Práctica 1

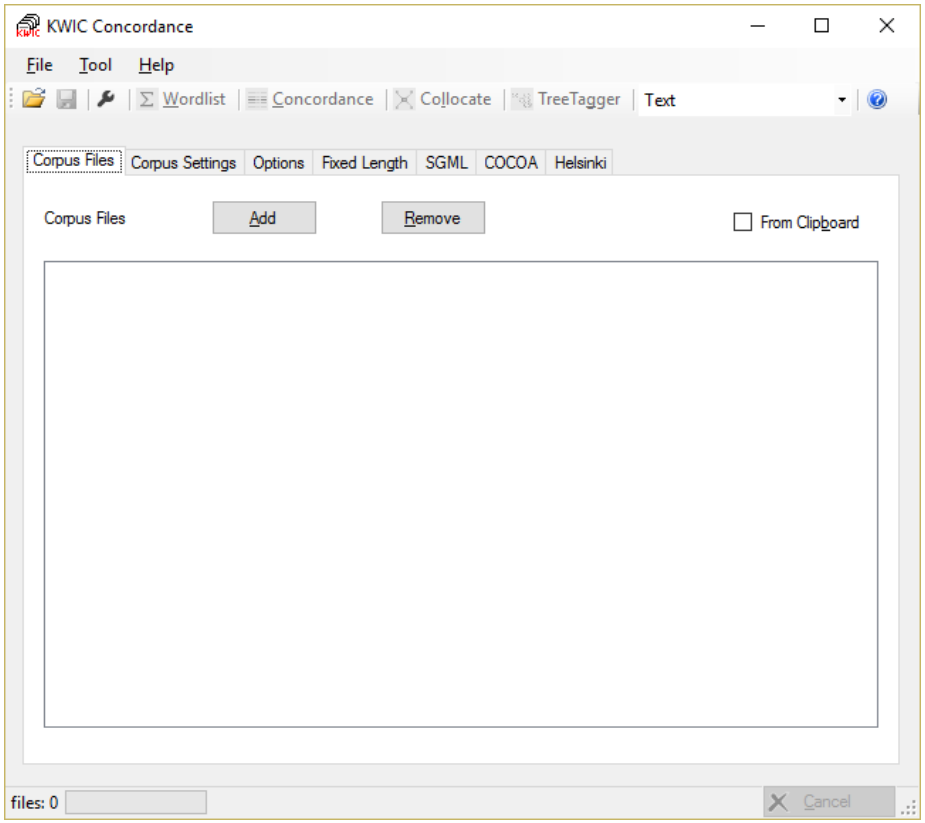
Herramientas de análisis del corpus
Extracción de terminología

KWIC Concordance for Windows

KWIC Concordance for Windows

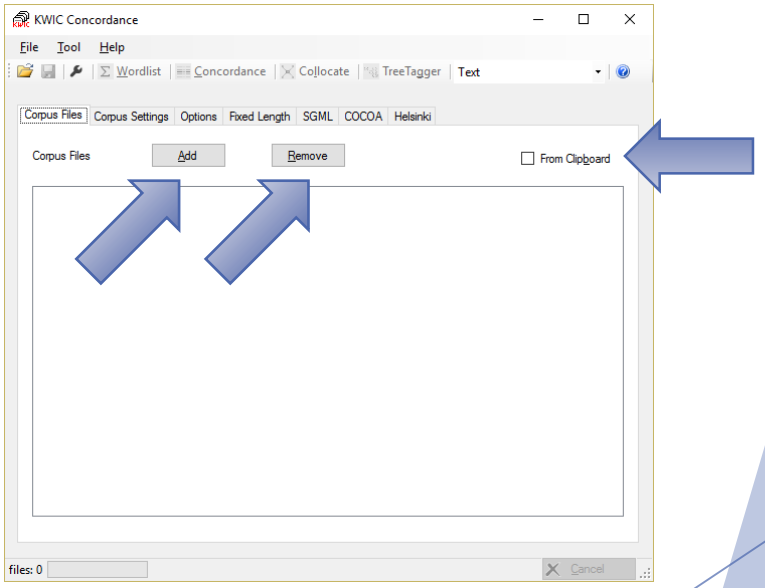
- Herramienta software lingüística
  - Listas de frecuencia de palabras
  - Concordancias
  - Colocaciones
  - N-gramas
  - Etiquetado gramatical (*part-of-speech*, POS) y lemas
- Sistema operativo Windows, licencia *freeware*
- Versión 5.3 (septiembre de 2016)
- http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html

Pantalla principal de KWIC Concordance



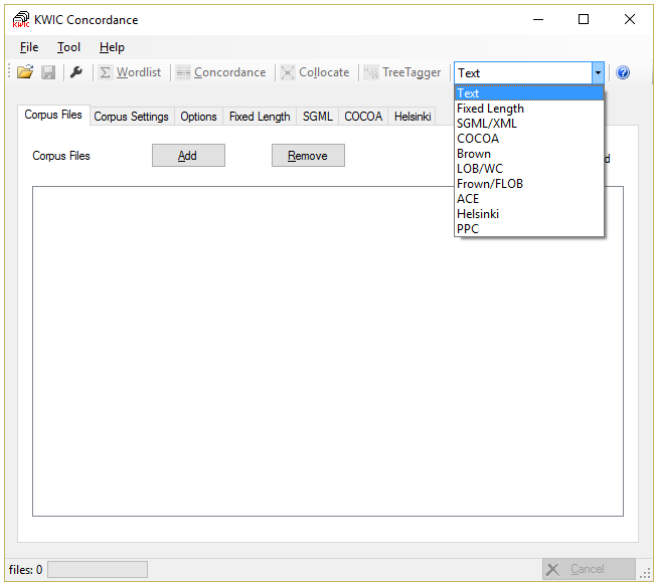
Añadir/eliminar ficheros del corpus

- 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

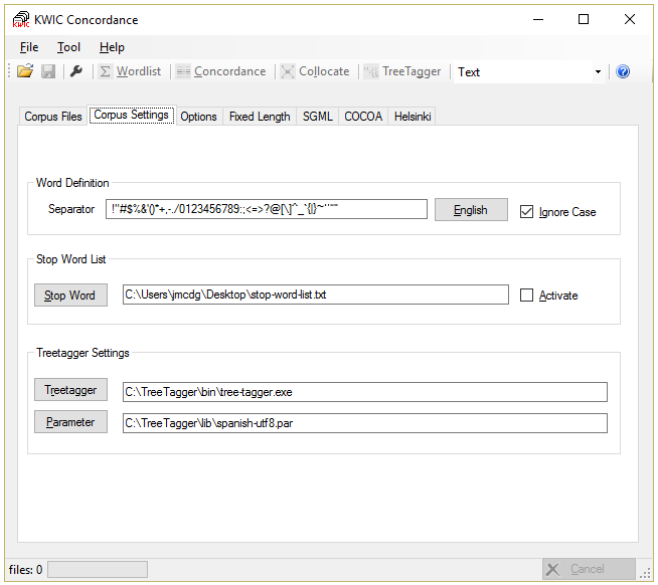


Resumen de formatos

- *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

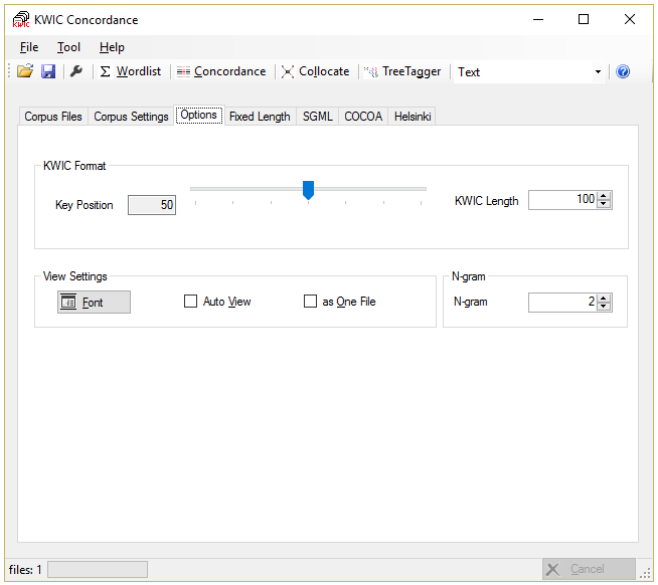


Configuración de TreeTagger

- 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*

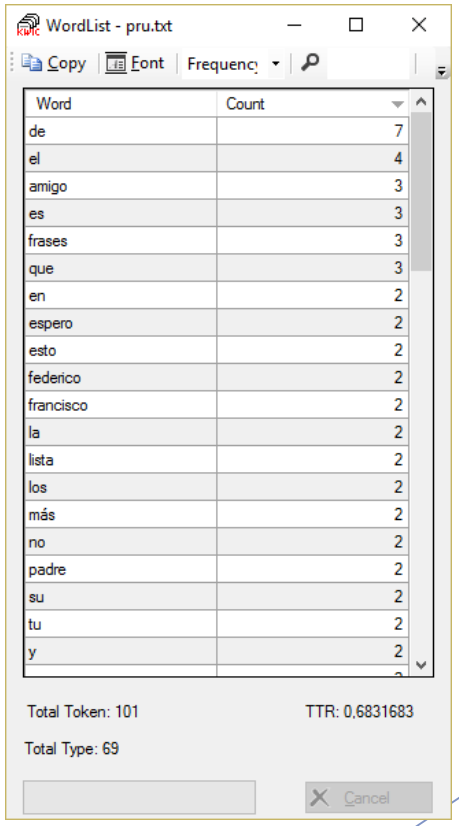
Opciones

- 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)



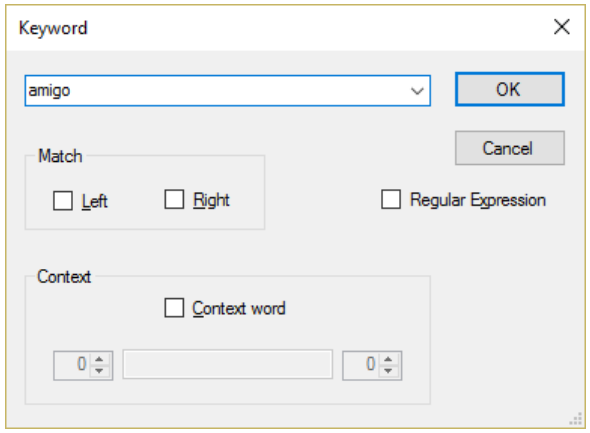
Lista de frecuencia de palabras

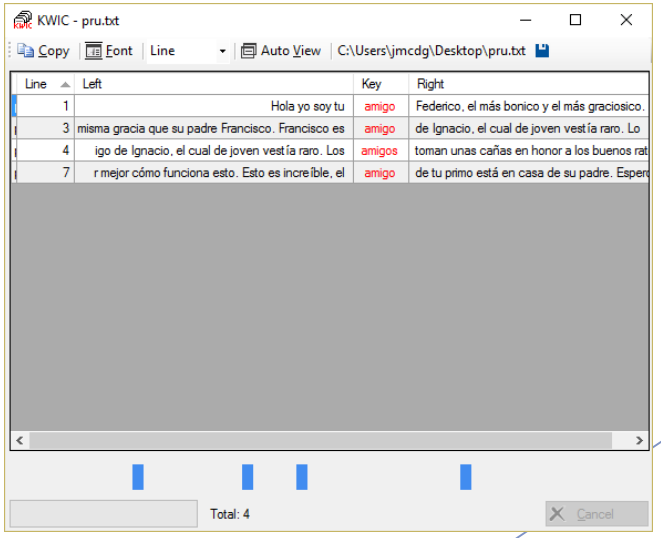
- Este comando devuelve la lista de palabras que conforman el corpus
- Puede ordenarse de acuerdo a diversos criterios:
  - Orden alfabético (directo e inverso)
  - Orden alfabético desde el final de la palabra
  - Frecuencia de aparición



Concordancia

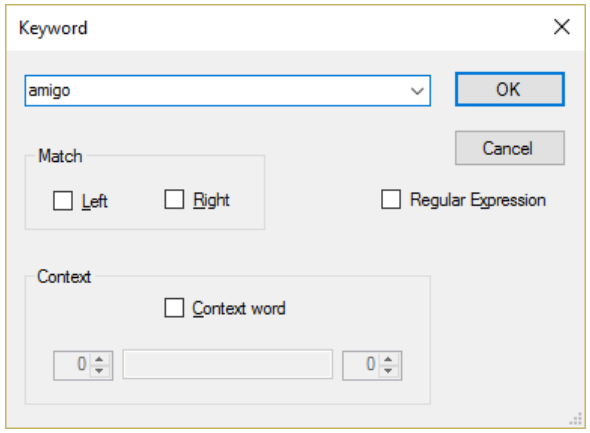
- Este comando devuelve las ocurrencias de un patrón de búsqueda en contexto dentro del corpus
- El resultado es tipo *Key Word In Context* (KWIC)
  - Las ocurrencias del patrón en el centro de la pantalla
  - Se puede ordenar de varias maneras

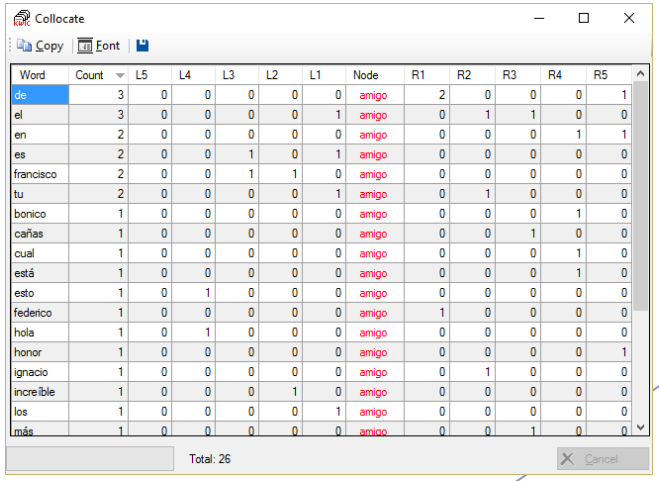




Colocaciones

- 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

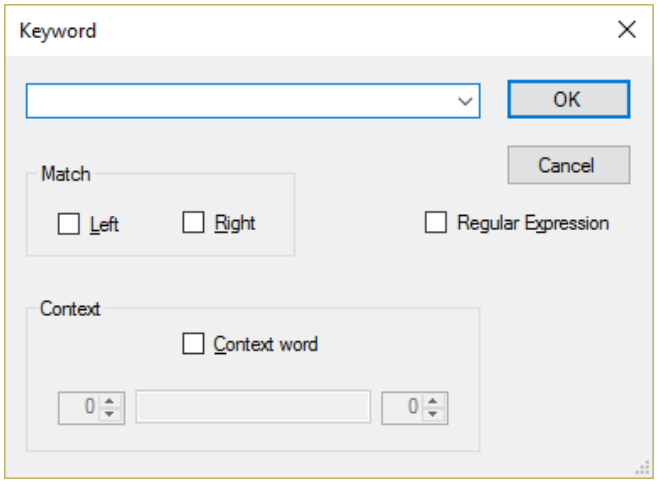




Word	Count	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
de	3	0	0	0	0	0	amigo	2	0	0	0	1
el	3	0	0	0	0	1	amigo	0	1	1	0	0
en	2	0	0	0	0	0	amigo	0	0	0	1	1
es	2	0	0	1	0	1	amigo	0	0	0	0	0
francisco	2	0	0	1	1	0	amigo	0	0	0	0	0
tu	2	0	0	0	0	1	amigo	0	1	0	0	0
bonico	1	0	0	0	0	0	amigo	0	0	0	1	0
cañas	1	0	0	0	0	0	amigo	0	0	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
federico	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

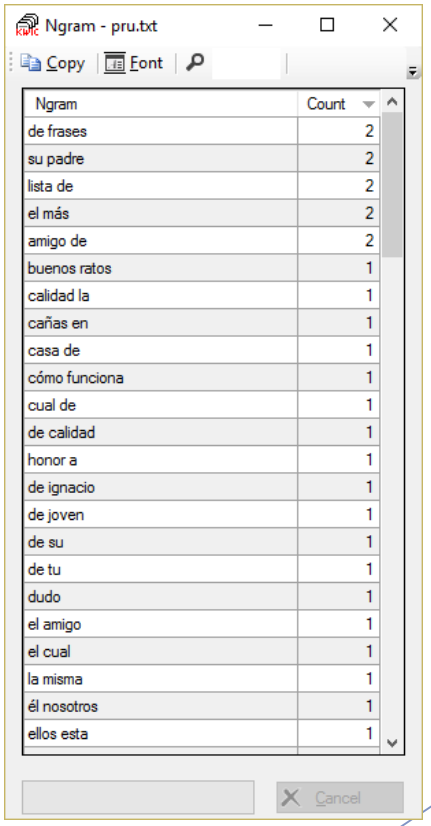
Diálogo “Keyword”

- 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

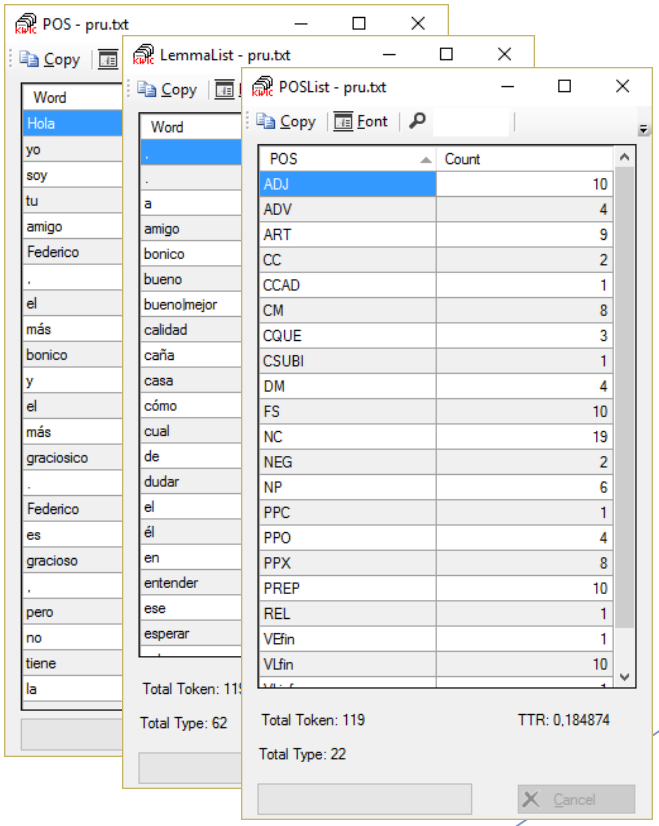
- Este comando devuelve los n-gramas que pueden hallarse en el corpus
- El resultado es una lista de n-gramas con su frecuencia de aparición
  - Un n-grama es una subsecuencia de *n* elementos en una secuencia dada
  - El valor “n” puede establecerse en la pestaña de opciones



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



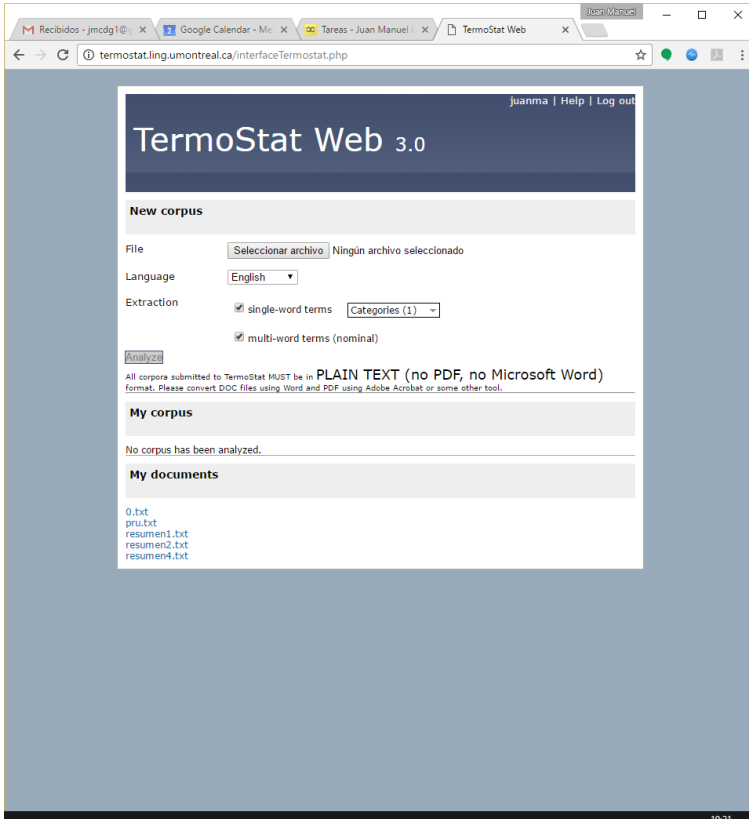
TermoStat

TermoStat

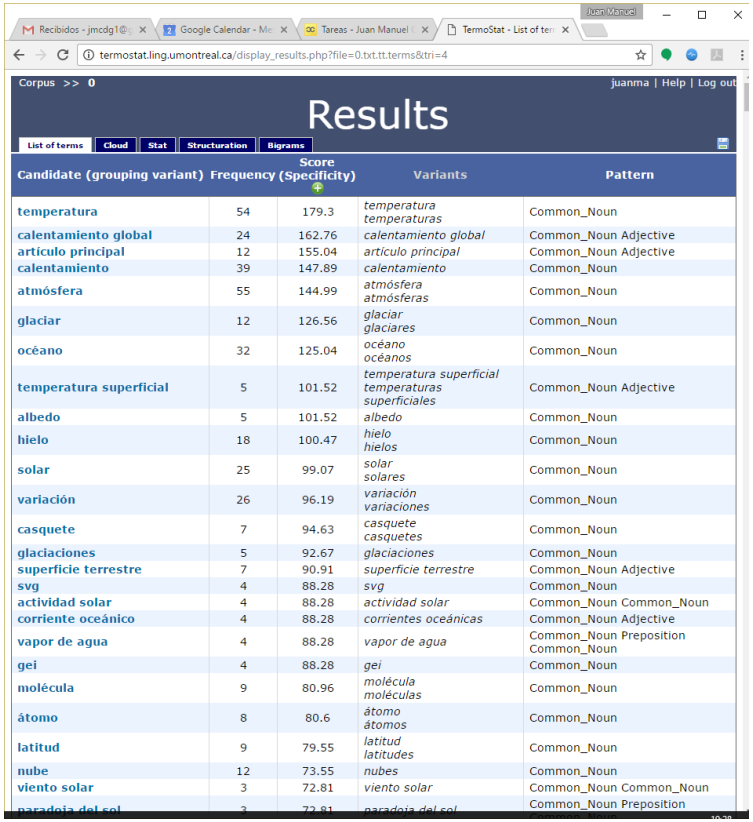
- Herramienta de análisis de corpus automática
  - Listas de frecuencia de palabras
  - Concordancias
  - Nube de términos
  - Ocurrencia de patrones gramaticales
  - Estructuración
  - Bigramas
- Aplicación web, accesible desde cualquier navegador
- Gratuita (necesario registro previo en el portal)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

Pantalla de login de TermoStat

Pantalla principal



Lista de términos



termostat.ling.umontreal.ca/display_results.php?file=0.txt.tt.terms&tri=4

Corpus >> 0

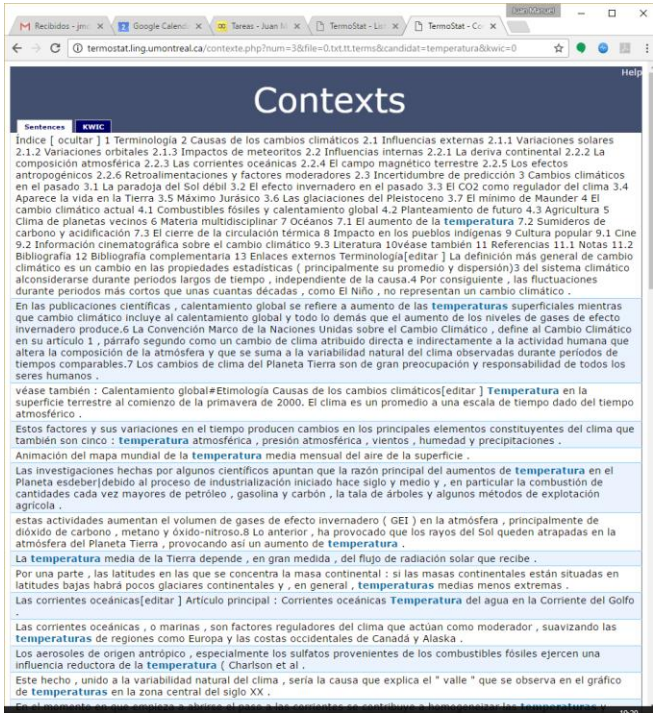
juanma | Help | Log out

Results

List of terms | Cloud | Stat | Structuration | Bigrams

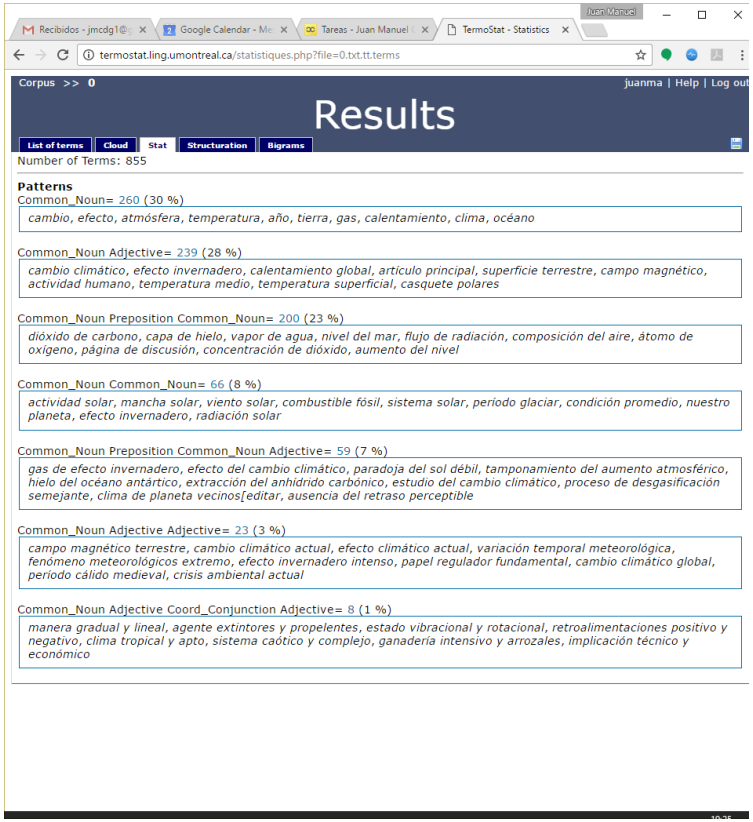
Candidate (grouping variant)	Frequency	Score (Specificity)	Variants	Pattern
temperatura	54	179.3	*temperatura* *temperaturas*	Common_Noun
calentamiento global	24	162.76	*calentamiento global*	Common_Noun Adjective
artículo principal	12	155.04	*artículo principal*	Common_Noun Adjective
calentamiento	39	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 Common_Noun
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	3	72.81	*paradoja del sol*	Common_Noun Preposition

Concordancias

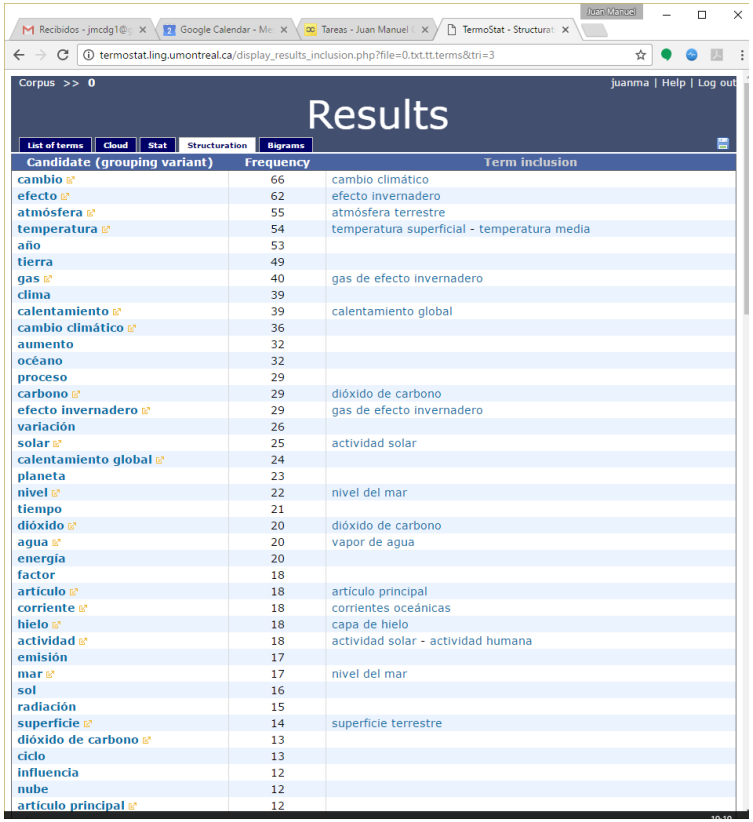


Nube de palabras

Patrones gramaticales

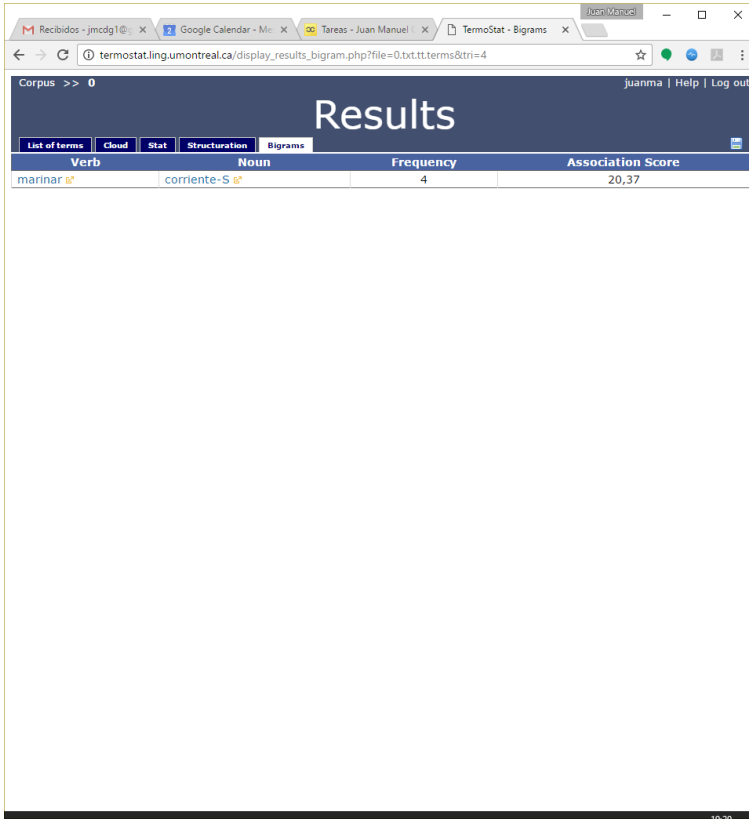


Estructuración



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


TermoStat - Bigrams
termostat.ling.umontreal.ca/display_results_bigram.php?file=0.txt.tt.terms&tri=4
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