LIFCACH 2.0

Lista de Frecuencias de Palabras del Castellano de Chile
Word Frequency List of Chilean Spanish

Contacto / *Contact*:
s s a d o w s k y @ g m a i l . c o m

CONTENIDOS DEL ARCHIVO ZIP / *CONTENTS OF THE ZIP FILE*

1. INFORMACIÓN SOBRE LA LIFCACH
   INFORMATION ABOUT THE LIFCACH

   `Sadowsky_&_Martinez_-_LIFCACH-2.0-README.rtf`

   El presente archivo.
   This file.

2. LISTA DE FRECUENCIAS, POR FUENTE, EN FORMATO EXCEL 2007
   FREQUENCY LIST, BY SOURCE, IN EXCEL 2007 FORMAT

   `Sadowsky_&_Martinez_-_LIFCACH-2.0.csv`

   Este archivo contiene la lista de frecuencias totales (la columna *Total Occurrences*), la categoría gramatical (la columna *POS*), las ocurrencias de cada token por millón de palabras (la columna *Per million words*) y las listas de frecuencias correspondientes a cada subcorpus.

   This file contains the list of total frequencies (the Total Occurrences *column), part of speech (the* POS *column), occurrences per million words (the* Per million words *column), and individual frequency lists for each sub-corpus.*

NOTAS / *NOTES*

1. Descripción / *Description*

The Word Frequency List of Chilean Spanish (LIFCACH) is a set of 102 frequency lists derived from the sub-corpora of the *Corpus Dinámico del Castellano de Chile* (Dynamic Corpus of Chilean Spanish, CODICACH), a corpus of contemporary written[1] Chilean Spanish developed by Sadowsky between 1997 and 2002; this corpus contained approximately 450 million words when the LIFCACH was created[2]. The LIFCACH also contains a non-weighted list of total frequencies (the *Total Occurrences* column), which is the sum of the frequencies of the 102 individual lists (in other words, the list of frequencies of the entire CODICACH corpus.)

The CODICACH is an opportunistic corpus with a bias toward press-based sources; it does not seek to be a BNC-style representative sampling of the overall written language. The modular nature of the CODICACH and of the 102 individual LIFCACH lists, however, allows researchers to use one or more of these lists alone, to combine them as needed, or to create their own frequency lists for Chilean Spanish by weighting each of the LIFCACH's individual lists as they see fit.

The LIFCACH 2.0 contains 476,776 lemmas[3] derived from the approximately 4.5 million types found in the 450 million running words contained in the CODICACH at the time the lists were created.

Version 1.1 added statistics on the frequency of occurrence of each lemma per million words of running text.

Version 2.0 corrects a handful of cases where a word with more than one part of speech classification was treated as more than one lexeme. Thanks to José Joaquín Atria for tracking this down.

2. Elaboración de la LIFCACH / *Creation of the LIFCACH*

The steps in creating the LIFCACH were as follows:

i. Type frequency lists based on the running words of each of the 102 sub-corpora of the CODICACH were generated.
ii. Each type frequency list was lemmatized and POS-tagged using the Universitat Politecnica de Catalunya's MS-Tools v2.0[4].
iii. Lemmas with a frequency of 1 were removed (approximately 300,000) in the case of the `…No-Hapax.xlsx` version. Eliminating these was considered an acceptable trade-off in exchange for a far more manageable file size.
iv. The resulting lemma frequency lists were assembled and total occurrences were calculated.

An important caveat regarding this methodology must be mentioned. The use of type frequency lists instead of running words in the POS tagging and lemmatizing process was a practical

[1] Although the CODICACH does contain two oral corpora, *ORAL_Entrevistas_Lgtcas* and *ORAL_TV*, these are of such negligible size that the CODICACH must be considered a corpus of written Spanish.
[2] The CODICACH currently contains approximately 830 million words.
[3] This is the number of non-hapax lemmas. The total number of lemmas in the LIFCACH, including hapax legomena, is 844,370.
[4] For more information on MS-Tools, contact Lluís Padró at padro@lsi.upc.es.

necessity, due to the speed of the software used and the computing resources available at the time the LIFCACH was created. However, this reduced the accuracy of the lemmatization process by eliminating context. As a result, the software had to analyze words such as *canto* without the information required to decide if a given instance of this word is a form of the verb *cantar* or the noun *canto*.

It should also be noted that the lemmatizing and tagging software that was used is based on European Spanish, a national dialect that is rather removed from Chilean Spanish.

3. Lista de categorías gramaticales / *Part of Speech List*

The following are the POS codes used in the frequency lists.

CÓDIGO/CODE	CATEGORÍA GRAMATICAL	PART OF SPEECH
AJ	Adjetivo	Adjective
AV	Adverbio	Adverb
C	Conjunción	Conjunction
D	Determinante	Determiner
I	Interjección	Interjection
N	Sustantivo	Common noun
NG	Nombre geográfico	Toponym
NP	Nombre propio	Proper noun
PN	Pronombre	Pronoun
PP	Preposición	Preposition
SG	Sigla	Abbreviation
V	Verbo	Verb

4. Listado de fuentes / *List of Sources*

Each frequency list in the LIFCACH is derived from a different sub-corpus of the CODICACH. The codes used for these lists are as follows:

CÓDIGO/*CODE*	DESCRIPCIÓN/*DESCRIPTION*
ACAD_CCAA	Academic Texts - Applied Sciences
ACAD_CCNN	Academic Texts - Natural Sciences
ACAD_CCSS	Academic Texts - Social Sciences
ACAD_Hum	Academic Texts - Humanities
DIAR_CEN_Estrella_Valpo	Newspaper – Central Chile – Estrella de Valparaíso
DIAR_CEN_Gran_Valpo	Newspaper – Central Chile – Gran Valparaíso
DIAR_CEN_Lider_San_Antonio	Newspaper – Central Chile – El Líder, San Antonio
DIAR_CEN_Mercurio_Valpo	Newspaper – Central Chile – El Mercurio, Valparaíso
DIAR_NOR_Estrella_Arica	Newspaper – North Chile – La Estrella, Arica
DIAR_NOR_Estrella_Iquique	Newspaper – North Chile – La Estrella, Iquique
DIAR_NOR_Estrella_Loa	Newspaper – North Chile – La Estrella, Loa
DIAR_NOR_Estrella_Norte_Antofagasta	Newspaper – North Chile – La Estrella, Antofagasta
DIAR_NOR_Mercurio_Antofagasta	Newspaper – North Chile – El Mercurio, Antofagasta
DIAR_NOR_Mercurio_Calama	Newspaper – North Chile – El Mercurio, Calama
DIAR_NOR_Nortino_Iquique	Newspaper – North Chile – El Nortino, Iquique
DIAR_SAN_Cuarta	Newspaper – Santiago – La Cuarta
DIAR_SAN_Estrategia	Newspaper – Santiago – Estrategia
DIAR_SAN_Firme	Newspaper – Santiago – La Firme

DIAR_SAN_Mercurio	Newspaper – Santiago – El Mercurio
DIAR_SAN_Metropolitano	Newspaper – Santiago – El Metropolitano
DIAR_SAN_Mostrador	Newspaper – Santiago – El Mostrador
DIAR_SAN_Primera_Linea	Newspaper – Santiago – Primera Línea
DIAR_SAN_Primera_Pagina-El_Area	Newspaper – Santiago – Primera Página / El Área
DIAR_SAN_Segunda	Newspaper – Santiago – La Segunda
DIAR_SAN_Tercera	Newspaper – Santiago – La Tercera
DIAR_SAN_Ultimas_Noticias	Newspaper – Santiago – Las Últimas Noticias
DIAR_SUR_Austral_Osorno	Newspaper – South Chile – Austral, Osorno
DIAR_SUR_Austral_Temuco	Newspaper – South Chile – Austral, Temuco
DIAR_SUR_Austral_Valdivia	Newspaper – South Chile – Austral, Valdivia
DIAR_SUR_Cronica	Newspaper – South Chile – Crónica
DIAR_SUR_El_Sur	Newspaper – South Chile – El Sur
DIAR_SUR_Enc_BioBio	Newspaper – South Chile – Enciclop. Bío-Bío
DIAR_SUR_Llanquihue_Pto_Montt	Newspaper – South Chile – El Llanquihue, Pto. Montt
ESPER_CartasDirector	Personal Writings – Letters to Editor
ESPER_ForosInet	Personal Writings – Internet Site Forums
ESPER_Clasificados	Personal Writings – Classified Ads
ESPER_ForosMedios	Personal Writings – Media Forums
ESPER_Usenet	Personal Writings – Usenet
LEX_Jurisprudencia	Legal – Jurisprudence
LEX_Leyes	Legal – Laws
LEX_Libros	Legal – Law Books
LEX_Misc	Legal – Miscellaneous
LIBR_Ficcion	Books – Fiction
LIBR_NoFiccion	Books – Non-Fiction
OBRC_CandiaCares_DicoCoa	Reference Works – Dictionary of Coa
OBRC_GonzalezParra_ManualProvrb	Reference Works – Book of Chilean Proverbs
ORAL_Entrevistas_Lgtcas	Oral – Linguistic Interviews
ORAL_TV	Oral – Television
PUB_Misc	Advertising – General 1
PUB_Publicidad	Advertising – General 2
REV_CMP_ChileTech	Magazine – Computers – ChileTech
REV_CMP_CompuChile	Magazine – Computers – CompuChile
REV_CMP_ComputerWorld	Magazine – Computers – ComputerWorld
REV_CMP_Informatica	Magazine – Computers – Informática
REV_CMP_Infoweek	Magazine – Computers – Infoweek
REV_CMP_Internet21	Magazine – Computers – Internet21
REV_CMP_Mouse	Magazine – Computers – Mouse
REV_DEP_All	Magazine – Sports
REV_ESP_Capital	Magazine – Specialty – Capital
REV_ESP_CiudadArquitectura	Magazine – Specialty – CiudadArquitectura
REV_ESP_Conicyt	Magazine – Specialty – Conicyt Scientific
REV_ESP_CopropInmob	Magazine – Specialty – Copropiedad Inmobiliaria
REV_ESP_DiarioSocCivil	Magazine – Specialty – Diario de la Sociedad Civil
REV_ESP_Educar	Magazine – Specialty – Educar
REV_ESP_LemuChile	Magazine – Specialty – LemuChile
REV_ESP_Lignum	Magazine – Specialty – Lignum
REV_ESP_Mensaje	Magazine – Specialty – Mensaje
REV_ESP_Notas_CESAF	Magazine – Specialty – Notas CESAF
REV_ESP_Publimark	Magazine – Specialty – Publimark
REV_ESP_Rev_Inf_Musical	Magazine – Specialty – Revista Musical
REV_ESP_Rev_Scielo	Magazine – Specialty – Scielo Scientific
REV_ESP_Rev_Social	Magazine – Specialty – Revista Social

REV_ESP_Rev_Trabajo_Social	Magazine – Specialty – Revista de Trabajo Social
REV_ESP_RevChil_Cirujia	Magazine – Specialty – Revista Chilena de Cirujía
REV_ESP_Revistas_Industriales	Magazine – Specialty – Industrial Magazines
REV_ESP_Sidhartha	Magazine – Specialty – Siddhartha
REV_GEN_Asuntos_Publicos	Magazine – General – Asuntos Públicos
REV_GEN_Cosas	Magazine – General – Cosas
REV_GEN_Cultura_Urbana	Magazine – General – Cultura Urbana
REV_GEN_El_Siglo	Magazine – General – El Siglo
REV_GEN_Ercilla	Magazine – General – Ercilla
REV_GEN_Hacer_Familia	Magazine – General – Hacer Familia
REV_GEN_Man	Magazine – General – Man
REV_GEN_Mujer_a_mujer	Magazine – General – Mujer a mujer
REV_GEN_Nos	Magazine – General – Nos
REV_GEN_Puerto_Paralelo	Magazine – General – Puerto Paralelo
REV_GEN_Punto_Final	Magazine – General – Punto Final
REV_GEN_Que_Pasa	Magazine – General – Qué Pasa
REV_GEN_Revista_ED	Magazine – General – Revista ED
REV_GEN_Rocinante	Magazine – General – Rocinante
REV_INF_Dirigible	Magazine – Children's – Dirigible
REV_INF_Icarito	Magazine – Children's – Icarito
REV_INF_Papas_Fritas	Magazine – Children's – Papas Fritas
REV_INF_Volare	Magazine – Children's – Volare
REV_JUV_All	Magazines – Youth
REV_LOC_All	Magazines – Local
RVDI_ECN_Diario_PyME	Financial Mags & Newspapers – Diario PyME
RVDI_ECN_El_Diario	Financial Mags & Newspapers – El Diario
RVDI_ECN_Emprendedores	Financial Mags & Newspapers – Emprendedores
RVDI_ECN_Negocios_Ambientales	Financial Mags & Newspapers – Negoc. Ambientales
SIT_INS_All	Government Sites 1
SIT_INS_Old	Government Sites 2

TEMUCO, CHILE - AUGUST 2012