CroDeriV 2.0: Initial Experiments

Krešimir Šojat, Matea Srebačić, Tin Pavelić
ksojat@ffzg.hr, msrebaci@unizg.hr, tpavelic@ffzg.hr
Faculty of Humanities and Social Sciences, University of Zagreb

PolTAL 2014, Warsaw
2014-09-18

Introduction

- Croatian – rich inflectional and derivational processes
- inflectional processes – **Croatian Morphological Lexicon**
  - hml.ffzg.hr
- derivational processes – **CroDeriV**
  - newly developed resource for Croatian
  - so far only verbs
  - croderiv.ffzg.hr
- expansion of CroDeriV to nouns
  - detection of derivational relations between lexemes of various POS → rich derivational families

CroDeriV

- 14 326 Croatian verbs analyzed for morphemes
  - 3 386 lexical morphemes
- **lexical entry:**
  - lemma
  - morphological structure
    - 11 slots
    - p = prefix, r = root, i = interfix, s = sufix, () = optional, *-ti* = infinitive ending

Prefixes	Roots	Suffixes
(p4-p3-p2-p1)	**(r2-i)-r1**	**(s3)-s2-s1-*ti***
	pis-	*ø-a-ti*
‘to write’		
pot-	*-pis-*	*ø-a-ti*
‘to sign’		
is-pot-	*-pis-*	*-iv-a-ti*
‘to sign one by one’		

  - linguistic metadata (aspect, reflexivity, etc.)
  - several steps of morphological analysis

1. verbs collected from online corpora and dictionaries
2. rule-based splitter applied to verbal lemmas
3. obtained results manually checked
4. verbs sharing the same lexical morpheme – mutually linked

- problems of automated processing:
  1. **homography**
  2. **graphical overlapping of prefixes and suffixes with roots**
  3. **phonological changes at morpheme boundaries** – allomorphes
  4. **phonological changes within lexical morphemes**

- results:
  - **recognition of all allomorphs of a particular morpheme**
  - **detection of all affixes that co-ocurr with particular roots**
  - **detection of complete derivational families of verbs**

The derivation of Croatian nouns

- **three basic derivational processes**
  - **suffixation** – most productive process in noun derivation
    pis(ati) ‘write’ + *-ac* > *pisac* ‘writer’
  - **compounding**
    roman ‘novel’ + *-o-* + *pisac* ‘writer’ > *romanopisac* ‘novelist’
  - **prefixation**
    su- + *radnik* ‘worker’ > *suradnik* ‘co-worker’
  - two combined processes:
    - **compounding + suffixation**
      vatr(a) ‘fire’ + *-o-* + *gas(iti)* ‘extinguish’ + *-ac* > *vatrogasac* ‘firefighter’
    - **prefixation + suffixation**
      po- + *mor(e)* ‘sea’ + *-ac* > *pomorac* ‘sailor’
  - two non-concatenative processes
    - **back-formation**
      dopisati ‘to add by writing’ > *dopis* ‘letter’
    - **conversion**
      mlada ‘young, female, adjective’ > *mlada* ‘bride, noun’

Experiment

- expansion of CroDeriV to nouns
  - list of nominal lemmas: HML’s nominal part tagged as common nouns

= test sample: 20 554 nouns

Step 1

- aim: **to detect single nominal suffixes and obtain an initial snapshot of their producitivity**
- methodology: **set of rules for the detection and segmentation of single suffixes applied to the test sample**
- results: **4993 nouns with correctly recognized stems + 22 single suffixes**

Suffix	*-nje*	*-a*	*-ica*	*-telj*	*-na*
No. of occurences	2776	224	108	104	100

  - suffix *-nje* – used in derivation of gerunds from verbal stems
  - 25% of all common nouns in the HML – verbal nouns
    > all verbs from HML are in CroDeriV
    > slight modification of rules
    > morphological structure of all gerunds in HML automatically determined and manually validated

Step 2

- aim: **recognition and segmentation of nouns derived from verbs via back-formation**
  - **noun = verb without suffixal part**
  - **same morphological structure, same derivational family**
- methodology: **suffixal part removed from the verbs in CroDeriV and matched with remaining list of nouns in test sample**
- results: **3367 common nouns tagged as candidates > 2167 nouns were correctly segmented and assigned to the corresponding derivational family in CroDeriV**

Step 3

- aim: **(1) detection of possible suffixal combinations > input for rules capable of dealing with multiple suffixes**
  (2) detection of nominal stems not recognized in previous steps
- methodology: **40% of the remaining set of nouns was randomly chosen and manually analyzed for morphemes > nominal stems obtained and matched with the list of stems and roots from CroDeriV**
- results:
  - **matching recall: 100%, precision: 33,88%**
    (based on the manual evaluation of 5520 randomly selected nouns)

Overall results:

- **1773 / 3386 roots from CroDeriV (53,4%) have been correctly assigned to at least one noun from HML**
  - **automated expansion of derivational families**
- **1753 new nominal roots obtained through manual evaluation**
  - **can be used in further processing**
- **this simple automated approach assigned the correct root to 59,48% of the nouns from the test sample**
- **two noun sets are obtained:**

(1) set of nouns derivationally related to verbs in CroDeriv
> **enrichment of already existing derivational families**

(2) set of nouns that are not derivationally related to verbs in CroDeriV
> **formation of new derivational families**

The research within the project XLike leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement nº 288342.