Tools for annotation and exploitation of parallel corpora: the case of the IJS-ELAN Slovene-English corpus

Tomaž Erjavec
Department of Intelligent Systems
Institute "Jožef Stefan"
Ljubljana, Slovenia

Talk given at
Centre for Corpus Linguistics
University of Birmingham
February 9th 2001

http://nl.ijs.si/et/talks/bham01/

Overview:

1. 'External tools'
2. The IJS-ELAN corpus Version 1
3. Word-class syntactic tagging
4. Extracting translation equivalents
5. Conclusions

1 External tools

Guidelines:

- open sources: GNU/Unix
- open standards: ISO/W3C
- availability, reusability, extensibility

IJS-ELAN corpus processing software:

- tokeniser & segmenter: MULTEXT(-East)
  capitalisation, abbreviations, compounds, numerals
- aligner: vanilla & Atril
- filters: Perl & Omnimark
- PoS tagger: TnT
  tagset, lexicon, rules
- morphological analysis: CLOG
  tagset, lexicon, rules
- translation extraction: Cognate, 21, Plug
  annotated parallel corpus, support resources

The SW collection as an exercise in testing a range of corpus processing software, thus furthering our mission in TELRI.

2 The IJS-ELAN corpus

The IJS-ELAN corpus:

- 1 million words
- bi-lingual: Slovene-English & English-Slovene
- sentence aligned
- tokenised
- encoded in accordance with TEI
- available on the Web

Tomaž Erjavec. The ELAN Slovene-English Aligned Corpus.
Proceedings of the Machine Translation Summit VII,
pages 349–357, Singapore, 1999.

Making the Corpus:

1. texts acquired, converted, segmented & aligned
2. tokenisation with `mtseg` & (most) errors corrected with Perl
3. conversion to standard format, addition of header information
4. packaging, conversions for WWW, concordancing

3 Composition of the corpus

15 component texts

Title	kB	kW
Constitution of the Republic of Slovenia	364	20
Speeches by the President of Slovenia, M. Kučan	1102	69
Functioning of the National Assembly	325	20
Slovenian Economic Mirror; 13 issues, 98/99	4056	239
National Environmental Protection Programme	1222	70
Europe Agreement	589	34
Europe Agreement - Annex II	483	25
Strategy for Integration into EU	1511	89
Programme for accession to EU - agriculture	543	29
Programme for accession to EU - economy	394	23
Vademecum by Lek	471	24
EC Council Regulation No 3290/94 - agriculture	1182	69
Linux Installation and Getting Started	3044	173
GNU PO localisation	353	13
G. Orwell: Nineteen Eighty-Four	6698	195

The team:

- Tomaž Erjavec: Orwell, Linux PO files, www.gov.si
- Roman Maurer: Linux book
- Andrej Skubic: Economic Mirror, www.gov.si
- Špela Vintar: EU texts

4 Corpus encoding

Corpus structure

1. Corpus = corpus header, corpus element+
2. corpus element = text header, text
3. text (body) = translation unit+
4. translation unit = segment, segment

SGML conformity

1. SGML declaration (7bit)
2. SGML DTD: a parametrisation of TEI

A TEI conformant corpus DTD

```
<!DOCTYPE tei.2 PUBLIC "-//TEI P3//DTD Main Document Type//EN" [
  <!ENTITY % TEI.prose    'INCLUDE'> <!-- base tag set -->
  <!ENTITY % TEI.analysis 'INCLUDE'> <!-- add: basic linguistic analysis -->
  <!ENTITY % TEI.linking  'INCLUDE'> <!-- add: pointer mechanisms -->
  <!-- add: local extensions -->
  <!ENTITY % TEI.extensions.ent SYSTEM "teitmx.ent">
  <!ENTITY % TEI.extensions.dtd SYSTEM "teitmx.dtd">
]>

teitmx.ent:
<!ENTITY % body 'IGNORE' >

teitmx.dtd:
<!ELEMENT %n.body;       - -  (tu+)>
<!ELEMENT tu             - -  (seg, seg)>
<!ATTLIST tu                  %a.global;>
```

This instantiation is encapsulated in the one file `ijs-elan.dtd`

5 Examples of translation units

```
<tu lang="sl-en" id="usta.301">
<seg lang="sl"><w type=dig>70.</w> <w>&ccaron;
<seg lang="en"><w>Article</w> <w type=dig>70</
</tu>
...
<tu lang="sl-en" id="spor.301">
<seg lang="sl"><w>ii</w><c>)</c> <w>za</w> <w>
<seg lang="en"><c type=open>(</c><w>ii</w><c t
</tu>
...
<tu lang="sl-en" id="kmet.301">
<seg lang="sl"><c>-</c> <w>razvoj</w> <w>pode&
<seg lang="en"><c>-</c> <w>Pillar</w> <w>IV</w
</tu>
...
<tu lang="sl-en" id="vade.301">
<seg lang="sl"><w>Na</w> <w>bole&ccaron;e</w>
<seg lang="en"><w>Apply</w> <w>a</w> <w>thin</
</tu>
...
<tu lang="en-sl" id="ligs.301">
<seg lang="en"><w>Many</w> <w>text</w> <w>proc
<seg lang="sl"><w>Za</w> <w>Linux</w> <w>je</w
</tu>
...
<tu lang="en-sl" id="gnpo.301">
<seg lang="en"><w>Usage</w><c>:</c> <w>%s</w>
<seg lang="sl"><w>Uporaba</w><c>:</c> <w>%s</w
</tu>
```

6 Plans for Version 2

Text:

- že omenjeni zavod za živinorejo, ki bo vsaj v prehodnem obdobju nadaljeval naloge na področju zootehnike;
- the above-mentioned Livestock-Breeding Centre, which will continue the tasks in zootechnics, at least for the transitional period;

Result:

```
<seg id="kmet.sl.87" corresp="kmet.en.87">
<c>-</c>
<w msd="Q" lemma="že">že</w>
<w msd="Vmp--pmp" lemma="omeniti">omenjeni</w>
<w msd="Ncmsa--n" lemma="zavod">zavod</w>
<w msd="Spsa" lemma="za">za</w>
<w msd="Ncfsa" lemma="živinoreja">živinorejo</w>
<c>,</c>
<w msd="Css" lemma="ki">ki</w>
<w msd="Vcif3s" lemma="biti">bo</w>
<w msd="Q" lemma="vsaj">vsaj</w>
<w msd="Spsl" lemma="v">v</w>
<w msd="Afpnsl" lemma="prehoden">prehodnem</w>
<w msd="Ncnsl" lemma="obdobje">obdobju</w>
<w msd="Vmps-sma" lemma="nadaljevati">nadaljeval</w>
<w msd="Ncfpa" lemma="naloga">naloge</w>
<w msd="Spsl" lemma="na">na</w>
<w msd="Ncnsl" lemma="področje">področju</w>
<w msd="Ncfsg">zootehnike</w>
<c>;</c>
</seg>
```

7 Experiments in tagging of Slovene

Sašo Džeroski and Tomaž Erjavec and Jakub Zavrel. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. LREC 2000, pages 1099–1104.

Slovene MULTEXT-East dataset:

PoS	Att	Val	1984	Lexicon
Pronoun	11	36	594	1,335
Adjective	7	22	169	279
Numeral	7	23	80	226
Verb	8	26	93	128
Noun	5	16	74	99
Preposition	3	8	6	6
Adverb	2	4	3	3
Conjunction	2	4	2	3
Interjection	0	0	1	1
Abbreviation	0	0	1	1
Particle	0	0	1	1
Σ	45	139	1,025	2,083
Punctuation	1	10	10	-

Table 1: Slovene morphosyntactic distribution

	Full	Train	Test
Sentences	5855	5204	651
Tokens	92399	81805	10594
Words	77772	68825	8947
Ambigs	87.2%	86.4%	70.2%
Diff pairs	18649	17166	3912
Diff words	16017	14831	3573
Diff MSDs	1004	976	543

Table 2: Corpus dataset

8 Evaluation

Token type	Tokens	RBT	MET	MBT	TnT
All	10594	85.95	86.36	86.42	89.22
on PoS	10594	95.64	94.66	95.31	96.59
Known	9049	92.88	91.56	93.58	95.08
on PoS	9049	98.75	97.02	98.76	98.51
on Type	8713	98.67	96.94	98.82	98.71
on Case	3557	87.74	88.16	88.89	93.06
on Number	4629	97.19	96.28	97.43	98.33
on Gender	4556	95.90	93.99	96.62	97.65
Unknown	1545	45.37	55.92	44.47	54.88
on PoS	1545	77.41	80.84	75.08	85.30

Table 3: Tagging accuracies

PoS	All tokens		Known		Unknown	
	n	%	n	%	n	%
Σ	10594	89.2	9049	95.0	1545	54.8
X	1647	100.0	1647	100.0	0	-
V	2454	95.8	2044	99.0	410	79.7
N	1901	81.4	1356	92.9	545	53.0
P	1062	79.0	1014	82.7	48	0.0
C	828	96.4	828	96.4	0	-
S	811	96.1	807	96.6	4	0.0
A	757	61.6	316	90.8	441	40.8
R	696	93.9	629	96.3	67	71.6
Q	336	88.6	332	89.7	4	0.0
M	98	65.3	72	83.3	26	15.3

Table 4: Tagging Accuracies by PoS, TnT

9 Tagset reduction by attribute removal

Tagset	Cardinality	MBT Accuracy
PoS Only	12	96.07
Type Only	38	95.57
All but Case	392	89.67
All but Gend	582	88.22
All but Numb	602	86.94
All but Type	665	87.27
Full MSDs	1021	86.93

Table 5: MBT accuracies on reduced tagsets

10 The TnT tagger

Qualities:

- availability
- accuracy
- robustness (large tagset!)
- speed

'Trigrams 'n Tags':

- Author: Thorsten Brants, Universität des Saarlandes
- Avalilable via WWW
- research license
- Solaris, Linux
- fast, robust, well-designed
- unknown word guessing

11 TnT parameters

```
%% Statistically tagged file, Sun Dec  5 16:32:44 1999
%% lexicon     : mte.lex
%% ngrams      : mte.123
%% corpus      : elan-sl.t
%% model       : trigrams
%% sparse data : linear interpolation
%%  lambda1 = 1.292668e-01
%%  lambda2 = 3.310223e-01
%%  lambda3 = 5.397110e-01
%% unknown mode: lexicon entry @UNKNOWN
%% case of characters is significant
%% using suffix trie up to length 10
%% unknown words are marked with an asterisk (*)
%% Thorsten Brants, thorsten@coli.uni-sb.de

%% 177776 (30.04%) unknown tokens
%%      7141 recognized as cardinals/ordinals
%% 102761 tokens taken from the backup lexicon
%% avg. 10.87 tags/token, 1.71 tags/known token

izhajajoč               Afpmsnn *
iz                      Spsg
Temeljne                Afpfsg  *
ustavne                 Afpfsg  *
listine                 Ncfsg   *
o                       Spsl
samostojnosti           Ncfsl   *
```

12 Lemmatising unknown words

- Lemmatisation (produce headword):
  infusions ⇒ *infusion, infuzije* ⇒ *infuzija*
- Stemming (strip inflectional suffix):
  infusions ⇒ *infusion, infuzije* ⇒ **infuzij*
- Radical stemming (retain root):
  infusions ⇒ *infus, infuzijski* ⇒ *infuz*

An approach using Inductive Logic Programming:
Sašo Džeroski, Tomaž Erjavec:
Learning to lemmatise Slovene words by learning morphological analysis and POS tagging.
J. Cussens and S. Džeroski, editors. Learning Language in Logic. Springer, Berlin, 2000.
Lecture Notes in Artificial Intelligence, 1925.

Algorithm:

1. Tag the text
2. Given the word-form and the tag, derive the lemma

Dataset:

- Training: body of the novel '1984'
- Testing: Appendix of '1984' (The principles of Newspeak)
- Open domain: Text on EU

13 1st step: tagging

Tagging with TnT:

- MULTEXT-East tagset
- Training on Slovene Orwell corpus
- Use TnT own unknown word guessing module (suffix trie from known hapax words)

Table 6: Validation results for the TnT tagger.

	Accuracy	Correct/Err
All tokens	83.7%	4065/789
All words	82.5%	3260/692
Known words	84.3%	3032/565
Unknown words	64.2%	228/127

Table 7: Validation results for the TnT tagger on nouns and adjectives.

	All	Known	Unknown
Nouns	73.8%	77.5%	58.3%
Adjectives	62.3%	60.7%	68.4%
Both	70.1%	72.2%	61.6%

14 2nd step: morphological analysis

Lemmatising with CLOG:

- MULTEXT-East tagset
- Training on Slovene Orwell inflectional lexicon
- Focus experiment on unknown Nouns and Adjectives

To learn rules for morphological analysis (lemmatisation) we used first-order decision list learning systems, that can train from positive examples only.

Suresh Manandhar, Sašo Džeroski, and Tomaž Erjavec.
Learning multilingual morphology with CLOG.
In David Page, editor, *Inductive Logic Programming; 8th International Workshop ILP-98, Proceedings*
Number 1446 in Lecture Notes in Artificial Intelligence, pages 135–144. Springer, 1998.

15 Rules of analysis

Prolog facts

- FOIDL for English Verb past:

```
past([b,a,r,k],[b,a,r,k,e,d]).
past([g,o],[w,e,n,t]).
```

- CLOG for Slovene Noun feminine singular genitive:

```
n0fsg([s,e,t,v,e],[s,e,t,e,v]).
n0fsg([b,o,l,e,z,n,i],[b,o,l,e,z,e,n]).
n0fsg([p,e,r,u,t,i],[p,e,r,u,t]).
n0fsg([m,i,z,e],[m,i,z,a]).
```

Learned rules

- FOIDL for English Verb past:

```
past([g,o],[w,e,n,t]) :- !.
past(A,B) :- split(A,C,[e,p]), split(B,C,[p,t]), !.
past(A,B) :- split(B,A,[d]),   split(A,C,[e]), !.
past(A,B) :- split(B,A,[e,d]).
```

- CLOG for Slovene Noun feminine singular genitive:

```
n0fsg(A,B):-mate(A,B,[],[],[t,v,e],[t,e,v]),!.
n0fsg(A,B):-mate(A,B,[],[],[e,z,n,i],[e,z,e,n]),!.
n0fsg(A,B):-mate(A,B,[],[],[i],[]),!.
n0fsg(A,B):-mate(A,B,[],[],[e],[a]),!.
```

16 Analysis evaluation

Table 8: Morphological analyser results on the validation set.

	All		Known		Unknown	
	Acc.	Correct/Err	Acc.	Correct/Err	Acc.	Correct/Err
hline Nouns	97.5%	936/24	99.1%	766/ 7	90.9%	170/17
Adjectives	97.3%	431/12	96.6%	339/12	100%	92/0
Both	97.4%	1367/36	98.3%	1105/19	93.9%	262/17

Table 9: Lemmatisation results on the validation set.

	All		Known		Unknown	
	Acc.	Correct/Err	Acc.	Correct/Err	Acc.	Correct/Err
Nouns	91.7%	880/ 80	95.4%	738/ 35	75.9%	142/ 45
Adjectives	87.6%	388/ 55	88.0%	309/ 42	85.9%	79/ 13
Both	90.4%	1268/135	93.1%	1047/ 77	79.2%	221/ 58

Table 10: Lemmatisation results on the open domain set.

	Token	Type	Lemma
Accuracy	81.3%	79.8%	75.6%
All	1322	796	595
Correct	1075	635	450
Error	247	161	145
Wrong	195	105	73
Mixed	-	38	62
Fail	52	18	10

17 Translation equivalents

In search of programs that:

- take as input a sentence aligned parallel corpus
- produce (multiword) translation equivalents
- are available, free, precise and robust
- work under Unix and don't depend on a GUI...

Three examples: Cognate; 21; Plug

Such programs could then be used in a tool chain or configuration (possibly only semi-automatic) that would enable rapid creation of high-quality bilingual (terminological) lexica.

An experiment on the IJS-ELAN corpus is described in:

Špela Vintar (1999). A Lexical Analysis of the ELAN Slovene-English Parallel Corpus.
Proceedings of the Workshop Language Technologies: Multilingual Aspects, within the framework of the SLE conference, July '99, Ljubljana: Filozofska fakulteta.

18 Cognates

Cognates are the simplest (but often quite productive) translation equivalents.

Perl module `String::Approx`:

- Available from CPAN
- Computes edit distance beteen two strings: ops are add, delete, substitute
- Implemented in C using the so-called Manber-Wu k-differences algorithm shift-add.
- GNU Library General Public License

19 Example

Špela Vintar: finding congnates in a bi-text:
in a translation unit return all pairs (w_{source}, w_{target})
that have edit distance below given treshold

An example, sorted by frequency of occurence:

```
     21 informatizacije informatization
...
      7 upravi          administration
      7 avtomatizacije  automatization
      6 procesov        processes
      6 organizacij     organizations
      6 avtomatizacijo  automatization
      5 uprave          have
      5 organizacijske  organizational
      5 organizacije    organization
...
      1 zagati          administration
      1 zadnje          made
      1 zadeva          demanding
      1 začeti          organizational
```

20 Twente word alignment software

The 21 program constructs a bilingual lexicon from a parallel sentence aligned corpus. The translations are ranked according to computed confidence. Uses various statistical measures, and is based on a symmetric translation model. Works for single words (tokens) only.

- Downloadable from the Web
- GNU General Public License
- C source code

21 Example translations

```
Total corpus 1:          Total corpus 2:
 9367 words               13244 words
 6750 words used          8997 words used
 1858 different words     1402 different words

Total:
 815 sentences
 737 sentences used

Results of Model A Iterative Proportional Fitting algorithm

Dictionary 1:

šest              širše
----------------  ------------------
non-profit  0.33  wider          0.86
6           0.33  self-governin 0.14
making      0.33


državama        države        državi
--------------  ------------  --------------
Italian   0.50  state   0.60  state     0.37
Hungarian 0.50  defence 0.34  highest   0.25
                monies  0.06  country   0.14
                              foreign   0.14
                              be        0.10
```

22 PLUG Word Aligner

PWA comprises two word alignment systems, the Linköping Word Aligner (LWA) and the Uppsala Word Aligner (UWA). Both were developed in the PLUG project (1997-2000). PWA integrates both systems in the modular corpus toolbox Uplug and includes tools for the automatic generation of monolingual word collocations (phrases) and for the automated evaluation of alignment results (the PLUG Scorer - PLS).

- Author: Jörg Tiedemann, Department of Linguistics, Uppsala University
- Research only license agreement
- Binaries for Linux, MS Windows, SunOS
- Perl, Tcl/Tk

23 Example translations

- Teh nekaj dni je v temelju spremenilo naš navidezno mali svet, spremenilo je Jugoslavijo.
- These few days have fundamentally changed our seemingly small world, and have changed Yugoslavia.

```
# columns: (id,source,target,align step,score)
# created_by: UWA
# date_created: Fri Jan 12 17:52:05 2001
kuca.4  spremenilo      changed         1       1.66
kuca.4  Jugoslavijo     Yugoslavia      2       0.727
kuca.4  svet            world           6       0.63
kuca.4  Teh             These           6       0.44
kuca.4  mali            small           7       0.559
```

24 Conclusions

- free tools are becoming available
- preliminary results seem promising
- supervised learning requires substantial resources

25 Link Index

- TELRI, IJS site: http://nl.ijs.si/et/telri/
- MULTEXT-East: http://nl.ijs.si/ME/
- IJS-ELAN corpus: http://nl.ijs.si/elan/
- Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets: http://nl.ijs.si/et/Bib/LREC00/
- A Lexical Analysis of the ELAN Slovene-English Parallel Corpus: http://www2.arnes.si/∼svinta/spela-en.htm
- The TnT tagger: http://coli.uni-sb.de/∼thorsten/tnt/
- The 21 aligner: http://parlevink.cs.utwente.nl/Projects/twentyone.html
- The PWA aligner: http://stp.ling.uu.se/∼corpora/plug/pwa/