Linguistic Resources of the LyS (Language and Information Society) Group

ReTeLe Workshop (SEPLN)

Marcos Garcia
marcos.garcia.gonzalez@udc.gal

Grupo LyS, Universidade da Coruña

September 13, 2016

DiCE

- Dictionary
- Corpus

SentiStrength

- Corpus
- Lexicon

EN-ES-CS Corpus

UD Galician-TreeGal

Sentiment Analysis Tools

DiCE: Diccionario de Colocaciones del Español

- Dictionary of collocations
- General Spanish
- Size:
  - Lemmas: 212 (e.g., *gana*)
  - Lexical units: 556
    - *sentimiento*
    - *sensación*
    - *actitud*
  - Collocations: 21,269
    - "Sara tenía muchas **ganas** de ir a aquella tienda"
    - "¿La merienda nos quitará las **ganas** de cenar?"
    - "El equipo va con todas las **ganas**"

DiCE: Diccionario de Colocaciones del Español

- Current version: *sentiment* nouns
- Encoding:
  - Semantic tag
  - Actantial structure
  - Usage examples (mostly from CREA)
  - Collocation frequency
  - Quasi-synonyms / quasi-antonyms
  - Syntactic schema
  - Collocations

DiCE: Diccionario de Colocaciones del Español

- Manually created
- Version: >1.0 (database)
- Documentation: `www.dicesp` + papers
- Availability: online
- Contact Person:
  - Margarita Alonso Ramos
  - `diccionariodecolocaciones@gmail.com`
- Resource Creation:
  - Creator: Margarita Alonso Ramos + several linguists
  - Funding projects: PGIDT99PXI10401B (Xunta de Galicia) + COLOCATE (MINECO) + FPUs

DiCE Corpus

- General Spanish
- Mostly from CREA (Corpus de Referencia del Español Actual, RAE)
- Size:
  - Collocations: 30,643 (506,054 words)
  - Lexical Units: 1,381 (18,345 words)
- Manually annotated
- Queries from CREA (and other resources)

DiCE Corpus

- Availability: online
- Contact Person:
  - Margarita Alonso Ramos
  - diccionariodecolocaciones@gmail.com
- Resource Creation:
  - Creator: Margarita Alonso Ramos + DiCE team (LyS Group)
  - Funding projects: PGIDT99PXI10401B (Xunta de Galicia) + FFI2008-06479-C02-01 (MICINN)

SentiStrength Spanish Corpus

- Polarity corpus for Spanish
- Type of language: Twitter
- Plain text
- Size:
  - 3,200 tweets (1,600 dev / 1,600 test)
- UTF-8 Unicode

SentiStrength Spanish Corpus

- Annotation (both POS and NEG values):
  - POS: 1 $\rightarrow$ 5
  - NEG: -1 $\rightarrow$ -5
- Creation:
  - Downloaded with Twitter API (2014)
  - 3 annotators (from 7)
- Doc.(paper): Vilares et al., 2015 (JIS journal)

SentiStrength Spanish Corpus

- Distribution:
  - License:
    - GNU GPL
    - `http://sentistrength.wlv.ac.uk/SpanishTweetsTestAndDevelopmentSetsDavidVilares.zip`
  - Contact Person: David Vilares (`david.vilares@udc.es`)
- Resource creation
  - Creator: David Vilares (+3 annotators)
  - Funding projects:
    - FFI2014-51978-CS (MINECO)
    - R2014/034 (Xunta de Galicia)
    - FPU13/01180
    - Inditex UDC 2014

SentiStrength Spanish Lexicon

- Polarity lexicon for Spanish
- Type of language: Twitter
- Plain text
- Size: 26,752 entries
- UTF-8 Unicode
- Anotation:
  - POS: 1 $\rightarrow$ 5
  - NEG: -1 $\rightarrow$ -5

SentiStrength Spanish Lexicon

- Creation: expanded from existing lexica (SentiStrenght + Brook et al., 2009)
- Doc. (paper): Vilares et al., 2015 (JIS journal)
- Distribution:
  - License: GNU GPL
  - `http://sentistrength.wlv.ac.uk/SpanishSentiDataDavidVilares.zip`
  - Contact Person: David Vilares (`david.vilares@udc.es`)
- Resource creation:
  - Creator: David Vilares
  - Funding projects:
    - FFI2014-51978-CS (MINECO)
    - R2014/034 (Xunta de Galicia)
    - FPU13/01180
    - Inditex UDC 2014

EN-ES-CS Corpus

- Tweets with code-switching
- English + Spanish (in the same tweet)
- Plain text (polarity) + download (API)
- Size:
  - Tweets: 3,062
  - English: 24,758 tokens / 5,565 unique / 3,576 OOV
  - Spanish: 16,174 tokens / 5,033 unique / 3,714 OOV

EN-ES-CS Corpus

- Annotation:
  - Manual (3 annotators)
  - Polarity (SentiStrenght: 5 POS + 5 NEG)
  - Agreement: ≈ 0.65
- Based on a previous dataset (Solorio et al., 2014)
- Version: 0.1
- Doc. (papers): WASSA 2015 + LREC 2016

EN-ES-CS Corpus

- Distribution:
  - License: LGPLLR (LGPL for Linguistic Resources)
  - `http://grupolys.org/software/CS-CORPORA`
  - Contact person: David Vilares (`david.vilares@udc.es`)
- Resource creation:
  - David Vilares, Carlos Gómez, Miguel A. Alonso (and Thamar Solorio team)
  - Funding projects:
    - FFI2014-51978-CS (MINECO)
    - R2014/034 (Xunta de Galicia)
    - FPU + Oportunius Grants (Xunta)

UD Galician-TreeGal Corpus

- Generic news (*xeral* subcorpus of XIADA)
- Galician (ILG/RAG spelling)
- Plain text (CoNLL-U format)
- Size:
  - 1,000 sentences
  - 24,219 tokens
- UTF-8 Unicode

UD Galician-TreeGal Corpus

- Annotation (from XIADA)
  - Token, lemma, POS-tag
- Added annotation (TreeGal)
  - UD POS-tags + UD (dependency labels)
- Creation:
  - Automatic conversion from XIADA
  - Cross-lingual parsing + manual correction
- Version: 0.3
- Doc.: guidelines + SEPLN paper

UD Galician-TreeGal Corpus

- Distribution:
  - License: LGPLLR
  - `http://grupolys.org/~marcos/resources/syntax/UD_Galician-TreeGal/`
  - Next UD release (?)
  - Contact person: Marcos Garcia (`marcos.garcia.gonzalez@udc.gal`)
- Resource creation:
  - Creator: Marcos Garcia (+XIADA) + revision (LyS Group)
  - Funding:
    - FJCI-2014-22853 (MICINN)
    - FFI2014-51978-C2 (MICINN)
    - Oportunius Grant (Xunta de Galicia)

Sentiment Analysis Tools

- MIOPIA (2015)
  - Syntax-based, language-dependent
  - `https://miopia.grupolys.org/`
- Samulan (2016)
  - Universal, Unsupervised, Uncovered
  - Syntax-based, language independent
  - `http://grupolys.org/software/UUUSA/`
- Future:
  - Multilingual parsers (for SA)
  - Universal + language-specific rules

Thanks!

Questions?

grupolys.org