

# Computational Linguistic Resources / NLP Applications for Romanian developed at ICIA

Some of the resources presented below are described in detailed enough articles for those interested. The standard documentation of all resources will contain links to all relevant articles or research reports. Corpora and lexical resources will contain the minimal documentation elements specified by TEI/CES header.

## I. Corpora

### I.1. Manually validated corpora (tokenized, lemmatized, tagged)

- **NAACL 2003:** English-Romanian parallel corpus, containing around 1.6 million tokens in the two languages. The corpus is segmented, morpho-syntactically annotated and lemmatized. The Romanian part has been manually checked at word level (during around six months) to correct the annotation/segmentation. When checking the segmentation, we marked the expressions using the “\_” character (e.g. *au dat-o\_în\_bară* with the lemma “*a o da\_în\_bară*”). The corpus content differs from the initial version (distributed by Rada Mihalcea) because we inserted the missing diacritics.
- **Orwell’s 1984:** English-Romania parallel corpus containing approx. 250 thousand tokens in the two languages. Both parts have been manually checked at word level, and, recently, we have operated some modifications that, besides the standard corrections, also include the harmonizing of the tagset used, recognition of the expressions that either are translated by a single word in Romanian or have a separate entry in the Princeton WN 2.0. Around 2000 occurrences of some English words have been manually annotated at the sense level. In the near future, we will annotate the English part with the FDG parser and these annotations will be imported in Romanian as well.
- **Republica (Plato’s Republic):** French-Romanian parallel corpus containing around 250 thousand tokens. It is morpho-syntactically annotated.
- **Ziare (Newspapers):** corpus containing various articles from different issues of the daily *Evenimentul Zilei* (from 1995 to 1996). It contains approx. 92 thousand tokens and is morpho-syntactically annotated.
- **ROCO (Ziare):** Romanian newspaper corpus containing around 6.7 million tokens, that are morpho-syntactically annotated.
- **SEMCOR-RO:** this is a translation of the SEMCOR corpus (about 1 million tokens), dependency-linked, word-sense disambiguated and word-aligned to the English original
- **See-Era/Net corpus:** this is a 8-language parallel corpus, extracted from JRC-Acquis, of about 1,3 million words/language, word-aligned to English. English and Romanian parts are dependency-linhd. The languages are: Bulgarian, Czech, English (the hub), French, German, Greek, Romanian, Slovene

### I.2. Automatically annotated corpora corpora (tokenized, lemmatized, tagged)

- **FrameNet:** Romanian translation of a part of the FrameNet corpus 1.1. The corpus is lemmatized and morpho-syntactically annotated, containing around 25 thousand tokens.
- **Timex:** Romanian translation of a part of the Timex corpus. It is a parallel (English-Romanian) corpus, containing approx. 72 thousand tokens in the Romanian part, it is lemmatized and morpho-syntactically annotated.

- **RoSemCor**: English-Romanian-Italian parallel corpus, annotated at the sense level using WSDTool.
- **JRC-Acquis** (19,200 documents in Romanian) about 30,000,000 tokens
- **Romanian Wikipedia (edition 2006)** : about 2,500,000 tokens.

**I.3. Un-annotated Corpora** – around 130 volumes (fiction, scientific literature in informatics, law, medicine, etc.); 18 of them are in at least two languages (one of them being Romanian). Most of them have explicit copyright restrictions.

## II. Dictionaries/Lexicons

- **WEB-DEX**: Romanian explanatory dictionary, XML-encoded following the CONCEDE requirements. Around 65,000 entries (based on DEX 1996). The interface allows for creating complicated queries for searching the dictionary (e.g. “what word of German origin is synonym with *condică* and begins with letter *r*?”).
- **tbl.wordform.ro**: ASCII file containing around 1,300,000 tokens of Romanian words. Each token is annotated with an extended tag (called MSD), tags that belong to a set of approximately 700 morpho-syntactic descriptions for Romanian. Besides the morpho-syntactic annotation, each token is hyphenated and lemmatized. **tbl.wordform.ro** was created by collecting all words from the manually validated Romanian corpora, alongside with their morpho-syntactic annotations and lemmatizations.
- **ROPMORPH**: paradigmatic morphology of Romanian. It contains 171 paradigms for nouns, 58 paradigms for verbs, 19 paradigms for demonstrative adjectives, articles and pronouns.
- **Paradigmatic morphology of Romanian**: unification-based description of paradigmatic morphology; it is a lexicon of around 40000 lemmas; the description includes a subset of derived words. The morphologic generator allows the generation of more than 1,200,000 tokens accompanied by the attribute-value description.
- **RoWordNet**: lexical semantic network of Romanian. It is aligned at the conceptual level with the English WordNet with Princeton WordNet 2.0, SUMO&MILO ontologies, the IRST DOMAINS taxonomy and includes the SentiWordNet subjectivity mark-up. It contains, currently, around 55,000 synsets and it is continuously extended.
- **French-Romanian Dictionary**: contains 16710 entries. Its XML implementation is compatible with the TEI-light specifications for bilingual dictionaries. The interface allows for search with the help of regular expressions.
- **EUROVOC** – multilingual thesaurus (21 languages, including Romanian) used for Acquis Communautaire documents indexing and classification.

Besides these reference lexical resources there are many bilingual dictionaries automatically extracted from our parallel corpora, including a seven-language dictionary extracted from the parallel corpus “1984” (partially validated).

### III. Programs useful for NLP applications

- **EGLU:** unification-based integrated programming environment for complex NLP systems. It includes a compiler for linguistic descriptions and modules for morphologic analysis and generation, for syntactic analysis (CKY), (head-driven) syntactic generation, lexical or structural transfer for automatic translation. The system is implemented in Common Lisp.
- **PAIL:** integrated training environment for Artificial Intelligence. It contains 10 modules: two in the domain of syntactic analysis, based on Augmented Transition Networks (ATN) and chart parsing. The two modules can be combined with the module of automatic theorem prover to create systems for natural language understanding. PAIL is implemented in Common Lisp.
- **DIC:** Electronic dictionaries compiler. It was created to automatically generate the XML Concede coding from the typographical format (MSWord) of DEX. With minimal transformations, it can be used for the compilation of other dictionaries that use lexicographical and typographical conventions similar to those used in DEX.
- **TTAG:** Automatic optimal tag set generation system (Perl implementation), starting from a MULTEXT-EAST-type specification of the lexical descriptions (dictionary tagset). In this context, optimality means the minimal number of morpho-lexical labels (without information loss) necessary for the training of a statistical morpho-lexical analyzer. Minimal tag sets have been generated for the six languages of the MULTEXT-EAST project, for which the necessary resources were available. The ratio of the MSD number and the corpus labels (c-tag) number varies, from a language to another, between 5 and 11. The drastic reduction of the morpho-lexical label number, without information loss, allows more robust statistical language model generation; these models are unaffected (or significantly less affected) by the training data insufficiency.
- **CrossTAG:** morpho-syntactic annotation system that uses a second tag set to improve the annotation of the first one. As a collateral effect, a correspondence between the two label sets results.
- **Maximum Entropy Tiered Tagger:** The application is used for annotating with MSD tags, using a reduced tag set in an intermediate stage. For highly-inflected languages, it shows better results than conventional part of speech annotation methods. Using the ME classifier, the application does not require a tag conversion dictionary (from c-tag to MSD) for every word form.
- **TTL:** PERL module for text segmentation at sentence/word level, morpho-syntactic annotation (simple and layered – tiered tagging) and lemmatization. It is language independent, using regular expressions and Markov models; alternatively, MtSeg and TnT applications can be used for segmentation/morpho-syntactic annotation.
- **COLLOC:** a multilingual collocation extraction (we experimented in Romanian, English and French) capable of finding non-contiguous collocations.
- **WSDTool:** PERL application that annotates at the sense level every content word of a parallel corpus in XCES format (ICIA variant). The user

can choose to annotate any combination of parts from the parallel corpus for which aligned wordnets exist.

- **ROG:** A generator of inflected forms given a lemma and a morpho-lexical description (MSD)
- **TREQ:** PERL application that extracts translation equivalents dictionaries from parallel corpora;
- **YAWA:** lexical aligner in PERL. It is language independent, except for the modules that realise alignments specific to the pairs of aligned languages. So far, it works just for Ro-En pair of languages. It requires a parallel corpus in XCES format, morpho-syntactically annotated and lemmatized, and translation dictionaries produced by TREQ.
- **MEBA:** lexical aligner implemented in C#. It uses Giza++ for translation equivalents extraction. It is language independent, using different thresholds and weights for every feature of a possible lexical alignment.
- **COWAL:** Lexical alignment combining system. It was tested using MEBA and YAWA results as input.
- **Hyphenator:** automatic hyphenation system for Romanian. PERL application.
- **DIAC:** system that automatically recovers missing diacritics in Romanian texts; although it was used only for Romanian, the system can be used for other languages too, if it has proper resources (a morpho-lexical disambiguated text, containing diacritics).
- **Multilingual thesaurus aligner:** the system was developed for EUROVOC English (integral) variant with the incomplete variant for Romanian. The main goal of this alignment was the automatic recovery of unique, language independent, CELEX codes for every term; these codes were absent in the Romanian variant of EUROVOC. Although it was developed for solving specific problems, the system has a general utility in aligning taxonomical conceptualizations with multiple lexicalizations. C# application.
- **MtKit:** integrated environment that performs the lexical annotation/alignment of XCES corpora. It allows the construction of statistical translation models and has an incorporated user-friendly graphical editor which ensures the visualisation of lexical alignments and of the properties specific to every alignment constituent (POS-tag, lemma, syntactic group it belongs to, sense number and the corresponding definition), as well as the modification of incorrect or incomplete alignments. It also allows the automatic calculus of alignment accuracy (precision, recall, F-measure) against a reference alignment (gold-standard). C# application.
- **Sentence Aligner:** sentence level alignment statistical application. It combines the method based on the length of the text to be aligned (Gale& Church type) with lexical methods (Moore type, but without limiting only to 1:1 alignments); very good accuracy. C# application.
- **LexPar:** PERL application that determines the structure of a connected, acyclic and planar graph of a given sentence. It uses language specific rules to reduce the searching space. Thus one obtains a linker with better chances for discovering only the syntactical meaningful links.
- **SynWSD:** PERL application (under development) that annotates at sense level all the content words of a given text. It is trained on texts previously

annotated with TTL/LEXPART and disambiguates texts annotated in the same way. It uses WordNet in XML BalkaNet format as a sense inventory.

- **XCESGen:** PERL scripts series that generates parallel corpora in XCES format:
  - Metacategories annotation: every word receives a category that defines a tag-set subset.
  - Chunking: adjacent word phrases are marked and named: noun phrases, verb phrases, prepositional phrases, etc.
  - lemma/morpho-syntactic label annotation: it uses TTL.
  - Sense annotation: with WSDTool/SynWSD.
  - Link annotation: with LexPar.
  
- **Language Identifier:** application that automatically identifies the language of a text written in one of the 21 European Union languages. C# application.
- **GoogleSearch:** tools library for searching on the Internet, using Google search engine. Type of application: Screen Scrapper.
- **Two prototypes of MT**
  - **STAR:** a phase-based statistical translation system, using the MOSES factored translation decoder; works very well on Romanian-English and English-Romanian (legal domain). It was trained on JRC-Acquis. We did very encouraging experiments, using the small See-Eera.Net training corpus, on Greek-English and English-Greek, Slovene-English and English-Slovene, Bulgarian-English and English-Bulgarian.
  - **RO-EBMT:** an example-based MT (legal domain) for Romanian-English and English-Romanian, trained on See-Era.Net. Very good BLEU score, yet lower than the score of STAR which used 15 times more training data.
- **NetSearch:** tools library for automatic extraction of web page properties (ip of the server host of the web page, text, internal and external links, etc.)
- **Web Crawler:** application traversing Internet websites from link to link and downloading the pages on the way. It can be modified for entire websites extraction or for automatic extraction of a parallel corpus (for this it has to use a multilingual documents alignment module). C# application (under development).
- **Racai Xml Web Services.** Web services set (using SOAP for communication, WSDL for describing the service and UDDI for registration). The services display some of the applications described before: TTL, MTagger, Sentence, Phrase and Word Aligner, WordNet Viewer, Language Identifier, STAR MT.