

Implementation of Croatian NERC System

Božo Bekavac, bbekavac@ffzg.hr

Institute of Linguistics

University of Zagreb

How we started with Intex ?

- Goal/motivation: to create NERC system for processing Croatian texts
- compliant with *MUC Named Entity Task Definition*
- XML output
- Testing a lot of different tools
- Intex - best choice (even in competition with some “specialized” NERC tools)

Why Intex? – comparison with other “NERC tools”

- Simplicity

- Graph

- S

-

- U

- But, it

standard pro

described in previous conferences

```
Rule: Company1
```

```
Priority: 25
```

```
(
```

```
  ( {Token.orthography == upperInitial} )+ //from tokeniser
```

```
  {Lookup.kind == companyDesignator} //from gazetteer lists
```

```
):match
```

```
-->
```

```
:match.NamedEntity = { kind=company, rule="Company1" }
```

Strategies and resources used in our work

- *Internal and External Evidence* (McDonald 1996)
dr. Fran Mihaljević vs. **klinika** dr. Fran Mihaljević
(external evidence will override internal evidence)
- *Gazetteers* (and other lists of names)
- *Global word sequence checking* (Mikheev 1999)
Ivan, X i Marko *Srna, Olič i Pršo*
- *Filtering of false candidates* (Stevenson;
Gaizauskas 1999) *Oi Atena 2004*

Strategies NOT used in work

- **Dynamic lexicon** (Mikheev et al. 1998), (McDonald 1996), (Piskorski 2000)
- stores recognized NEs in secure context in order to detect NEs in non-predictive contexts
- Such information are relevant only inside certain discourse which is processed
- eg. *Jagoda, Višnja, Dunja...*
Jagoda je omiljena mnogim ljudima.
(A lot of people love (S)strawberry.)

Dynamic lexicon (2)

- somewhere in discourse: dr. Jagodu Ivić
- stores all word forms of recognized NEs
- In example:

Jagoda je omiljena mnogim ljudima.

marks *Jagoda* as person!

- Same method with: *Sunce, Zagreb...*
(eng. General Mils, Washington...)

Dynamic lexicon (3)

- *One Sense per Discourse* (Gale, Church, Yarowsky 1992): *it is extremely likely that polysemous words will share the same sense (98 %)*
- generates/recognizes all combinations of tokens from left to right (and acronymes)
- *Privredna banka Zagreb* → (*Privredna banka, Privrednoj banci, ..., banka Zagreb, ..., PBZ*)

Dynamic lexicon (4)

- help in resolving names with conjunction
- *odvjetnička tvrtka Vončina i Šavorić*

VS:

Pliva i Lura potpisali su ugovor o suradnji.

- checking in discourse for secure evidence:
...tvrtka Pliva ili Pliva d.d.
- (*China International Trust and Investment Corp*)

Graphs (1)

The screenshot displays the Intex software interface. The main window shows a text editor with the following content:

Text: C:\Intex\Croatian\Corpus\tekst.txt
Text units are lines/paragraphs.
39 lines/paragraphs, 986 (320 diff) tokens: 626 (300) simple forms + 0 (0) tags + 148 (10) digits + 21
276 simple words, 24 unknown tokens.

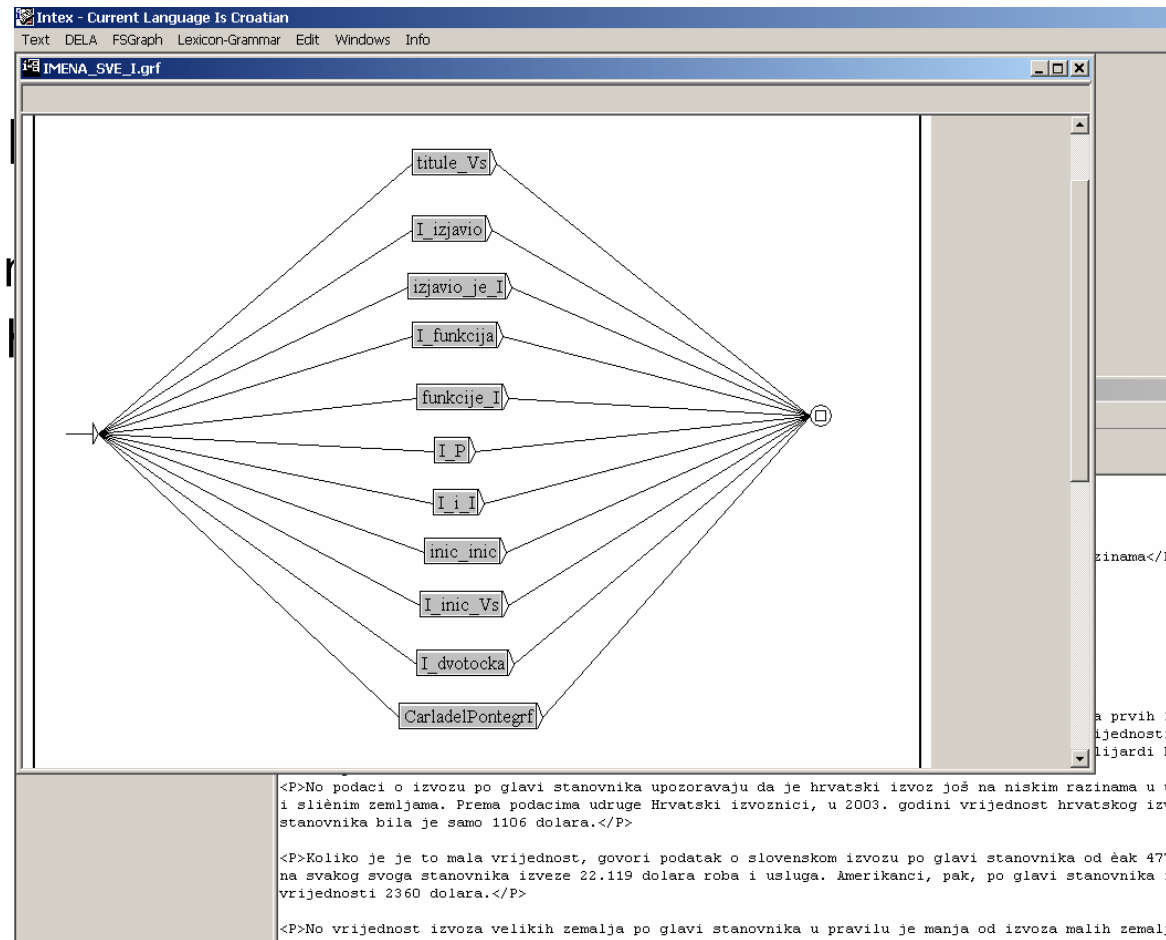
<P>Hrvatski izvoz napokon je prošle godine počeo rasti brže godine, izvoz u kunama rastao 15,7 posto a uvoz 5,7 posto. 44 milijardi kuna ili 7,25 milijardi američkih dolara, dok 15 milijardi dolara.</P>
<P>No podaci o izvozu po glavi stanovnika upozoravaju da je u sličnim zemljama. Prema podacima udruge Hrvatski izvoznici stanovnika bila je samo 1106 dolara.</P>

The graph window, titled 'vanj_dokaz.grf', shows a network of nodes and edges. The nodes are labeled with various verb forms, such as <ustvrditi>, <izjaviti>, <reći>, <kazati>, <potvrditi>, <istaknuti>, <dodati>, <zaključiti>, <odgovoriti>, <naglasiti>, <pojasniti>, <objasniti>, <poručiti>, <govoriti>, <priopćiti>, <istaknuti>, <dodati>, <uvjeriti>, <objaviti>, <ocijeniti>, <najaviti>, <reagirati>, <upozoriti>, <savjetovati>, <podsjetiti>, <uručiti>, <roditi>, <napomenuti>, <napomenula>, <pogledati>, <smatrati>, <spominjati>, <otkriti>, <imenovati>, <položiti>, <vratiti>, <uputiti>, <ponuditi>, <optužiti>, <upravljati>, <pozdraviti>, <odgovoriti>, <predati>, <spomenuti>, <vidjeti>, <nazvati>, <obrazložiti>, <pitati>, <pobjeći>, <objesiti>, <preminuti>, <potvrditi>, <primiti>, <ubiti>, <željeti>, <angažirati>, <doznati>, <izabrati>, <pitati>, <izreći>, <odbiti>, <pisati>, <pjevati>, <odgovoriti>, <potpisati>, <pozdraviti>, <predložiti>, <preći>, <razgovarati>, <uhitati>, <razriješiti>, <zahvaliti>, <obznaniti>, <čekati>, <najaviti>, <predati>, <očistovati>, <dodavati>.

The graph shows a sequence of nodes: [O] -> [I] -> [PRE] -> [MOT] -> vanj_dokaz. The nodes are connected by directed edges, representing the flow of information or the structure of the text.

Graphs (2)

- Will fill
- o suradn
- i da rast l



upozorio

Basic resources created for Croatian

- Sentence segmenter
- DELAF dictionary for Croatian
- Gazetteers
- Lists of names
- Grammars for numbers, dates...
- Of course, NERC system

- PRESENTATION

Performance of system

- F-measure of Croatian NERC system on informative domain **0.9**
- Errors like:
...mjesto predsjednika Vijeća za građanski nadzor sigurnosnih službi **kada** se ono...
LEMMA: kada (N), kada (Adv)
- non-informative domain: precision **0.79**, but recall 0.47, F-measure 0,59 → comparable with (Poibeau; Kosseim 2001)

Conclusion

- Intex/NooJ already now satisfies requirements for NERC system of highly inflective language
- Implementation of dynamic lexicon would make it even more suitable

Future work - perspectives

- migration of created resources to NooJ
- developing a new resources for NooJ
- Joint project with French side (Université de Franche-Comté):

Application of Local Grammar to Croatian texts with NooJ (EGIDE programme)

more people (two departments) starts to use NooJ in Croatia !!!