Recursos Lingüísticos de SINAI

L. Alfonso Ureña López
Mª. Teresa Martín Valdivia
Miguel Á. García Cumbreras
Eugenio Martínez Cámara
Salud María Jiménez Zafra
Alberto Andreu Marín
Mª Ángeles Sánchez Morales

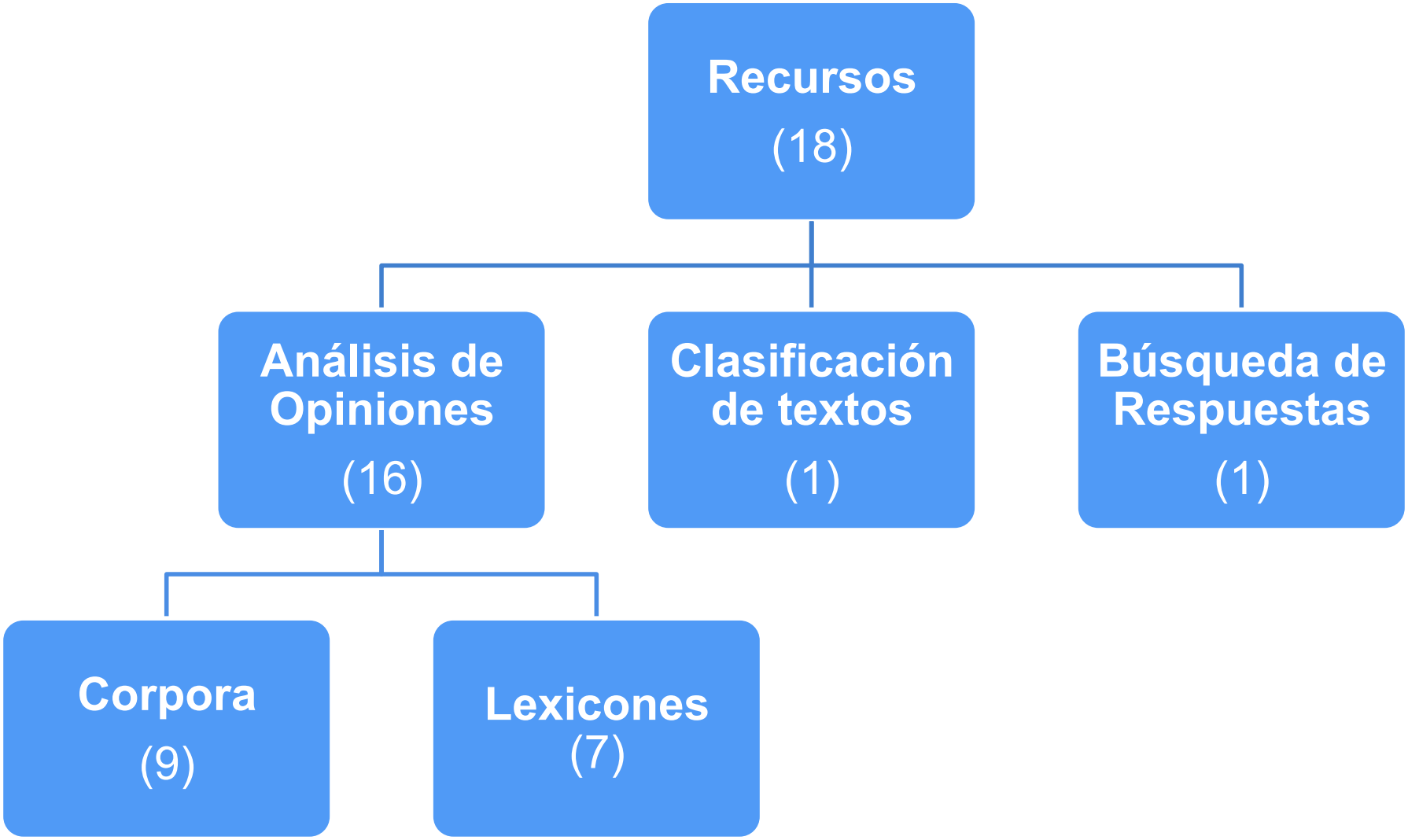
SINAI

- Creado en el año 2000
- Actualmente está formado por 9 miembros:
  - 7 Investigadores
    - L. Alfonso Ureña López – Investigador Principal
  - 1 Estudiante de doctorado
  - 1 Becario en prácticas
- El grupo cuenta con más de 400 publicaciones en revistas, congresos y libros
- Sus miembros forman parte de varios comités de programa, dirección de sociedades científicas, así como colaboran como revisores en numerosas revistas
- URL: http://sinai.ujaen.es

Principales Líneas de Investigación

- Procesamiento del Lenguaje Natural
- Análisis de Opiniones
- Recuperación de información Multilingüe, Multimodal y Geográfica
- Desambiguación
- Búsqueda de Respuestas
- Categorización de documentos
- Minería de datos y de texto
- Extracción de Información
- Aprendizaje Automático aplicado al PLN
- Integración de conocimiento basado en ontologías
- …

Generación de Recursos



ANÁLISIS DE OPINIONES

Análisis de Opiniones

Corpora

- COPOS: Corpus de opiniones en el dominio médico; 743 opiniones
- COST: Corpus de tweets en español; 34634 tweets
- COAH: Corpus de opiniones de hoteles; 1816 opiniones
- COAR: Corpus de opiniones de restaurantes; 2202 opiniones
- OCA: Corpus de opiniones cinematográficas en árabe; 500 opiniones
- EVOCA: Versión en inglés de OCA
- MCE Corpus: Versión en inglés del corpus MuchoCine (Cruz et al. 2008); 3878 opiniones
- SFU-SRC-Neg: Corpus SFU de opiniones etiquetado a nivel de negación; 400 opiniones
- SinaiSACorpus: 1,943 opiniones sobre cámaras de fotos extraídas de Amazon

Lexicones

- SOL: 3,151 negativas, 1,397 positivas
- iSOL: 2509 positivas, 5626 negativas
- eSOL: Adaptación de iSOL al dominio del cine
- eSOLDomain: Adaptación de iSOL a 8 dominios diferentes
- CRiSOL: Base de conocimiento de diversos recursos de AO filtrados por iSOL
- Emoti-sp: Conjunto de emoticonos positivos (70) y negativos (46)
- Hashtags-sp: Conjunto de hashtags positivos (172) y negativos (127)

COPOS Corpus

- Linguality type: Monolingual
- Language: Spanish
- Text format: Text/plain
- Size
  - Number of sentences: 2,009
  - Number of words: 32,365
- Character encoding: UTF-8

COPOS Corpus

- Classification
  - Text type: Text/plain
  - Register: Informal
- Annotation
  - Semantic Annotation – Polarity
    - Tagset: number of stars (0, 1, 2, 3, 4, 5)
    - Segmentation level: Document
- Creation
  - Creation mode: Automatic
- Version: 1

COPOS Corpus

- Documentation:

Plaza-del-Arco, Flor Miriam, Martín-Valdivia, M. Teresa, Jiménez-Zafra, Salud María, Molina-González, M. Dolores & Martínez-Cámara, Eugenio (2016). COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques. Procesamiento del Lenguaje Natural, 57, 83-90

Distribution

- Availability: Restricted use

COPOS Corpus

- License
  - License: -
  - Restrictions: Research purposes
  - Download location:
  - Distribution Access/Medium:

Contact Person: M. Teresa Martín-Valdivia (maite@ujaen.es)

COPOS Corpus

Resource Creation

- Resource creator: SINAI Research Group, Universidad de Jaén
- Funding project: Partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) from the Spanish Government and by a Grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983)

COST

- Linguality type: Monolingual
- Language: Spanish
- Text format: CSV
- Size
  - Number of documents: 34,634
  - Number positive documents: 17,317
  - Number negative documents: 17,317
- Character encoding: UTF-8

COST

- Classification
  - Text type: tweets
  - Register: informal
- Annotation: Opinion at document level
  - Tagset: Positive, Negative
- Creation
  - Creation mode: Automatic (distant supervision)
- Version: 1.0
- Documentation: Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., & Mitkov, R. (2015). Polarity classification for Spanish tweets using the COST corpus. *Journal of Information Science* 41(3): 263-272.

COST

- **Distribution**
- Availability: Available - Unrestricted Use
- License
  - License: CC - By
  - Restrictions
  - Download location: Hidden
  - Distribution Access/Medium: Downloadable
- **Contact Person:** L. Alfonso Ureña López and Eugenio Martínez Cámara
- **Resource Creation**: March 2011
- Resource creator: SINAI (Eugenio Martínez Cámara, M. Teresa Martín Valdivia, L. Alfonso Ureña López) and Ruslan Mitkov

COST

- Funding project:
  - Name: Análisis de Tendencias y Temáticas a través de Opiniones y Sentimientos (ATTOS, TIN2012-38536-C03-0)
    - URL: http://attos.ujaen.es
    - Funding Type: National Funds
    - Funder: España. Ministerio de Economía y Competitividad
    - Funding Country: España

COST

- Funding project:
  - Name: Análisis de Opinión en Redes Sociales y Contenidos Generados por Usuarios (AORESCU, P11-TIC-7684 MO)
    - URL:
    - Funding Type: Regional Funds
    - Funder: Junta de Andalucía
    - Funding Country: España

COST

- Funding project:
  - Name: Captura y Procesamiento de Grandes Volúmenes de Datos Provenientes de varias Fuentes de Micro-blogging (CEATIC-2013-01)
    - URL:
    - Funding Type: University Funds
    - Funder: Centro de Estudios Avanzados en TIC. Universidad de Jaén
    - Funding Country: España

COST

- Funding project:
  - Name: FIRST Project (FP7-287607)
    - URL: http://www.first-asd.eu/
    - Funding Type: European Funds
    - Funder: European Comission. FP7-2007-2013
    - Funding Country: Europe

COAH Corpus

- Linguality type: Monolingual
- Language: Spanish
- Text format: xml
- Size
  - Number of sentences: 1,816
  - Number of words: 23,9749
- Character encoding: UTF-8

COAH Corpus

- Classification
  - –Text type: xml
  - –Register: Informal
- Annotation: Opinion at document level
  - Tagset: 5 levels, from strong negative to strong positive
- Creation
  - –Creation mode: Automatic
- Version: 1.0

COAH Corpus

- Documentation: Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A. (2014). Cross-domain sentiment analysis using spanish opinionated words. Natural Language Processing and Information Systems, Lecture Notes in Computer Science, vol. 8455, pp. 214-219. Springer International Publishing. DOI: 10.1007/978-3-319-07983-7_28

COAH Corpus

- **Distribution**
- Availability: Available – Restricted use
- License
  - License: Creative Commons
  - Restrictions: Research purposes
  - Download location: http://sinai.ujaen.es/coah/
  - Distribution Access/Medium: Downloadable

Contact Person: M. Dolores Molina (mdmolina@ujaen.es) / Eugenio Martínez (emcamara@ujaen.es)

COAH Corpus

Resource Creation

- Resource creator: SINAI Research Group, Universidad de Jaén
- Funding project: Partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) from the Spanish Government and by a Grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983)

COAR Corpus

- Linguality type: Monolingual
- Language: Spanish
- Text format: xlsx
- Size
  - Number of sentences: 2,202
  - Number of words:
- Character encoding: UTF8

COAR Corpus

- Classification
  - –Text type: Text/Plain
  - –Register: Informal
- Annotation: Opinion at document level
  - Tagset: 5 levels, from strong negative to strong positive
- Creation
  - –Creation mode: Automatic
- Version: 1.0
- Documentation:

COAR Corpus

Distribution

- Availability: Available – Restricted use
- License
  - License: Creative Commons
  - Restrictions: Research purposes
  - Download location: http://sinai.ujaen.es/coar/
  - Distribution Access/Medium: Downloadable

Contact Person: M. Dolores Molina (mdmolina@ujaen.es) / Eugenio Martínez (emcamara@ujaen.es)

COAR Corpus

Resource Creation

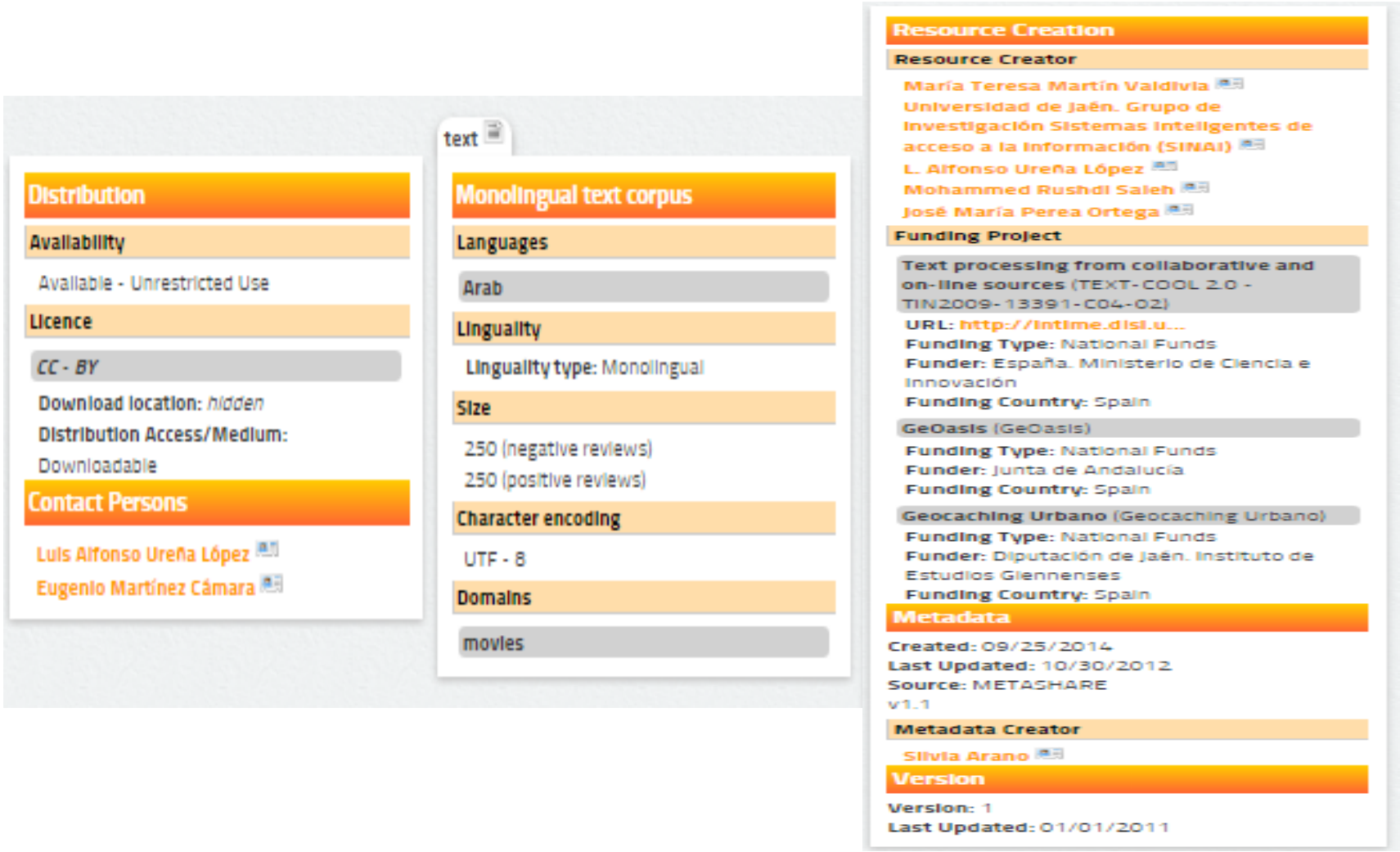
- Resource creator: SINAI Research Group, Universidad de Jaén
- Funding project: Partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) from the Spanish Government and by a Grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983)

OCA Corpus

URL en MetaShare:
http://metashare.elda.org/repository/browse/oca-corpus/



OCA Corpus


text
Distribution
Availability
Available - Unrestricted Use
Licence
CC - BY
Download location: hidden
Distribution Access/Medium:
Downloadable
Contact Persons
Luis Alfonso Ureña López
Eugenio Martínez Cámara
Monolingual text corpus
Languages
Arab
Linguality
Linguality type: Monolingual
Size
250 (negative reviews)
250 (positive reviews)
Character encoding
UTF - 8
Domains
movies
Resource Creation
Resource Creator
María Teresa Martín Valdivia
Universidad de Jaén. Grupo de Investigación Sistemas Inteligentes de acceso a la Información (SINAI)
L. Alfonso Ureña López
Mohammed Rushdi Saleh
José María Perea Ortega
Funding Project
Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)
URL: http://intime.dlsi.u...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain
GeOasis (GeOasis)
Funding Type: National Funds
Funder: Junta de Andalucía
Funding Country: Spain
Geocaching Urbano (Geocaching Urbano)
Funding Type: National Funds
Funder: Diputación de Jaén. Instituto de Estudios Giennenses
Funding Country: Spain
Metadata
Created: 09/25/2014
Last Updated: 10/30/2012
Source: METASHARE
v1.1
Metadata Creator
Silvia Arano
Version
Version: 1
Last Updated: 01/01/2011

EVOCA Corpus

URL en MetaShare:
http://metashare.upf.edu/repository/browse/evoca-corpus/e78964566a7911e48763000c291ecfc81335c45da13642bcbdd40bf8062c97ef/



25 1 4

EVOCA Corpus

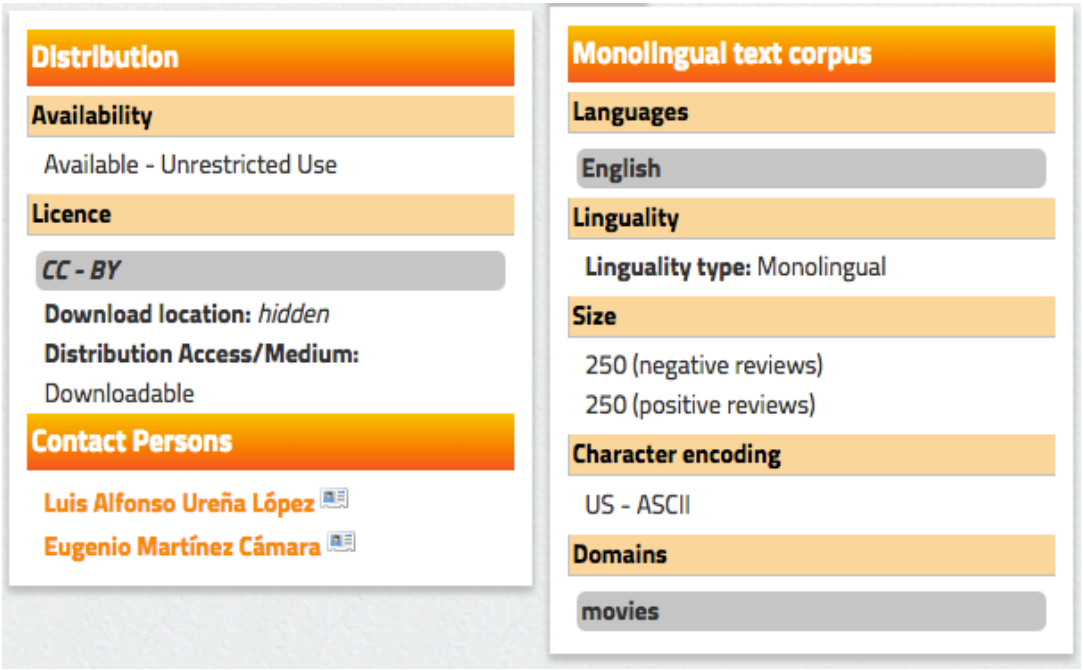
EVOCA Corpus

http://sinai.ujaen.es/es/

ID: http://sinai.ujaen.es/evoca-corpus-en/

EVOCA (English Version of OCA) is an English corpus generated from the translation of the Arabic corpus OCA. This corpus contains reviews of movies and is divided
into 250 positive reviews and 250 negative. This corpus was translated in April 2011.

EVOCA Corpus



Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY

Download location: *hidden*

Distribution Access/Medium:
Downloadable

Contact Persons

Luis Alfonso Ureña López

Eugenio Martínez Cámara

Monolingual text corpus

Languages

English

Linguality

Linguality type: Monolingual

Size

250 (negative reviews)
250 (positive reviews)

Character encoding

US - ASCII

Domains

movies

Resource Creation

Resource Creator

María Teresa Martín Valdivia
Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)
L. Alfonso Ureña López
Mohammed Rushdi Saleh
José María Perea Ortega

Funding Project

Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)
URL: http://intime.dlsi.u...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

GeOasis (GeOasis)
Funding Type: National Funds
Funder: Junta de Andalucía
Funding Country: Spain

Geocaching Urbano (Geocaching Urbano)
Funding Type: National Funds
Funder: Diputación de Jaén. Instituto de Estudios Giennenses
Funding Country: Spain

Metadata

Created: 09/25/2014
Last Updated: 10/30/2012
Source: METASHARE
v1.1

Metadata Creator

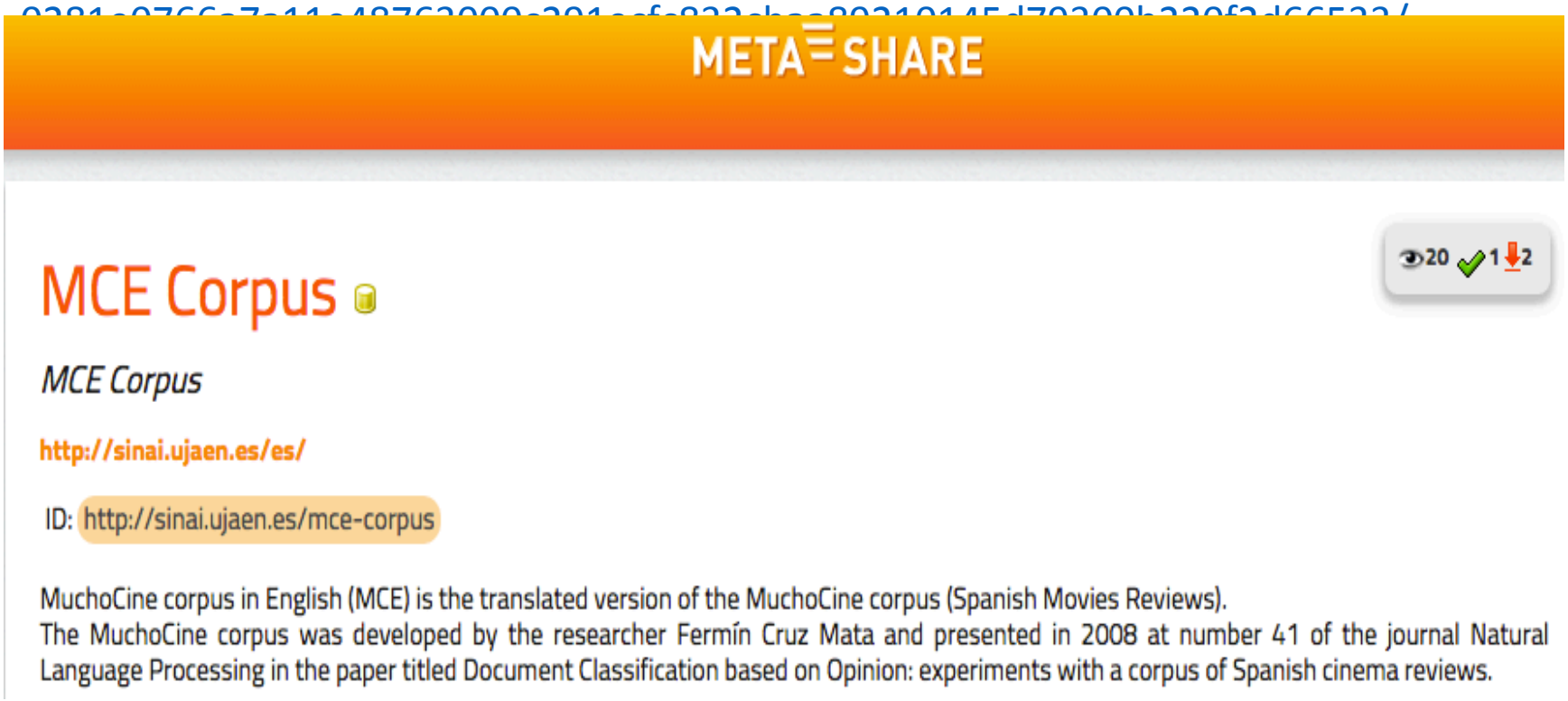
Silvia Arano

Version

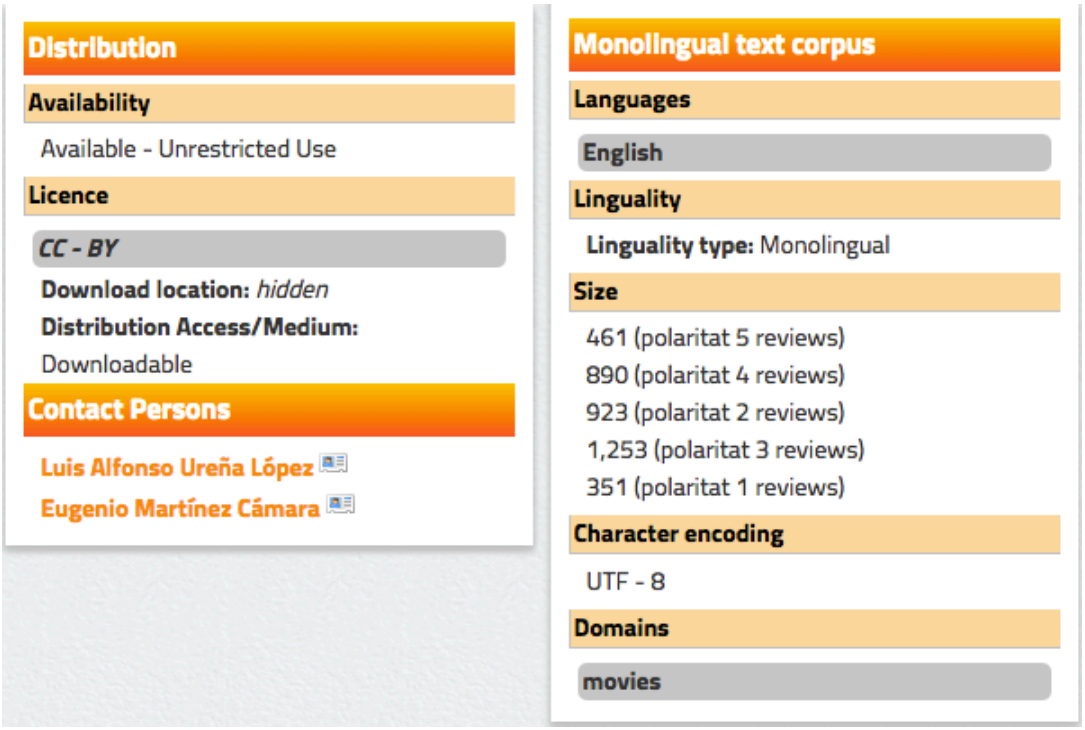
Version: 1
Last Updated: 01/01/2011

MCE Corpus

URL en MetaShare:
http://metashare.upf.edu/repository/browse/mce-corpus/



MCE Corpus



Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY

Download location: *hidden*

Distribution Access/Medium:
Downloadable

Contact Persons

Luis Alfonso Ureña López

Eugenio Martínez Cámara

Monolingual text corpus

Languages

English

Linguality

Linguality type: Monolingual

Size

461 (polaritat 5 reviews)
890 (polaritat 4 reviews)
923 (polaritat 2 reviews)
1,253 (polaritat 3 reviews)
351 (polaritat 1 reviews)

Character encoding

UTF - 8

Domains

movies

Resource Creation

Resource Creator

Eugenio Martínez Cámara
Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)
José María Perea Ortega
L. Alfonso Ureña López
María Teresa Martín Valdivia

Funding Project

Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)
URL: http://intime.dlsi.u...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Flexible interactive reading support tool (First - FP7-287607)
URL: http://www.first-asd...
Funding Type: Eu Funds
Funder: European Union

Metadata

Created: 09/25/2014
Last Updated: 10/30/2012
Source: METASHARE
v1.1

Metadata Creator

Silvia Arano

Version

Version: 1
Last Updated: 01/01/2012

SFU Review$_{SP}$-NEG Corpus

- Linguality type: Monolingual
- Language: Spanish
- Text format: Text/XML
- Size
  - Number of sentences: 9,455
  - Number of words: 221,874
- Character encoding: UTF-8

SFU Review$_{SP}$-NEG Corpus

- Classification
  - Text type: Text/XML
  - Register: Informal
- Annotation
  - Morphosyntactic Annotation – PoS Tagging
    - Tagset: Freeling
    - Segmentation level: Word
  - Semantic Annotation – Polarity
    - Tagset: positive, negative
    - Segmentation level: Document

SFU Review$_{SP}$-NEG Corpus

- –Semantic Annotation – Polarity negation
  - –Tagset: positive, negative, neutral
  - –Segmentation level: Sentence
- –Semantic Annotation - Negation
  - –Tagset: Other
  - –Segmentation level: Sentence
- –Syntactic Annotation – Negation
  - –Tagset: other
  - –Segmentation level: Sentence

- •Creation
  - –Creation mode: Mixed
- •Version: 1.0

SFU Review$_{SP}$-NEG Corpus

- Documentation:

Martí, M. Antònia, Taulé, Mariona, Nofre, Montserrat, Marsó, Laia, Martín-Valdivia, M. Teresa & Jiménez-Zafra, Salud María (2016). La negación en español: análisis y tipología de patrones de negación. Procesamiento del Lenguaje Natural, 57, 41-48

Distribution

- Availability: Available – Restricted use

SFU Review$_{SP}$-NEG Corpus

- License
  - –License: Creative Commons
  - –Restrictions: Research purposes
  - –Download location: http://sinai.ujaen.es/sfu-review-sp-neg/
  - –Distribution Access/Medium: Downloadable

Contact Person: Salud María Jiménez-Zafra (sjzafra@ujaen.es)

SFU Review$_{SP}$-NEG Corpus

Resource Creation

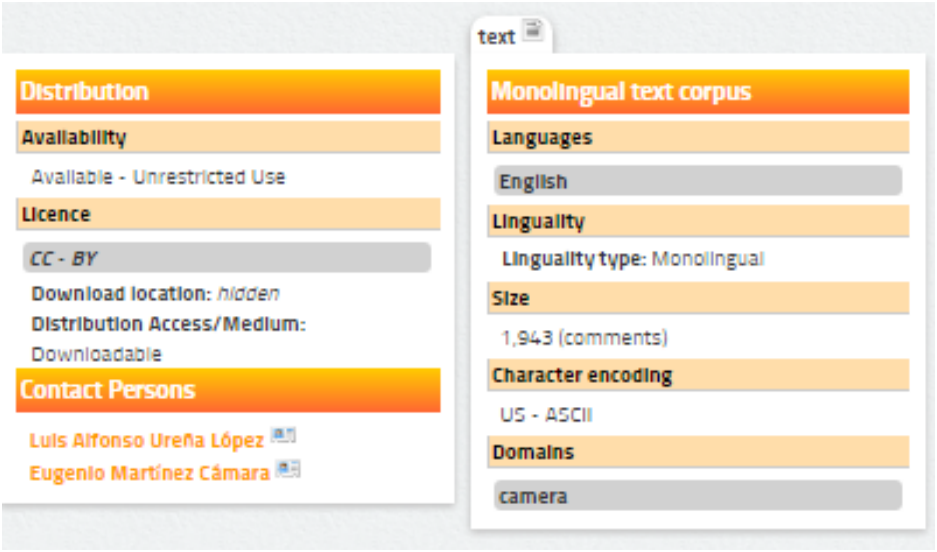
- Resource creator: SINAI Research Group, Universidad de Jaén
- Funding project: Partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and SOMEMBED-SLANG project (TIN2015-71147-C2-2) from the Spanish Government and by a Grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983)

SinaiSA Corpus

http://metashare.elda.org/repository/browse/sinaisacorpus/41619eee6a7a11e48763000c291ecfc877a3f7d5e94a4a53a1ab8793ab9360c5/



SinaiSA Corpus





SOL

URL en MetaShare:
http://metashare.upf.edu/repository/browse/sol/89938ea26a7a11e48763000c291ecfc8304b503d91c64fb99f43b21a0d26cbcd/



SOL

Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY

Download location: *hidden*

Distribution Access/Medium: Downloadable

Contact Persons

Luis Alfonso Ureña López

Eugenio Martínez Cámara

Lexical Conceptual Resource General Information

Lexicon

Encoding

Encoding level: Semantics

Linguistic information: Lemma

Creation

Creation mode details: This lexicon was automatically translated with the Reverso program.

Creation mode: Automatic

Original Sources

- Bing Liu's Opinion Lexicon at **http://www.cs.uic.ed...**

Creation Tools

- Reverso program

SOL

Monolingual text lexicalConceptualResource

Languages

Spanish

Linguality

Linguality type: Monolingual

Text Format

text/plain

Size

3,151 (negative) Words

1,397 (positive) Words

Character encoding

UTF - 8

Domains

general

Metadata

Created: 09/25/2014

Last Updated: 10/30/2012

Source: METASHARE

v1.1

Metadata Creator

Silvia Arano

Version

Version: 1

Last Updated: 01/01/2013

Resource Creation

Resource Creator

M. Dolores Molina González

Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)

María Teresa Martín Valdivia

L. Alfonso Ureña López

Eugenio Martínez Cámara

Funding Project

Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)

URL: http://intime.dlsi.u...

Funding Type: National Funds

Funder: España. Ministerio de Ciencia e Innovación

Funding Country: Spain

Trends and subjects analysis through opinions and feelings (ATTOS - TIN2012-38536-C03-0)

URL: http://gplsi.dlsi.ua...

Funding Type: National Funds

Funder: España. Ministerio de Ciencia e Innovación

Funding Country: Spain

iSOL

Resource Creation

Resource Creator

M. Dolores Molina González
Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)
María Teresa Martín Valdivia
L. Alfonso Ureña López
Eugenio Martínez Cámara

Funding Project

Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)
URL: http://intime.dlsi.u...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Trends and subjects analysis through opinions and feelings (ATTOS - TIN2012-38536-C03-0)
URL: http://gplsi.dlsi.ua...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Metadata

Created: 09/25/2014
Last Updated: 10/30/2012
Source: METASHARE
v1.1

Metadata Creator

Silvia Arano

Version

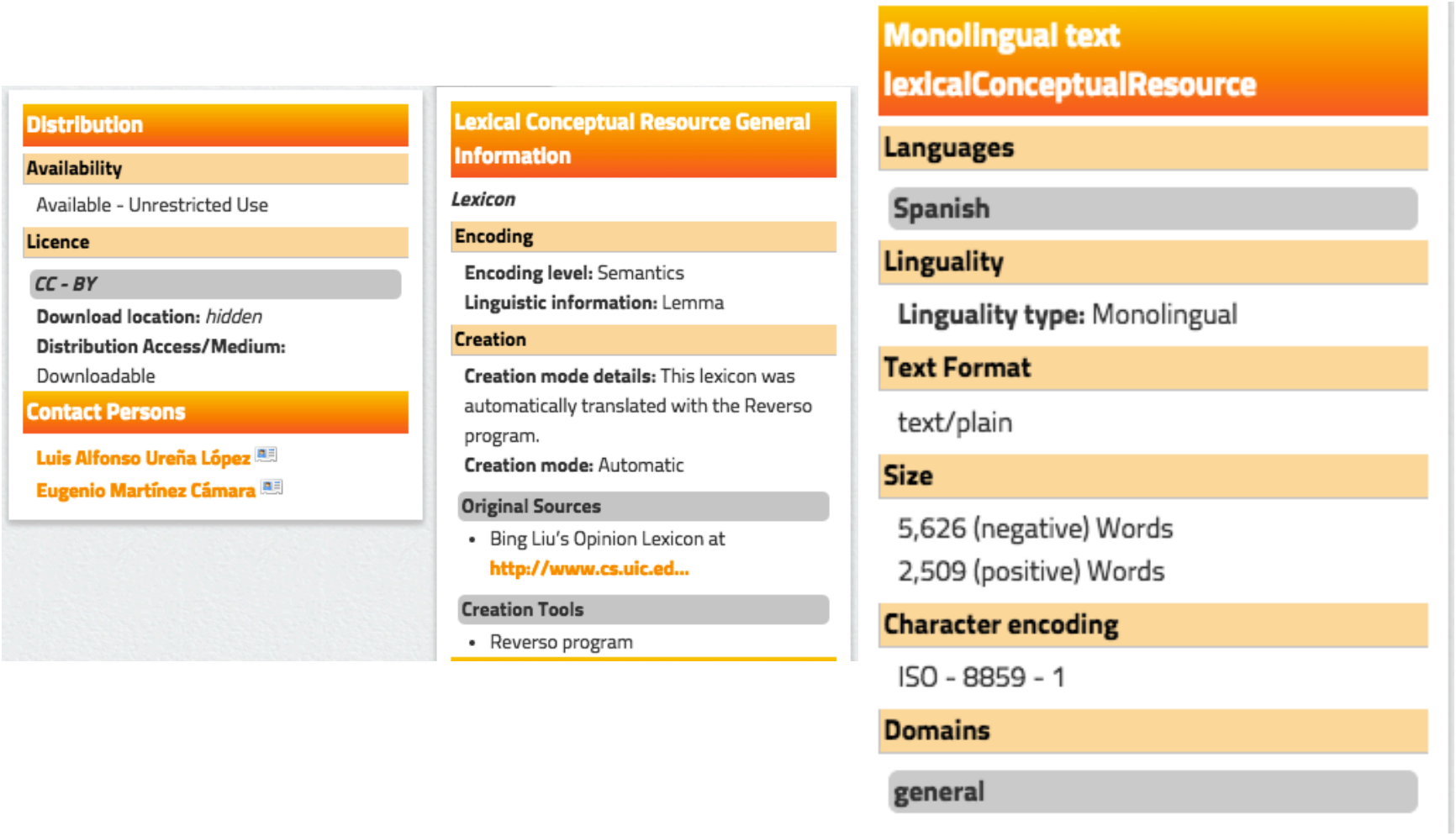
Version: 1
Last Updated: 01/01/2013

iSOL

URL en MetaShare:
http://metashare.upf.edu/repository/browse/isol/88feb0fc6a7a11e48763000c291ecfc826e3f9a2650446018e36fabb6926e294/



iSOL



Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY

Download location: *hidden*

Distribution Access/Medium: Downloadable

Contact Persons

Luis Alfonso Ureña López

Eugenio Martínez Cámara

Lexical Conceptual Resource General Information

Lexicon

Encoding

Encoding level: Semantics

Linguistic information: Lemma

Creation

Creation mode details: This lexicon was automatically translated with the Reverso program.

Creation mode: Automatic

Original Sources

- Bing Liu's Opinion Lexicon at http://www.cs.uic.ed...

Creation Tools

- Reverso program

Monolingual text lexicalConceptualResource

Languages

Spanish

Linguality

Linguality type: Monolingual

Text Format

text/plain

Size

5,626 (negative) Words

2,509 (positive) Words

Character encoding

ISO - 8859 - 1

Domains

general

iSOL

Resource Creation

Resource Creator

M. Dolores Molina González
Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)
María Teresa Martín Valdivia
L. Alfonso Ureña López
Eugenio Martínez Cámara

Funding Project

Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)
URL: http://intime.dlsi.u...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Trends and subjects analysis through opinions and feelings (ATTOS - TIN2012-38536-C03-0)
URL: http://gplsi.dlsi.ua...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Metadata

Created: 09/25/2014
Last Updated: 10/30/2012
Source: METASHARE
v1.1

Metadata Creator

Silvia Arano

Version

Version: 1
Last Updated: 01/01/2013

eSOL

URL en MetaShare:
http://metashare.upf.edu/repository/browse/esol/9a7895e265c911e48763000c291ecfc857fe6ffba28f4aa591d7a7e60a3c34b4/



22 1 5

eSOL ab

eSOL

http://sinai.ujaen.es/es/

ID: http://sinai.ujaen.es/esol-2/

iSOL is a list of domain-dependent opinion signal words in Spanish. The domain is the set of words of movie reviews.
The elaboration of the list was performed using a corpus-based approach. In this case it selected the Spanish Movie Reviews corpus.

eSOL

Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY

Download location: *hidden*

Distribution Access/Medium:

Downloadable

Contact Persons

Luis Alfonso Ureña López

Eugenio Martínez Cámara

Lexical Conceptual Resource General Information

Lexicon

Encoding

Encoding level: Semantics

Linguistic information: Lemma

Creation

Creation mode: Automatic

eSOL

Monolingual text lexicalConceptualResource

Languages

Spanish

Linguality

Linguality type: Monolingual

Text Format

text/plain

Size

5,639 (negative words) Words

2,535 (positive words) Words

Character encoding

ISO - 8859 - 1

Domains

movie reviews

Resource Creation

Resource Creator

M. Dolores Molina González

Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)

María Teresa Martín Valdivia

José María Perea Ortega

Eugenio Martínez Cámara

Funding Project

Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)

URL: http://intime.dlsi.u...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Trends and subjects analysis through opinions and feelings (ATTOS - TIN2012-38536-C03-0)

URL: http://gplsi.dlsi.ua...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Flexible interactive reading support tool (First - FP7-287607)

URL: http://www.first-asd...
Funding Type: Eu Funds
Funder: European Union

Metadata

Created: 09/25/2014

Last Updated: 10/30/2012

Source: METASHARE

v1.1

Metadata Creator

Silvia Arano

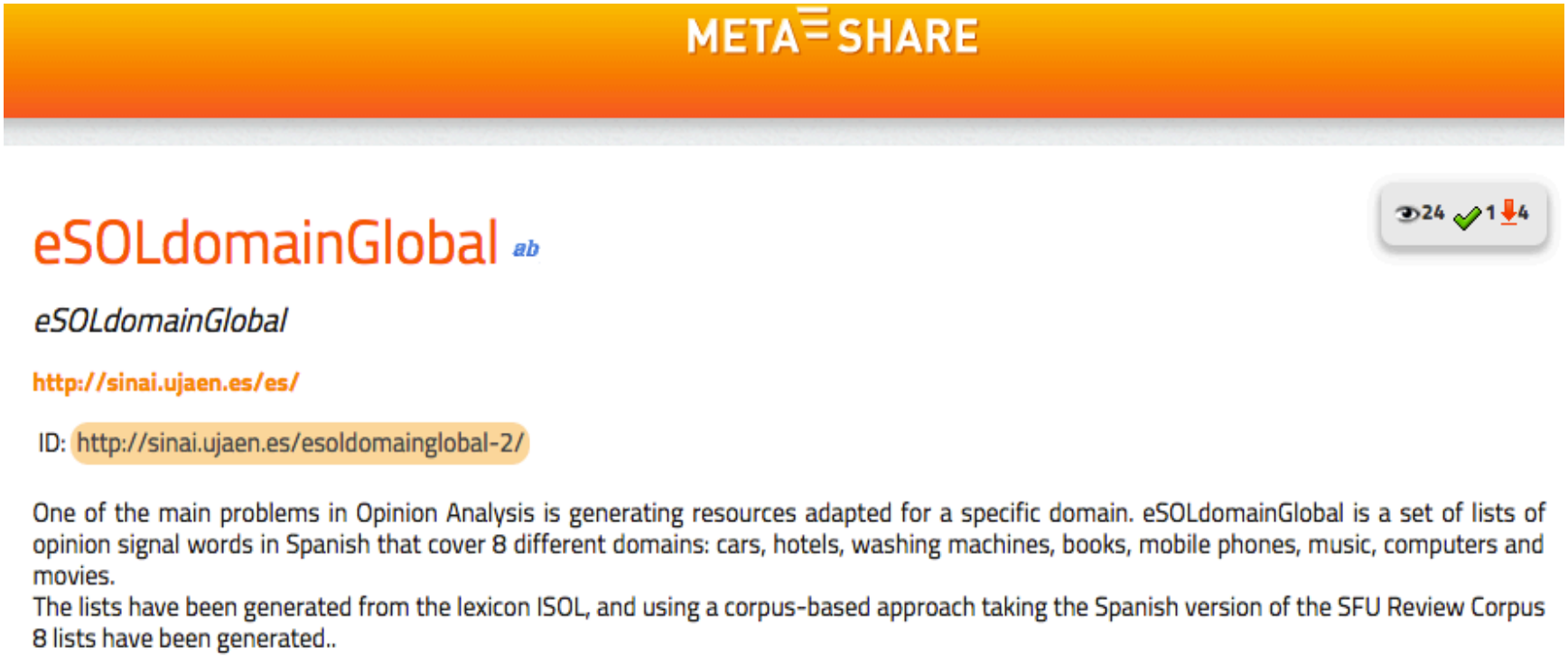
Version

Version: 1

Last Updated: 01/01/2012

eSOLdomainGlobal

URL en MetaShare:
http://metashare.upf.edu/repository/browse/esoldomainglobal/886da1206a7a11e48763000c291ecfc830955df93dd4442c8eb5360b00498c8d/



eSOLdomainGlobal

Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY

Download location: *hidden*

Distribution Access/Medium:

Downloadable

Contact Persons

Luis Alfonso Ureña López

Eugenio Martínez Cámara

Lexical Conceptual Resource General Information

Lexicon

Encoding

Encoding level: Semantics

Linguistic information: Lemma

Creation

Creation mode: Automatic

Metadata

Created: 09/25/2014

Last Updated: 10/30/2012

Source: METASHARE

v1.1

Metadata Creator

Silvia Arano

Version

Version: 1

Last Updated: 01/01/2013

eSOLdomainGlobal

Monolingual text lexicalConceptualResource

Languages

Spanish

Linguality

Linguality type: Monolingual

Text Format

text/plain

Size

2,535 (movies: positive) Words
5,648 (movies: negative) Words
2,527 (computers: positive) Words
5,644 (computers: negative) Words
2,538 (music: positive) Words
5,645 (music: negative) Words
2,528 (cars: positive) Words
2,529 (mobile phones: positive) Words
5,657 (mobile phones: negative) Words
2,520 (washing machines: positive) Words
5,639 (washing machines: negative) Words
2,529 (books: positive) Words
5,651 (books: negative) Words
5,648 (cars: negative) Words
2,517 (hotels: positive) Words
5,636 (hotels: negative) Words

Character encoding

ISO - 8859 - 1

Domains

- hotels
- books
- washing machines
- music
- mobile phones
- movies
- computers
- cars

Resource Creation

Resource Creator

M. Dolores Molina González
Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)
María Teresa Martín Valdivia
L. Alfonso Ureña López
Eugenio Martínez Cámara

Funding Project

Text processing from collaborative and on-line sources (TEXT-COOL 2.0 - TIN2009-13391-C04-02)
URL: http://intime.dlsi.u...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Trends and subjects analysis through opinions and feelings (ATTOS - TIN2012-38536-C03-0)
URL: http://gplsi.dlsi.ua...
Funding Type: National Funds
Funder: España. Ministerio de Ciencia e Innovación
Funding Country: Spain

Flexible interactive reading support tool (First - FP7-287607)
URL: http://www.first-asd...
Funding Type: Eu Funds
Funder: European Union

CRiSOL

- Resource type: List of opinion bearing words
- Linguality type: Monolingual
- Languages: Spanish
- Text format: xml
- Size
  - Number of entries: 8135 words
    - Positive: 2,509
    - Negative: 5,626
    - With SentiWordNet score: 4,434
- Character encoding: utf-8

CRiSOL

- Domains: General domain
- Encoding
  - Encoding level: Opinion (semantic)
  - Linguistic information: Word
- Creation mode: Automatic
- Version: 1.0
- Documentation: Molina González, M. Dolores, Martínez Cámara, Eugenio, & Martín Valdivia, M. Teresa. (2015). CRiSOL: Base de Conocimiento de Opiniones para el Español. Procesamiento Del Lenguaje Natural, 55, 143-150. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5226

CRiSOL

Distribution

- Availability: Available – Unrestrictive Use
- License
  - License CC-By
  - Restrictions
  - Download location: http://sinai.ujaen.es/wp-content/uploads/2015/09/crisol.tar.gz
  - Distribution Access/Medium: Downloadable

Contact Person: M. Dolores Molina González, Eugenio Martínez Cámara

Resource Creation

- Resource creator: M. Dolores Molina González, Eugenio Martínez Cámara and M. Teresa Martín Valdivia

CRiSOL

- Funding project:
  - Name: Análisis de Tendencias y Temáticas a través de Opiniones y Sentimientos (ATTOS, TIN2012-38536-C03-0)
    - URL: http://attos.ujaen.es
    - Funding Type: National Funds
    - Funder: España. Ministerio de Economía y Competitividad
    - Funding Country: España

CRiSOL

- Funding project:
  - Name: Análisis de Opinión en Redes Sociales y Contenidos Generados por Usuarios (AORESCU, P11-TIC-7684 MO)
    - URL:
    - Funding Type: Regional Funds
    - Funder: Junta de Andalucía
    - Funding Country: España

CRiSOL

- Funding project:
  - Name: Captura y Procesamiento de Grandes Volúmenes de Datos Provenientes de varias Fuentes de Micro-blogging (CEATIC-2013-01)
    - URL:
    - Funding Type: University Funds
    - Funder: Centro de Estudios Avanzados en TIC. Universidad de Jaén
    - Funding Country: España

Emoti-SP

- Resource type: Lexicon
- Linguality type: Monolingual
- Languages: Spanish
- Text format: Text/plain
- Size
  - Number of entries: 70 positive emoticons and 46 negative emoticons
- Character encoding: UTF-8

Emoti-SP

- Domains: General
- Encoding
  - Encoding level: Semantic
  - Linguistic information: Token
- Creation mode: Mixed
- Version: 1
- Documentation:

SINAI-ESMA: An unsupervised approach for Sentiment Analysis in Twitter. Salud María Jiménez Zafra, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, L. Alfonso Ureña López

Emoti-SP

Distribution

- Availability: Available – Restricted use
- License
  - License: Creative Commons
  - Restrictions: Research purposes
  - Download location: http://sinai.ujaen.es/emoti-sp/
  - Distribution Access/Medium: Email

Contact Person: Eugenio Martínez Cámara (emcamara@uajen.es)

Emoti-SP

Resource Creation

- Resource creator: SINAI Research Group, Universidad de Jaén
- Funding project: Partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), ATTOS project (TIN2012-38536-C03- 0) from the Spanish Government, AORESCU project (P11-TIC-7684 MO) from the regional government of Junta de Andalucía and CEATIC- 2013-01 project from the Universidad de Jaén.

Hashtags-SP

- Resource type: Lexicon
- Linguality type: Monolingual
- Languages: Spanish
- Text format: Text/plain
- Size
  - Number of entries: 172 positive hashtags and 127 negative hashtags
- Character encoding: UTF-8

Hashtags-SP

- Domains: General
- Encoding
  - Encoding level: Semantic
  - Linguistic information: Token
- Creation mode: Mixed
- Version: 1
- Documentation:

SINAI-ESMA: An unsupervised approach for Sentiment Analysis in Twitter. Salud María Jiménez Zafra, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, L. Alfonso Ureña López

Hashtags-SP

Distribution

- Availability: Available – Restricted use
- License
  - License: Creative Commons
  - Restrictions: Research purposes
  - Download location: http://sinai.ujaen.es/hashtags-sp/
  - Distribution Access/Medium: Email

Contact Person: Salud María Jiménez-Zafra (sjzafra@ujaen.es)

Hashtags-SP

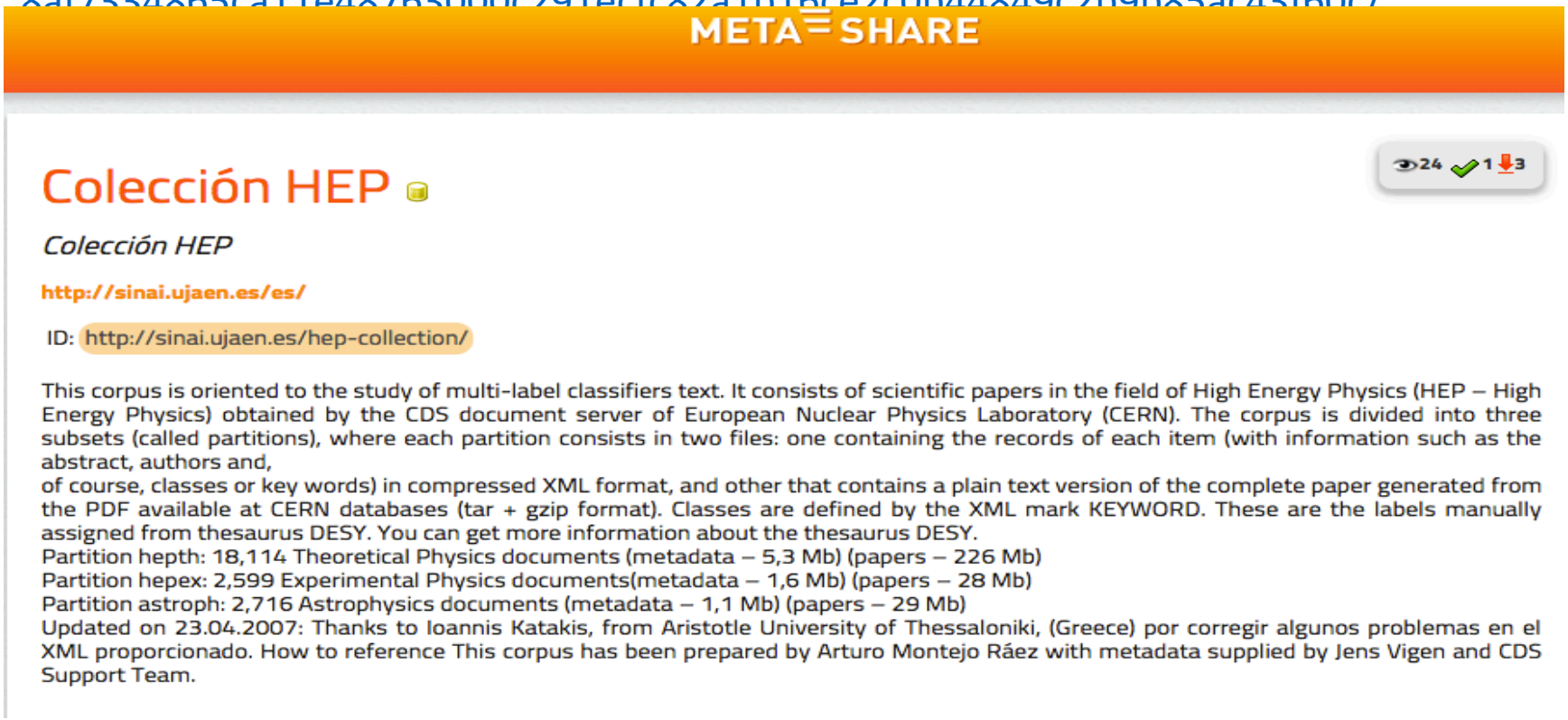
Resource Creation

- Resource creator: SINAI Research Group, Universidad de Jaén
- Funding project: Partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), ATTOS project (TIN2012-38536-C03- 0) from the Spanish Government, AORESCU project (P11-TIC-7684 MO) from the regional government of Junta de Andalucía and CEATIC- 2013-01 project from the Universidad de Jaén.

CLASIFICACIÓN DE TEXTOS

Colección HEP

URL en MetaShare:
http://metashare.upf.edu/repository/browse/coleccion-hep/8af7334865ca11e48763000c291ecfc82a1b16ce2c0b44849c2b9b85ac43f60c/



META=SHARE

24 1 3

Colección HEP

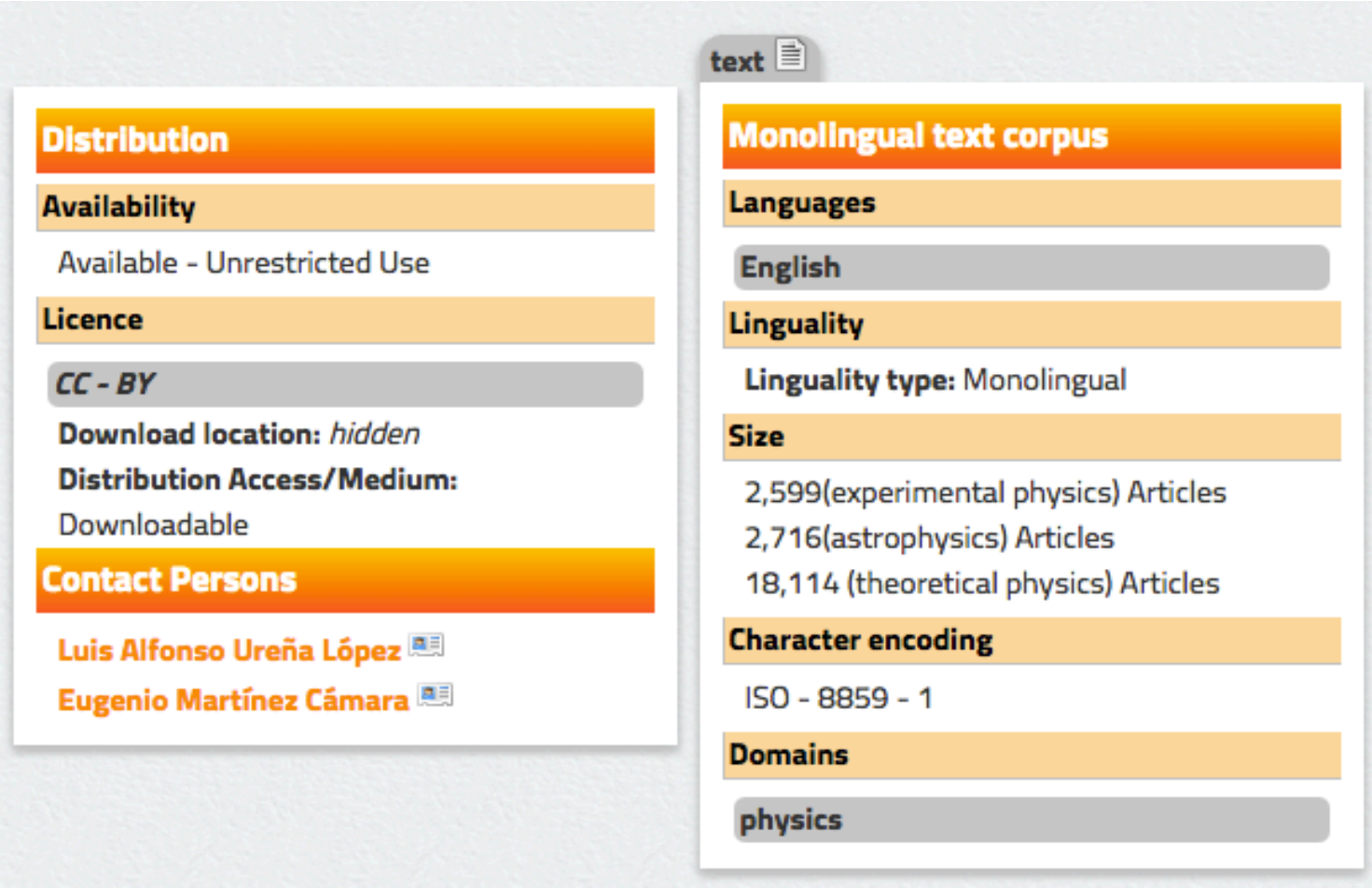
Colección HEP

http://sinai.ujaen.es/es/

ID: http://sinai.ujaen.es/hep-collection/

This corpus is oriented to the study of multi-label classifiers text. It consists of scientific papers in the field of High Energy Physics (HEP – High Energy Physics) obtained by the CDS document server of European Nuclear Physics Laboratory (CERN). The corpus is divided into three subsets (called partitions), where each partition consists in two files: one containing the records of each item (with information such as the abstract, authors and,
of course, classes or key words) in compressed XML format, and other that contains a plain text version of the complete paper generated from the PDF available at CERN databases (tar + gzip format). Classes are defined by the XML mark KEYWORD. These are the labels manually assigned from thesaurus DESY. You can get more information about the thesaurus DESY.
Partition hepth: 18,114 Theoretical Physics documents (metadata – 5,3 Mb) (papers – 226 Mb)
Partition hepex: 2,599 Experimental Physics documents(metadata – 1,6 Mb) (papers – 28 Mb)
Partition astroph: 2,716 Astrophysics documents (metadata – 1,1 Mb) (papers – 29 Mb)
Updated on 23.04.2007: Thanks to Ioannis Katakis, from Aristotle University of Thessaloniki, (Greece) por corregir algunos problemas en el XML proporcionado. How to reference This corpus has been prepared by Arturo Montejo Ráez with metadata supplied by Jens Vigen and CDS Support Team.

Colección HEP



Colección HEP

Resource Creation

Resource Creator

L. Alfonso Ureña López

Ralf Steinberger

Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)

Arturo Montejo Ráez

Funding Project

Recuperación de Respuestas en Documentos Digitalizados (R2D2 - TIC2003-07158-C04-04)

Funding Type: National Funds

Funder: España. Ministerio de Ciencia y Tecnología

Funding Country: Spain

Metadata

Created: 09/25/2014

Last Updated: 10/30/2012

Source: METASHARE

v1.1

Metadata Creator

Silvia Arano

Version

Version: 1

Last Updated: 01/01/2004

BÚSQUEDA DE RESPUESTAS

Spanish QC

URL en MetaShare:
http://metashare.upf.edu/repository/browse/spanish-qc/



Spanish QC

13 1 2

Spanish QC

http://sinai.ujaen.es/es/

ID: http://sinai.ujaen.es/spanish-qc-2/

This resource are 6305 spanish tagged questions for question answering classification, following the taxonomy defined in the paper "X. Li and D. Roth. Learning Question Classifiers", and having the following general and detailed categories:
ABBR: abbreviation, expansion
DESC: definition, description, manner, reason
ENTY: animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM: description, group, individual, title
LOC: city, country, mountain, other, state
NUM code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight.
Starting from a set of labeled questions for English, it has created this resource with various questions in Spanish labeled and checked by 3 people.

Spanish QC

Distribution

Availability

Available - Unrestricted Use

Licence

CC - BY

Download location: *hidden*

Distribution Access/Medium:

Downloadable

Contact Persons

Luis Alfonso Ureña López

Eugenio Martínez Cámara

Monolingual text corpus

Languages

Spanish

Linguality

Linguality type: Monolingual

Size

6,305 (questions)

Character encoding

ISO - 8859 - 1

Domains

general

Spanish QC

Resource Creation

Resource Creator

L. Alfonso Ureña López

Fernando Martínez Santiago

Universidad de Jaén. Grupo de investigación Sistemas Inteligentes de acceso a la información (SINAI)

Miguel Ángel García Cumbreras

Funding Project

Recuperación de Respuestas en Documentos Digitalizados (R2D2 - TIC2003-07158-C04-04)

Funding Type: National Funds

Funder: España. Ministerio de Ciencia y Tecnología

Funding Country: Spain

Metadata

Created: 09/25/2014

Last Updated: 10/30/2012

Source: METASHARE

v1.1

Metadata Creator

Silvia Arano

Version

Version: 1

Last Updated: 01/01/2006

Gracias por su atención

¿Alguna pregunta?