Cleansing and expanding the HURTLEX-EL with a multidimensional categorization of offensive words

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou
Alexandra Saivanidou and Stella Markantonatou

Institute for Language and Speech Processing, Athena R.C.

{vivianstamou, iakovi.alexiou, antyklimi, moloueleftheria,
alexansaivan, stiliani.markantonatou}@gmail.com

ATHENA
Research & Innovation
Information Technologies

In a snapshot

- Publically available cleansed version of the Modern Greek (MG) branch of HURTLEX containing 737 entries marked for **context dependence and type of target of offence**
- The lexicon was developed with manual inspection in three rounds under a prescriptive approach
- Development of diagnostic criteria for offensive word identification; social and cultural aspects were highlighted

Offensive Speech resources in MG

Lexicons:

- Efthymiou et al. (2014): categories of derogatory words
- Christopoulou (2012) and Xydopoulos (2012): methods for offensive language identification

Datasets:

- Offensive Greek Tweet Dataset (OGTD): the first MG dataset annotated for offensive language (Pitenis et al. 2020)
- Hate speech has been investigated by:
  - Lekea and Karampelas (2018): terrorist argument; 1265 words (unpublished)
  - Perifanos and Goutsos (2021): HS detection; Twitter; multimodal approach; hateful, xenophobic and racist tweets; no annotated corpus/lexicon published

No detailed annotation guidelines
No labelled corpora representing a wide range of registers

HURTLEX

- **Domain-independent lexicon**; 53 languages; lists offensive, aggressive and hateful words
- Supports the development of resources for **underrepresented languages**
- Its kernel contains about 1000 manually selected words classified in **17 fine-grained thematic categories**; then semiautomatically enriched via Multiwordnet & Babelnet

The revised Modern Greek branch of HURTLEX

Step 1: Lexicon Cleansing by two linguists:

- foreign words (either English or French)
- meaningless multitoken terms i.e., πουτίγκα κεφάλι, Lit. pudding head
- words with errors: φυσιογνωμονική 'physiognomic' instead of φυσιογνωμική 'physiognomic'
- multitoken terms with agreement errors: σεξουαλικά επίθεση for σεξουαλική επίθεση 'sexual assault'
- different inflectional forms of the same lemma were merged into the lemma
- archaic words: αιχμαλωτίζων 'capturer'

→ **2143 words were retained out of the 3114 original HURTLEX entries**

Step 2: Annotation by four under-graduate linguists:

- One of the three labels was assigned: context-independent, context-dependent, non-offensive
- General offensiveness diagnostics mainly about profane and obscene language were used
- Interannotator agreement score 0.77 (Fleiss kappa)–indicating an already substantial agreement

Step 3: Annotation and Bottom up development of Diagnostics by four under-graduate linguists:

- 3 rounds of discussions among annotators and group leaders to detail the criteria
- Final annotation scheme:
  1. **context-independent, context-dependent, non-offensive**
  2. **development of 17 diagnostics**
  3. **target of offence: individual, group, non-human, event-property-state**
- Interannotator agreement score 0.96 (Fleiss kappa)

Only offensive terms shared by all annotators were included in the cleansed MG HURTLEX.

Word Distribution

Most populated diagnostics: Behavior, Crime & immoral behavior, Animals:

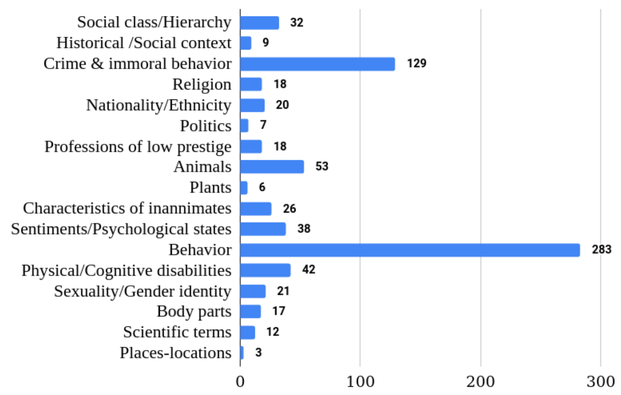

Figure 1. Word distribution per diagnostic.

The Diagnostics

	Classes	OL Target	Context Ind.	Context Dep.
1	Social class/ hierarchy	indv., groups		+
2	Historical/ social context	indv., groups, ESP		+
3	Crime immoral behavior	indv., groups, ESP	+	+
4	Religion	indv., groups, ESP		+
5	Nationality ethnicity	indv., groups	+	+
6	Politics	indv., groups, ESP	+	+
7	Professions of low prestige/sexual occup.	indv., groups	+	+
8	Animals	indv., groups, non-human		+
9	Plants	indv., groups, non human		+
10	Characteristics of inannimates	indv., groups, non-human		+
11	Sentiments, psychological states	indv., ESP	+	+
12	Behavior	indv., groups, ESP	+	+
13	Physical/ cognitive disabilities, appearance	indv., groups, non humans	+	+
14	Sexuality gender identity	indv., groups, ESP	+	+
15	Body parts	indv., groups, ESP, non-human	+	+
16	Scientific terms	indv., groups, ESP, non-human		+
17	Places-locations	indv., groups, ESP, non-human	+	

Table 1. Presentation of the OL diagnostics.

Comparison to orginal HURTLEX

Historically/culturally marked MG diagnostics deviate from the HURTLEX categories.

- HURTLEX "SVP: seven deadly sins of the Christian tradition" vs "Religion reflects tendencies of Greek society: terms about different religions or religious states"
- HURTLEX "IS—social class/ hierarchy" vs "Social class/ hierarchy: terms about social and economic (dis)advantages, e.g., νεόπλουτος 'nouveau riche'"
- New diagnostic 2 "Historical / social context: contemporary terms particular to Greek history, e.g., κλέφτες 'armatole / militiamen' (Greek armed groups of the Ottoman occupation era)". HURTLEX distributes these terms in "Potential negative connotations (QAS)", "Derogatory words (CDS)" and, "Felonies and words related to crime and immoral behavior(RE)"
- New diagnostic 5 "Nationality/ ethnicity: terms about nationalities/minorities within the Greek ethnicity reflecting social and cultural differentiation, e.g., 'Jew', 'gypsy'

Aknowledgments

We acknowledge support of this work by the project "PHILOTIS: State-of-the-art technologies for the recording, analysis and documentation of living languages" (MIS 5047429), which is implemented under the "Action for the Support of Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).