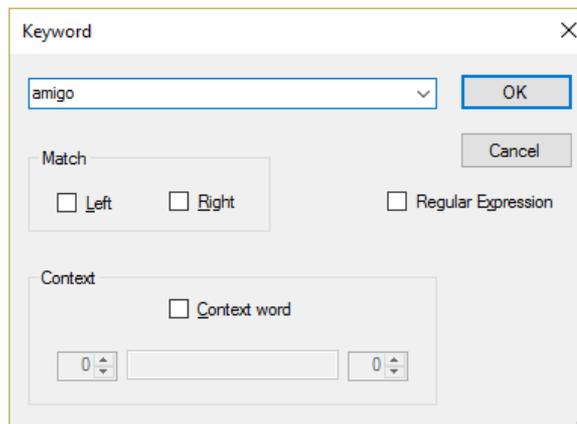
The screenshot shows a window titled "WordList - pru.txt" with a menu bar containing "Copy", "Font", "Frequenc:", and a search icon. Below the menu bar is a table with two columns: "Word" and "Count". The table lists 20 words and their corresponding counts. At the bottom of the window, there are statistics: "Total Token: 101", "TTR: 0,6831683", and "Total Type: 69". A "Cancel" button is visible in the bottom right corner.

Word	Count
de	7
el	4
amigo	3
es	3
frases	3
que	3
en	2
espero	2
esto	2
federico	2
francisco	2
la	2
lista	2
los	2
más	2
no	2
padre	2
su	2
tu	2
y	2

Total Token: 101      TTR: 0,6831683  
Total Type: 69

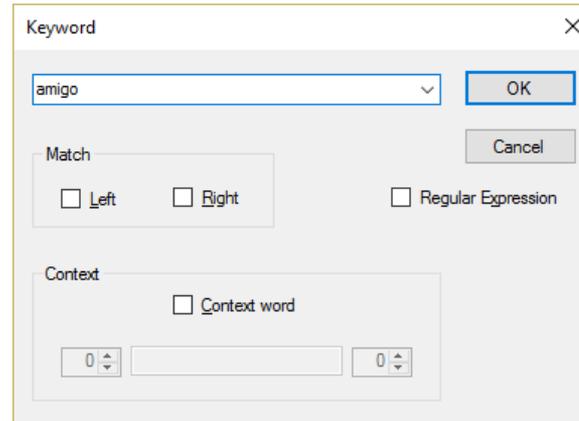
# Concordancia

- ▶ Este comando devuelve las ocurrencias de un patrón de búsqueda en contexto dentro del corpus
- ▶ El resultado es tipo *Key Word In Context* (KWIC)
  - ▶ Las ocurrencias del patrón en el centro de la pantalla
  - ▶ Se puede ordenar de varias maneras



# Colocaciones

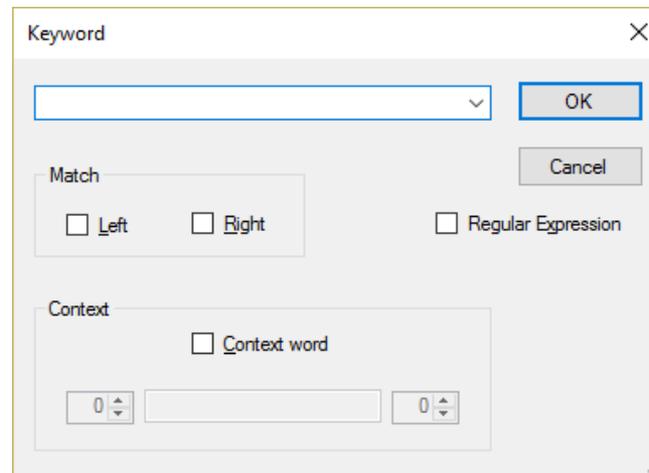
- ▶ Este comando muestra las colocaciones léxicas del patrón de búsqueda
- ▶ Devuelve una tabla que incluye la posición de las palabras en relación al patrón (L5-R5)
  - ▶ Nos da información sobre la co-ocurrencia de las palabras



Word	Count	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
de	3	0	0	0	0	0	amigo	2	0	0	0	1
el	3	0	0	0	0	1	amigo	0	1	1	0	0
en	2	0	0	0	0	0	amigo	0	0	0	1	1
es	2	0	0	1	0	1	amigo	0	0	0	0	0
francisco	2	0	0	1	1	0	amigo	0	0	0	0	0
tu	2	0	0	0	0	1	amigo	0	1	0	0	0
bonico	1	0	0	0	0	0	amigo	0	0	0	1	0
cañas	1	0	0	0	0	0	amigo	0	0	1	0	0
cual	1	0	0	0	0	0	amigo	0	0	0	1	0
está	1	0	0	0	0	0	amigo	0	0	0	1	0
esto	1	0	1	0	0	0	amigo	0	0	0	0	0
fedenco	1	0	0	0	0	0	amigo	1	0	0	0	0
hola	1	0	1	0	0	0	amigo	0	0	0	0	0
honor	1	0	0	0	0	0	amigo	0	0	0	0	1
ignacio	1	0	0	0	0	0	amigo	0	1	0	0	0
increíble	1	0	0	0	1	0	amigo	0	0	0	0	0
los	1	0	0	0	0	1	amigo	0	0	0	0	0
más	1	0	0	0	0	0	amigo	0	0	1	0	0

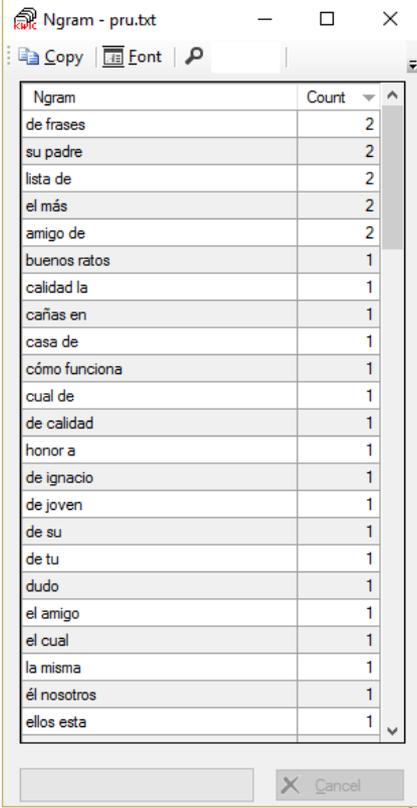
# Diálogo “Keyword”

- ▶ Especifica el patrón para la concordancia y la tabla de colocación
  - ▶ Coincidencia exacta a la derecha y/o izquierda
  - ▶ Uso de expresiones regulares
  - ▶ Establecimiento de restricciones sobre la co-ocurrencia del patrón con determinada palabra



# N-grama

- ▶ Este comando devuelve los n-gramas que pueden hallarse en el corpus
- ▶ El resultado es una lista de n-gramas con su frecuencia de aparición
  - ▶ Un n-grama es una subsecuencia de  $n$  elementos en una secuencia dada
  - ▶ El valor “n” puede establecerse en la pestaña de opciones



The screenshot shows a window titled "Ngram - pru.txt" with a menu bar containing "Copy", "Font", and a search icon. Below the menu is a table with two columns: "Ngram" and "Count". The table lists various n-grams and their corresponding counts. At the bottom of the window, there is a "Cancel" button.

Ngram	Count
de frases	2
su padre	2
lista de	2
el más	2
amigo de	2
buenos ratos	1
calidad la	1
cañas en	1
casa de	1
cómo funciona	1
cual de	1
de calidad	1
honor a	1
de ignacio	1
de joven	1
de su	1
de tu	1
dudo	1
el amigo	1
el cual	1
la misma	1
él nosotros	1
ellos esta	1

# TreeTagger

- ▶ Este comando muestra las anotaciones gramaticales y lemas
- ▶ El resultado son tres tablas
  - ▶ POS: lista de términos, etiquetado y lema
  - ▶ POSList: lista de etiquetas y frecuencia
  - ▶ LemmaList: lista de lemas, frecuencia y variantes

The screenshot shows three overlapping windows from the TreeTagger application. The top window, 'POS - pru.txt', displays the original text with grammatical annotations. The middle window, 'LemmaList - pru.txt', shows the lemmas and their frequencies. The bottom window, 'POSList - pru.txt', shows the part-of-speech tags and their counts.

Word	POS	Count
Hola		
yo	ADJ	10
soy	ADV	4
tu	ART	9
amigo	CC	2
Federico	CCAD	1
.	CM	8
el	CQUE	3
más	CSUBI	1
bonico	DM	4
y	FS	10
el	NC	19
más	NEG	2
gracioso	NP	6
.	PPC	1
Federico	PPO	4
es	PPX	8
gracioso	PREP	10
.	REL	1
pero	VEfin	1
no	VLfin	10
tiene		
la		

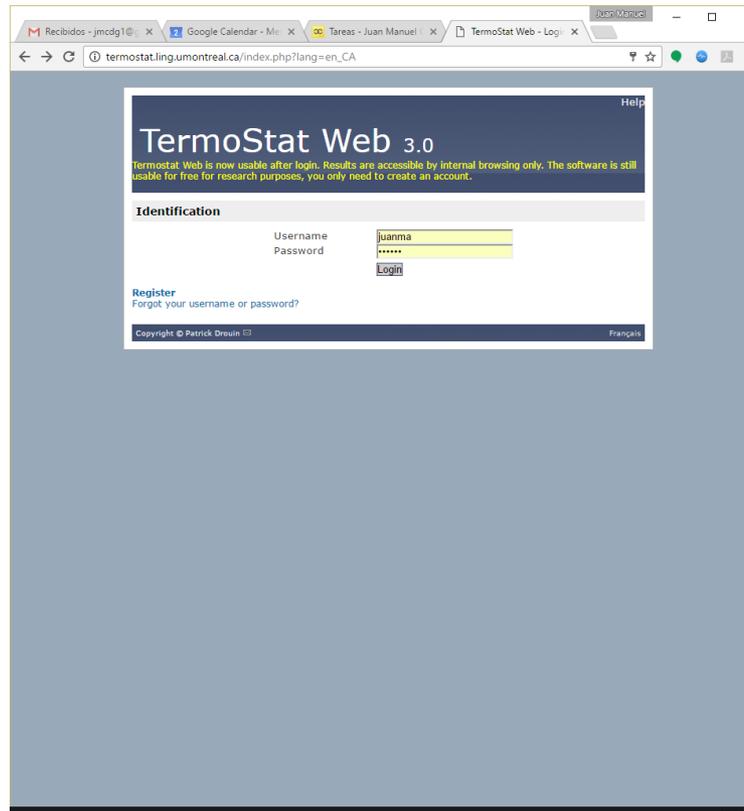
Summary statistics from the POSList window:  
Total Token: 119  
Total Type: 22  
TTR: 0,184874

# TermoStat

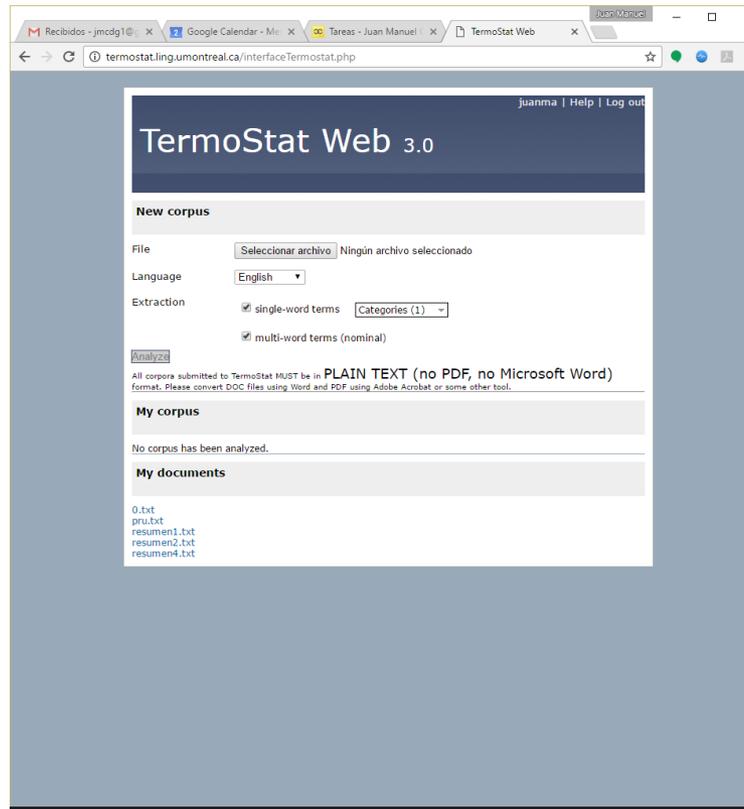
# TermoStat

- ▶ Herramienta de análisis de corpus automática
  - ▶ Listas de frecuencia de palabras
  - ▶ Concordancias
  - ▶ Nube de términos
  - ▶ Ocurrencia de patrones gramaticales
  - ▶ Estructuración
  - ▶ Bigramas
- ▶ Aplicación web, accesible desde cualquier navegador
- ▶ Gratuita (necesario registro previo en el portal)
- ▶ <http://termostat.ling.umontreal.ca/>

# Pantalla de login de ThermoStat



# Pantalla principal



# Lista de términos

Candidate (grouping variant)	Frequency (Specificity)	Score	Variants	Pattern
temperatura	54	179.3	temperatura temperaturas	Common_Noun
calentamiento global	24	162.76	calentamiento global	Common_Noun Adjective
artículo principal	12	155.04	artículo principal	Common_Noun Adjective
calentamiento	39	147.89	calentamiento	Common_Noun
atmósfera	55	144.99	atmósfera atmósferas	Common_Noun
glaciar	12	126.56	glaciar glaciares	Common_Noun
océano	32	125.04	océano océanos	Common_Noun
temperatura superficial	5	101.52	temperatura superficial temperaturas superficiales	Common_Noun Adjective
albedo	5	101.52	albedo	Common_Noun
hielo	18	100.47	hielo hielos	Common_Noun
solar	25	99.07	solar solares	Common_Noun
variación	26	96.19	variación variaciones	Common_Noun
casquete	7	94.63	casquete casquetes	Common_Noun
glaciaciones	5	92.67	glaciaciones	Common_Noun
superficie terrestre	7	90.91	superficie terrestre	Common_Noun Adjective
svg	4	88.28	svg	Common_Noun
actividad solar	4	88.28	actividad solar	Common_Noun Common_Noun
corriente oceánico	4	88.28	corrientes oceánicas	Common_Noun Adjective
vapor de agua	4	88.28	vapor de agua	Common_Noun Preposition
gei	4	88.28	gei	Common_Noun
molécula	9	80.96	molécula moléculas	Common_Noun
átomo	8	80.6	átomo átomos	Common_Noun
latitud	9	79.55	latitud latitudes	Common_Noun
nube	12	73.55	nubes	Common_Noun
viento solar	3	72.81	viento solar	Common_Noun Common_Noun
paradoja del sol	2	72.81	paradoja del sol	Common_Noun Preposition

# Concordancias

**Contexts**

Sentences KWIC

Índice [ocultar] 1 Terminología 2 Causas de los cambios climáticos 2.1 Influencias externas 2.1.1 Variaciones solares 2.1.2 Variaciones orbitales 2.1.3 Impactos de meteoritos 2.2 Influencias internas 2.2.1 La deriva continental 2.2.2 La composición atmosférica 2.2.3 Las corrientes oceánicas 2.2.4 El campo magnético terrestre 2.2.5 Los efectos antropogénicos 2.2.6 Retroalimentaciones y factores moderadores 2.3 Incertidumbre de predicción 3 Cambios climáticos en el pasado 3.1 La paradoja del Sol débil 3.2 El efecto invernadero en el pasado 3.3 El CO<sub>2</sub> como regulador del clima 3.4 Aparece la vida en la Tierra 3.5 Máximo Jurásico 3.6 Las glaciaciones del Pleistoceno 3.7 El mínimo de Maunder 4 El cambio climático actual 4.1 Combustibles fósiles y calentamiento global 4.2 Planteamiento de futuro 4.3 Agricultura 5 Clima de planetas vecinos 6 Materia multidisciplinar 7 Océanos 7.1 El aumento de la temperatura 7.2 Sumideros de carbono y acidificación 7.3 El cierre de la circulación térmica 8 Impacto en los pueblos indígenas 9 Cultura popular 9.1 Cine 9.2 Información cinematográfica sobre el cambio climático 9.3 Literatura 10 Véase también 11 Referencias 11.1 Notas 11.2 Bibliografía 12 Bibliografía complementaria 13 Enlaces externos Terminología[editar] La definición más general de cambio climático es un cambio en las propiedades estadísticas (principalmente su promedio y dispersión) del sistema climático al considerarse durante periodos largos de tiempo, independiente de la causa.4 Por consiguiente, las fluctuaciones durante periodos más cortos que unas cuantas décadas, como El Niño, no representan un cambio climático.

En las publicaciones científicas, calentamiento global se refiere a aumento de las temperaturas superficiales mientras que cambio climático incluye al calentamiento global y todo lo demás que el aumento de los niveles de gases de efecto invernadero produce.6 La Convención Marco de la Naciones Unidas sobre el Cambio Climático, define al Cambio Climático en su artículo 1, párrafo segundo como un cambio de clima atribuido directa e indirectamente a la actividad humana que altera la composición de la atmósfera y que se suma a la variabilidad natural del clima observadas durante periodos de tiempos comparables.7 Los cambios de clima del Planeta Tierra son de gran preocupación y responsabilidad de todos los seres humanos.

véase también : Calentamiento global#Etimología Causas de los cambios climáticos[editar] Temperatura en la superficie terrestre al comienzo de la primavera de 2000. El clima es un promedio a una escala de tiempo dado del tiempo atmosférico.

Estos factores y sus variaciones en el tiempo producen cambios en los principales elementos constituyentes del clima que también son cinco : temperatura atmosférica, presión atmosférica, vientos, humedad y precipitaciones.

Animación del mapa mundial de la temperatura media mensual del aire de la superficie.

Las investigaciones hechas por algunos científicos apuntan que la razón principal del aumento de temperatura en el Planeta esdeber[debido al proceso de industrialización iniciado hace siglo y medio y, en particular la combustión de cantidades cada vez mayores de petróleo, gasolina y carbón, la tala de árboles y algunos métodos de explotación agrícola.

estas actividades aumentan el volumen de gases de efecto invernadero ( GEI ) en la atmósfera , principalmente de dióxido de carbono , metano y óxido-nitroso.8 Lo anterior, ha provocado que los rayos del Sol queden atrapadas en la atmósfera del Planeta Tierra , provocando así un aumento de temperatura .

La temperatura media de la Tierra depende , en gran medida , del flujo de radiación solar que recibe .

Por una parte , las latitudes en las que se concentra la masa continental ; si las masas continentales están situadas en latitudes bajas habrá pocos glaciares continentales y , en general , temperaturas medias menos extremas.

Las corrientes oceánicas[editar] Artículo principal : Corrientes oceánicas Temperatura del agua en la Corriente del Golfo .

Las corrientes oceánicas , o marinas , son factores reguladores del clima que actúan como moderador , suavizando las temperaturas de regiones como Europa y las costas occidentales de Canadá y Alaska .

Los aerosoles de origen antrópico , especialmente los sulfatos provenientes de los combustibles fósiles ejercen una influencia reductora de la temperatura ( Charlson et al.

Este hecho , unido a la variabilidad natural del clima , sería la causa que explica el " valle " que se observa en el gráfico de temperaturas en la zona central del siglo XX .

**Contexts**

Sentences KWIC

7 Océanos 7.1 El aumento de la temperatura 7.2 Sumideros de carbono y acidificación 7.3 superficiales mientras que cambio climático incluye al

global#Etimología Causas de los cambios climáticos[editar] Temperatura en la superficie terrestre al comienzo de del clima que también son cinco : Animación del mapa mundial de la que la razón principal del aumentos de Tierra , provocando así un aumento de La temperatura media de la Tierra depende , en medias menos extremas . del agua en la Corriente del Golfo de regiones como Europa y las costas ( Charlson et al . en la zona central del siglo XX y favorece la fusión completa de todo polares , llevando el planeta a un entre el ecuador y los Polos . , precipitación , viento , presión ) y otros , hay que tener la de equilibrio era de - 41 ° C en la Tierra sería de - 20 emite radiación a un máximo de 0,48 mucho menor , y remite la radiación superficial es de unos 15 ° C , diaria promedio del aire en casilla meteo Temperatura en esa región . es alta , se favorece su intercambio también . es baja , el CO2 se acumula . era muy superior a la actual y que llegaron a 5 ° C . hace unos 150 años ( siempre dentro medias que son siempre las más altas , desencadenándose la próxima glaciación , o dióxido de carbono , deshielos , superficial hasta 465 ° C , capaz de , ni las bajas temperaturas superficiales que superficiales que alcanzan mínimas de - 86 de los océanos asciende , se vuelve

# Nube de palabras



# Patrones gramaticales

Corpus >> 0 juanma | Help | Log out

## Results

List of terms Cloud Stat Structuration Bigrams

Number of Terms: 855

**Patterns**

**Common\_Noun= 260 (30 %)**  
cambio, efecto, atmósfera, temperatura, año, tierra, gas, calentamiento, clima, océano

**Common\_Noun Adjective= 239 (28 %)**  
cambio climático, efecto invernadero, calentamiento global, artículo principal, superficie terrestre, campo magnético, actividad humano, temperatura medio, temperatura superficial, casquete polares

**Common\_Noun Preposition Common\_Noun= 200 (23 %)**  
dióxido de carbono, capa de hielo, vapor de agua, nivel del mar, flujo de radiación, composición del aire, átomo de oxígeno, página de discusión, concentración de dióxido, aumento del nivel

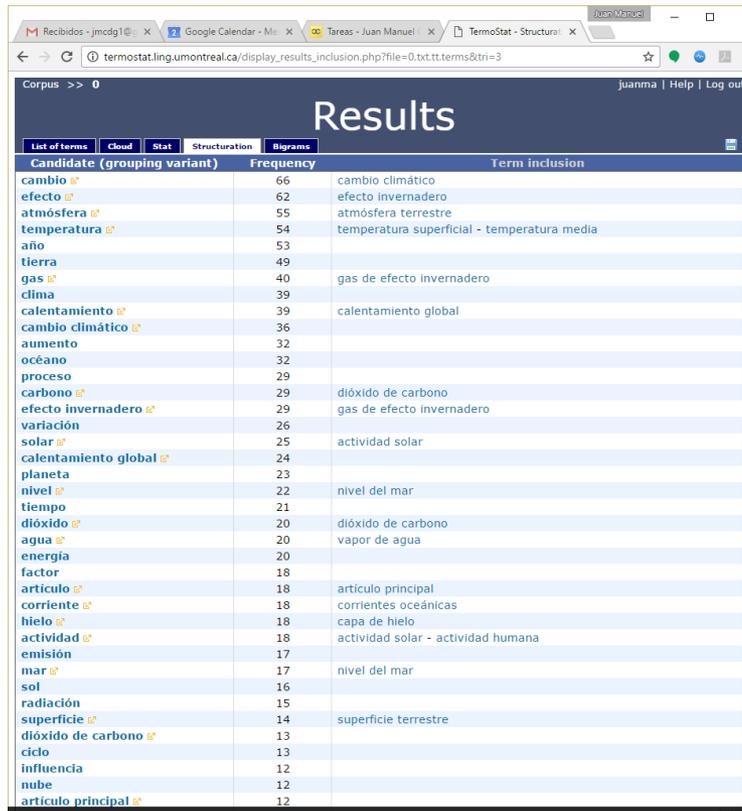
**Common\_Noun Common\_Noun= 66 (8 %)**  
actividad solar, mancha solar, viento solar, combustible fósil, sistema solar, periodo glacial, condición promedio, nuestro planeta, efecto invernadero, radiación solar

**Common\_Noun Preposition Common\_Noun Adjective= 59 (7 %)**  
gas de efecto invernadero, efecto del cambio climático, paradoja del sol débil, tamponamiento del aumento atmosférico, hielo del océano antártico, extracción del anhídrido carbónico, estudio del cambio climático, proceso de desgasificación semejante, clima de planeta vecinos[editar, ausencia del retraso perceptible

**Common\_Noun Adjective Adjective= 23 (3 %)**  
campo magnético terrestre, cambio climático actual, efecto climático actual, variación temporal meteorológica, fenómeno meteorológicos extremo, efecto invernadero intenso, papel regulador fundamental, cambio climático global, periodo cálido medieval, crisis ambiental actual

**Common\_Noun Adjective Coord\_Conjunction Adjective= 8 (1 %)**  
manera gradual y lineal, agente extintores y propelentes, estado vibracional y rotacional, retroalimentaciones positivo y negativo, clima tropical y apto, sistema caótico y complejo, ganadería intensivo y arrozales, implicación técnico y económico

# Estructuración



Candidate (grouping variant)	Frequency	Term inclusion
cambio	66	cambio climático
efecto	62	efecto invernadero
atmósfera	55	atmósfera terrestre
temperatura	54	temperatura superficial - temperatura media
año	53	
tierra	49	
gas	40	gas de efecto invernadero
clima	39	
calentamiento	39	calentamiento global
cambio climático	36	
aumento	32	
océano	32	
proceso	29	
carbono	29	dióxido de carbono
efecto invernadero	29	gas de efecto invernadero
variación	26	
solar	25	actividad solar
calentamiento global	24	
planeta	23	
nivel	22	nivel del mar
tiempo	21	
dióxido	20	dióxido de carbono
agua	20	vapor de agua
energía	20	
factor	18	
artículo	18	artículo principal
corriente	18	corrientes oceánicas
hielo	18	capa de hielo
actividad	18	actividad solar - actividad humana
emisión	17	
mar	17	nivel del mar
sol	16	
radiación	15	
superficie	14	superficie terrestre
dióxido de carbono	13	
ciclo	13	
influencia	12	
nube	12	
artículo principal	12	

# Bigramas

Corpus >> 0 juanma | Help | Log out

## Results

[List of terms](#) [Cloud](#) [Stat](#) [Structuration](#) [Bigrams](#)

Verb	Noun	Frequency	Association Score
marinar	corriente-S	4	20,37