The ELAN Slovene-English Aligned Corpus

Tomaž Erjavec
Department for Intelligent Systems
Jožef Stefan Institute
Jamova 39
SI-1000 Ljubljana
Slovenia

Abstract

Multilingual parallel corpora are a basic resource for research and development of MT. Such corpora are still scarce, especially for lower-diffusion languages. The paper presents a sentence-aligned tokenised Slovene-English corpus, developed in the scope of the EU ELAN project. The corpus contains 1 million words from fifteen recent terminology-rich texts and is encoded according to the Guidelines for Text Encoding and Interchange (TEI). Our document type definition is a parametrisation of the TEI which directly encodes translation units of the bi-texts, in a manner similar to that of translation memories. The corpus is aimed as a widely-distributable dataset for language engineering and for translation and terminology studies. The paper describes the compilation of the corpus, its composition, encoding and availability. We highlight the corpus acquisition and distribution bottlenecks and present our solutions. These have to do with the workflow in the project, and, not unrelatedly, with the encoding scheme for the corpus.

1 Introduction

While translation systems are being developed intensely for major languages, there has been much less effort to secure MT or CAT tools for languages with a smaller number of speakers. Slovene is a South-Slavic language with cca. 2 million speakers and is spoken predominately in Slovenia. For the development as well as assessment of MT technology applied to a new language, it is extremely useful to have widely available and reusable annotated corpora of the language in question. For MT research the most valuable type of corpus is one consisting of texts and their translations, i.e., a parallel bilingual corpus. Further utility is obtained if the corpus is aligned at least at the sentence level, and the texts tokenised and part-of-speech annotated.

The vaule of aligned corpora is well attested in practice, with several recent European projects devoted to producing them, e.g., MLCC (Armstrong et al., 1998) (nine EU languages), Crater (McEnery et al., 1997) (Spanish, French and English), or ENPC (Johansson et al., 1996) (English-Norwegian). Such corpora have also been produced for non-European languages, e.g., the HKUST Chinese-English corpus (Wu and Xia, 1995).

For the Slovene language, the only available parallel corpus has so far been the one released on the TELRI CD-ROM (Erjavec et al., 1998), which comprises Plato's Republic and the MULTEXT-East corpus (Erjavec and Ide, 1998). The parallel part of the MULTEXT-East corpus consists of the novel '1984' by George Orwell in English and translations. The MULTEXT-East corpus derives most its value from the fact that it contains parallel texts in many languages, and is heavily annotated: the markup includes gross document structure, sentence and sub-sentence markup (names, quotes. ...), disambiguated lemmas and morphosyntactic descriptions of its words, and alignment of sentences with the ones from the English original. To facilitate the reusability of the corpus, it is annotated in accordance with international recommendations for written text corpora targeted towards language engineering research, in particular the Corpus Encoding Specification, CES (Ide, 1998). While the encoding of MULTEXT-East corpus is such that it is suitable for further processing, annotation and exploitation, this parallel English-Slovene corpus nevertheless consists of only one novel.

The European Language Activity Network (EU MLIS project ELAN) provided an opportunity to somewhat remedy this lack. Our contribution to ELAN was, in part, to collect and annotate a 1 million word Slovene-English / English-Slovene corpus and make it widely available as a standardised dataset for bilingual language research on the Slovene language. The IJS-ELAN corpus contains fifteen recent texts, from interesting areas of text production. The texts and corpus encoding have been chosen so as to have minimal restriction on further use, and could thus be made widely available as a standardised dataset for bilingual language engineering research.

The article is structured as follows: Section 2 reports on the corpus compilation project and the processing issues involved. Section 3 presents the 15 component texts of the corpus. Section 4 turns to our parametrisation of the TEI and the markup used in the corpus. Section 5 discusses the availability and distribution of the corpus and Section 6 gives conclusions and direction for further work.

2 Acquisition and Processing

The project operated under tight time and labour constraints, so it was imperative to maximise the results by minimising the most costly steps in the production process. These include obtaining permission of the copyright holders to use the texts for the purposes of the project, obtaining the digital originals of the texts themselves, converting, segmenting and aligning the texts / translations, tokenising the texts and packaging them in a standard format and writing the text and corpus headers. The two decisive factors in stream-lining the cost were the adopted work-flow in the project and the encoding of the corpus.

A large amount of labour can be required for converting the original documents to a format understood by tools to process the text, in particular those that segment, align and tokenise the text. Such tools are, at least for academic projects, usually developed inhouse, an endeavour taking substantial time. Also, these processes are always noisy (incorrect segmentation and alignment) and require a laborious post-editing phase. Here, tools that offer a good visual validation and correction environment are welcome.

Recently, tools for translators, and esp. translation memory software have become successful commercial products. Translation memory software stores aligned segments, usually sentences, of previous bi-texts. When presented with a new original it compares its sentences with those stored in the translation memory and offers their translations for (edited) inclusion to the translator. Translation memories can be produced semi-automatically via an interactive process of segmentation and alignment. They thus closely resemble classical aligned corpora.

Rather then acquiring the texts and converting, segmenting, aligning and, especially, hand-validating locally, these tasks were, for the majority of the texts, performed by external collaborators of the project. The software used was mostly Déjà Vu, a commercial translation memory program, which offers an interactive alignment environment. The output of this process gives texts which are, to a large extent, stripped of original markup and presented in a simple tabular format, one translation unit per line. We thus obtained aligned bi-texts, which are, however, missing all structural information above the translation unit segments, i.e., sentences.

The bi-texts were then cleaned up with Perl filters (character set normalisation, removal of spurious formatting), and then tokenised into words and punctuation marks. This step was performed with the MULTEXT tool 'mtseg' (Cristo, 1996), with resources for English and Slovene developed in the MULTEXT-East project (Dimitrova et al., 1998). The tokenisation also flags numerals, compounds, abbreviations, etc. Again, this step introduced errors, which were, to a large extent, corrected with Perl filters. The tokenised aligned texts were converted into a TEI conformant encoding; here the header information is added to the bi-texts, and the alignments are encoded as SGML/TEI elements. The last step involved packaging the corpus distribution.

3 Corpus Composition

The small scale of the project prohibited any attempt towards making an English-Slovene reference-type corpus except maybe at the level of encoding. The composition of the Elan Slovene-English corpus was motivated in part by considerations of usability, and in part by ease of acquisition. For usability, the corpus contains recent (90's) texts rich in terminology and from active topic areas. Ease of acquisition also played a decisive role in choosing the particular texts; we only considered texts where the original and translation were already available electronically, in one of a few formats: HTML, RTF, and SGML (QUERTZ DTD). A factor in selecting the component texts was the willingness of the copyright holders to allow the inclusion of their texts in the corpus, with minimal restrictions on further distribution. Finaly, having collaborators in the project who chose the kinds of texts they were themselves most interested in studying also gave a certain coherence to the corpus.

The corpus has fifteen components, which are mostly complete bi-texts, but with omissions of predominately non-textual data (numerical charts etc). In the corpus, each bi-text is given its ID and constitutes, along with its header, one element of the corpus.

The texts are usefully divided into those that have a Slovene original and an English translation, and those whose original is English, and the translation is into Slovene. Apart from there being linguistic differences due to the opposition original/translation, the two parts also have a quite different composition.

The Slovene - English half has been, for the most part, acquired from various branches of the Slovene government. It consists of eleven texts, containing somewhat more than half of the corpus material. The

Slovene-English texts, together with their IDs, approximate sizes in kilo-bytes and -words, and year of publication, are as follows:

usta 364 Kb, 20 kW, 1997
Constitution of the Republic of Slovenia
Ustava Republike Slovenije
Constitutional Court of the Republic of Slovenia

kuca 1102 Kb, 69 kW, 1990-95
Speeches by the President of Slovenia, M. Kučan
Govori predsednika RS, M. Kučana
The Office of the President of the Republic of Slovenia

parl 325 Kb, 20 kW, 1998
Functioning of the National Assembly
Delovanje Državnega zbora
The National Assembly of the Republic of Slovenia

ecmr 4056 Kb, 239 kW, 1998/1999
Slovenian Economic Mirror; 13 issues
Ekonomsko ogledalo; 13 številk
Institute of Macroeconomic Analysis and Development of the Republic of Slovenia

ekol 1222 Kb, 70 kW, 1999
National Environmental Protection Programme
Nacionalni program varstva okolja
Office of the Government of the Republic of Slovenia for European Affairs

spor 589 Kb, 34 kW, 1996
Europe Agreement
Evropski sporazum
Office of the Government of the Republic of Slovenia for European Affairs

anx2 483 Kb, 25 kW, 1996
Europe Agreement - Annex II
Evropski sporazum - Priloga II
Office of the Government of the Republic of Slovenia for European Affairs

stra 1511 Kb, 89 kW, 1997
Slovenia's Strategy for Integration into EU
Strategija Slovenije za vključevanje v EU
Office of the Government of the Republic of Slovenia for European Affairs

kmet 543 Kb, 29 kW
Slovenia's programme for accession to EU - agriculture
Državni program za prilagajanje zakonodaje - kmetijstvo
Office of the Government of the Republic of Slovenia for European Affairs

ekon 394 Kb, 23 kW
Slovenia's programme for accession to EU - economy
Državni program za prilagajanje zakonodaje - gospodarstvo
Office of the Government of the Republic of Slovenia for European Affairs

vade 471 Kb, 24 kW, 1995
Vademecum by Lek, 1995
Vademecum Lekove domače lekarne
Lek d.d.; OTC Division

The English-Slovene part of the corpus contains almost half of the corpus material, but is composed of only four elements, with two of these being full-length books. It also has different text types from the Slovene-English part: two components deal with computers, one with pharmaceuticals and one with a rather grim projection of the future, from the past:

vino 1182 Kb, 69 kW, 1994
EC Council Regulation No 3290/94 - agriculture
Uredba sveta ES št. 3290/94 - kmetijstvo
Office of the Government of the Republic of Slovenia for European Affairs

ligs 3044 Kb, 173 kW, 1999
Linux Installation and Getting Started
Namestitev in začetek dela z Linuxom
Linux Documentation Project: -en: Specialized Systems Consultants / -sl: Linux User Group of Slovenia, LUGOS

gnpo 353 Kb, 13 kW, 1999
GNU PO localisation files
GNU PO lokalizacije datoteke
Free Software Foundation, Linux Documentation Project

orwl 6698 Kb, 195 kW, 1948
G. Orwell: Nineteen Eighty-Four
G. Orwell: 1984
The Slovene translation of the book was published by Knjižnica Kondor, Mladinska knjiga in 1983 (translator: Alenka Puhar). The first digital versions of the English and Slovene (as well as the Serbian and Croat translations) were keyed in at the School of Oriental and African Studies at London University, then became a part of the Oxford Text Archive and were published, with minimal changes, on the ECI-I CDROM. This served as the basis of the marked-up MULTEXT-East version.

4 Corpus Encoding

The IJS-ELAN corpus uses an SGML Document Type Definition, which is a parametrisation of the Text Encoding Initiative Guidelines (TEI P3, (Sperberg-McQueen and Burnard, 1994; Erjavec, 1999). Although TEI makes explicit recommendations for encoding aligned parallel corpora these, however, did not seem suitable for IJS-ELAN. Instead, we encoded the corpus in a manner similar to Translation Memory Exchange, TMX (Melby, 1998). In this section we first present the recommended TEI and TMX formats and then introduce our own. We then explain the overall structure of the corpus, and then move on to the corpus headers and texts, especially the token level markup.

4.1 TEI and TMX

The TEI P3 book (Sperberg-McQueen and Burnard, 1994) discusses alignment of parallel texts in multilingual corpora in section 14.4.2 and offers four different methods of encoding. The first choice is whether the elements to be aligned are points or intervals and the second whether the alignment itself is encoded as cross references of the segmentation markup or in a free standing linkage element. The basic assumptions for TEI parallel corpora is that the integrity of the two (or more) text documents is retained, as alignment is encoded on the meta-level. A closely related view is held by the Corpus Encoding Standard, CES (Ide, 1998), a TEI-derived encoding specification for corpora targeted at language engineering. The CES takes the TEI stand-off markup one step further: the original documents, so called *primary data* are in the process of alignment left completely unmodified; all the alignment information, possibly with segmentation, is held in a separate SGML document. This document contains an optional header followed by the <linkList> element with pointers into the primary data. To illustrate:

```
<!DOCTYPE cesDoc PUBLIC "-//CES//DTD cesDoc//EN">
 <cesDoc lang=en version="4.3">
  <cesHeader type="text" lang="en" ...
...
   <text lang="en">
    <body id="Oen">
     <div id="Oen.1" type="part" n=1>
...
      <s id="Oen.1.1.5.6">It was even
conceivable that they  watched everybody all the
time.</s>
      <s id="Oen.1.1.5.7">But at any rate
they could plug in your wire whenever they wanted
to.</s>
...

<!DOCTYPE cesDoc PUBLIC "-//CES//DTD cesDoc//EN">
 <cesDoc lang=sl version="4.3">
  <cesHeader type="text" lang="en" ...
...
   <text lang="sl">
    <body id="Osl">
     <div id="Osl.1" type="part" n=1>
...
      <s id="Osl.1.2.6.6">Mogo&ccaron;e je bilo
celo, da vsakogar ves &ccaron;as opazujejo; bodi
tako ali druga&ccaron;e, priklju&ccaron;ili so se
lahko na tvoj oddajnik, kadarkoli so hoteli.</s>
```

```
...

<!DOCTYPE cesAlign PUBLIC
                   "-//CES//DTD cesAlign//EN">
 <cesAlign version="1.0">
  <linkList>
   <linkGrp targType="s">
...
    <link xtargets="Osl.1.2.6.6 ;
                    Oen.1.1.5.6 Oen.1.1.5.7">
...
   </linkGrp>
  </linkList>
 </cesAlign>
```

It should be noted that while the pointers above refer to IDs (elements) in the primary data, this need not necessarily be the case; both TEI and CES make recommendations for extended pointer mechanisms, which allow reference to arbitrary locations in the documents.

Both proposals, and especially CES, treat the alignment information as stand-off, i.e., the documents to be aligned still exist in their own right and can contain arbitrarily complex markup above the level which is being aligned. This view is gaining in popularity and is useful in a number of applications (Thompson and McKelvie, 1997). A hidden assumption here seems to be, that the primary data is first acquired and converted to TEI or CES encoding, possibly trying to make use of original markup. The texts are then aligned, and the alignment converted to stand-off markup into this primary data. In our case the situation was clearly different, as the translation memories did not preserve the document structure.

Translation memories take translation units directly as their primary 'corpus' elements; the original documents are no longer of any direct interest. The aligned segments are simply encoded inside such units, and if any additional information about the segments is need, e.g., when they were created, which subject area they belong to, etc., this is encoded in the translation unit as well, and not in the document header.

The proposal to standardise translation memories, called the Translation Memory Exchange, TMX (Melby, 1998), is being developed in the scope of LISA, the Localisation Industry Standards Association. While it is useful to compare TMX with TEI, the two are quite different and TMX is not directly usable for aligned corpus encoding for research and interchange. Rather than encoding 'for scholarly purposes', TMX is application oriented; it is heavily concerned with preserving the markup of the digital original (e.g., RTF), while TEI documents will most likely contain mostly descriptive markup; the markup of the original will either be converted to such markup or discarded. On a related point, TMX in implemented XML and mandates the use of UNICODE, while TEI uses the more conservative SGML and allows character set entities. Below is one of the few currently available examples of TMX translation units:

```
<tu tuid="0002" srclang="*all*">
  <prop type="Domain">Cooking</prop>
  <tuv lang="EN"><seg>menu</seg></tuv>
  <tuv lang="FR-CA"><seg>menu</seg></tuv>
  <tuv lang="FR-FR"><seg>menu</seg></tuv>
</tu>
```

In contrast to the TEI recommendations, the text from which the above segment came from does not exist anymore, at least not directly. Even if the translation unit segments of one language or both languages are in sequence, this does not mean that the 'original' (whatever this term in the context or (re)encoding means) could be recreated; there could be gaps, the structural markup of the original is absent (e.g., <div>) and the translation units could have been edited, where the edits only make sense in the context of the translation unit.

The consequences of adopting such an approach for encoding of a parallel aligned corpus are similar; the primary data become the translation units, the usage of which does not involve any (possibly complex) pointer resolution. The originals are, to a certain extent, lost.

4.2 The IJS-ELAN DTD

We use a simple instantiation of TEI, which keeps the benefits of the 'off the shelf' TEI encoding (header, sub-segment markup) but treats corpus texts as a direct collection of translation units.

This document type is very similar to the one used in PLUG: Parallel Corpora in Linköping, Uppsala, and Göteborg, (Ahrenberg et al., 1999; Tiedemann, 1998).The main difference between the approaches lies in the method of DTD construction: PLUG use their own XML DTD whereas we parametrise the TEI in conformance with the procedures outlined in Chapter 29 of TEI P3 (Sperberg-McQueen and Burnard, 1994, pp.737-744).

The TEI "Chicago Pizza Model" allows the construction of a particular TEI SGML DTD by a) choosing one base tagset, b) adding additional tagsets and c) defining local extensions. The following SGML prolog implements our DTD:

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

The two `teitmx` extension files are quite short. The entity extension file ignores the standard definition of the TEI <body>, while the DTD extension file redefines <body> to be composed of translation units only; each translation unit has two (TEI.ANALYSIS) segments and the standard global attributes, of which we use the identifier ID and language LANG:

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

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

If it is felt as necessary, the above <tu> definition could be expanded to contain more information about the translation unit in question: terms appearing in the translation unit could be extracted, and further annotated or the revision description of the translation unit could be included. The addition of such meta-information would make the encoding even more similar to standard translation memories.

As can be seen, the structure of the corpus bi-text is extremely simple. This makes it suitable for direct processing with limited tools or computer expertise. The above encoding also to a large extent enforces the condition that the usage of the corpus will be at most over its sentences (segments), as all super-segmental markup is lost and the texts would require a substantial amount of effort to recreate in their entirety.

To enable the TEI parametrisation to work with an SGML conformant system, the TEI distribution is also needed. Because many SGML tools have problems coping with the SGML complexity used in TEI, we have also made a one-file normalised DTD enforcing the same markup as the paramterisation. This DTD is used for local processing and is included in the corpus distribution. The DTD was produced automatically via the 'Pizza Chef' on-line service at
⟨*http://firth.natcorp.ox.ac.uk/TEI/pizza.html*⟩.

We have also implemented a strict version of the one-file DTD, which allows only the elements and nestings encountered in our corpus. It is thus much more prescriptive than the TEI-derived DTD and served as a validation aid.

4.3 Top level corpus structure

The corpus as a whole is a valid SGML document, and therefore contains the following components:

1. the SGML Declaration, which defines local processing options. It makes the usual (TEI) assumptions about capacity points but limits the character set to ASCII: all the language specific characters in the corpus, e.g., č, and Ć are encoded as SGML entities, e.g., `č` and `Ć`. The declaration also prohibits tag

minimisation, so the corpus encoding is XML-like.

2. the SGML DTD, the Document Type Definition, which defines the annotation grammar of the corpus. As explained above it is a parametrisation of TEI. It also defines all and only the SGML character entities that appear in the corpus, e.g., `&`, `<`, `č`, `ß`

3. the SGML Document itself, which contains the SGML Prolog, the corpus header and (SYSTEM entity references to) all the corpus components, i.e., headers and texts.

Each of the fifteen corpus elements is stored in two files, one containing the component header and the other the aligned bi-text. As we expect that many users will be interested only in parts of the corpus, a significant amount of information identical across texts is kept in the text headers, and not solely in the header of the corpus.

4.4 The headers

The corpus as a whole, as well as each component has its TEI header. This header contains detailed information about the file itself, the source of its text, its encoding, and revision history.

Because the corpus is bilingual, we tried our hand at extending the bilinguality into the headers. This is achieved by doubling header elements but distinguishing their localisation via the `lang` attribute. This step, however, is not completely satisfactory, as it leads to some use/mention conflicts: does a header element marked as `lang=sl` *contain* Slovene text, or is it *describing* text in Slovene?

To give an impression of the information encoded in the header, we give below some examples from the corpus headers. The first is the beginning of the corpus and corpus header:

```
<teiCorpus.2>
  <teiheader type="corpus" lang="slen"  id="ijs-e
   creator="et" status="update" date.created="199
   date.updated="1999-06-22"
  >
    <filedesc>
      <titlestmt>
        <title lang="en">The IJS-ELAN Slovene/Eng
        <title lang="sl">Slovenskoangle&scaron;ki
```

The tags declaration in the corpus header:

```
<tagsdecl>
 <tagusage gi=text occurs=15>Element 'Text'. Att
 <tagusage gi=body occurs=15>Element 'Body'. Con
 <tagusage gi=tu occurs=31900>Element 'Translati
 <tagusage gi=seg occurs=63800>Element 'Translat
 <tagusage gi=s occurs=13386>Element 'Sentence'.
 <tagusage gi=w occurs=1091745>Element 'Word'. A
 <tagusage gi=c occurs=167243>Element 'Punctuati
</tagsdecl>
```

Part of the responsibility statement from a text header:

```
  <respstmt>
   <name>Jasna Belc, SVEZ</name>
   <resp lang="sl">Zagotovitev digitalnega origin
   <resp lang="en">Provision of digital original<
   <name>&Scaron;pela Vintar, FF</name>
   <resp lang="sl">Poravnava</resp>
   <resp lang="en">Alignment</resp>
```

The bibliography of a source texts in a text header:

```
<bibl lang="en" default="yes">
  <title lang="en">Linux Installation and Getting
  <xref type="URL">http://metalab.unc.edu/LDP/LDP
  <xref type="URL">ftp://metalab.unc.edu/pub/Linu
  <publisher>Specialized Systems Consultants
    <xref type="URL">http://www.ssc.com/</xref>
  </publisher>
</bibl>
```

4.5 The texts

Each text (<body> element, to be precise) is composed of translation units, <tu> elements, each having two segments: the original and translation. The definition of the segment element is taken directly from the TEI.ANALYSIS module, and allows significant subsegment-level markup. Our corpus currently encodes word and punctuation elements, i.e., it is tokenised. Below we give some translation units from the corpus:

```
<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>
```

The token level markup is, of course, not meant for reading, but to facilitate software to exploit the corpus further, rather than having to do tokenisation itself, usually a first step before further processing.

We have assigned some possibly useful values to the TYPE attribute of the token elements. As the programs that assigned these types, as well as the corpus texts contain errors, so do the types of some tokens. The token elements have the following values of `type`, with some examples taken from the corpus:

<w type=comp> Compound (lexical multiword unit), e.g., *medtem ko, vice versa, New York*

<w type=dig> Digit (numeric expression), e.g., *1984, 3., IV, 20%, 1993-1996, 25/76, 16MB*

<w type=abbr> Abbreviation (ending in a period), e.g., *tar., et al., S.u.S.E., dipl.*

<w> implied type for 'normal' words, e.g., *Slovenije, market, 's, Article, živinorejo, INAVGURACIJSKI, Hurt-Andreatta, Hrup51, E-poštni, D'you*

The punctuation element, <c> can be marked with `type` `open` or `close`, e.g., <c type=open>[</c>. This type is interesting for quotes: the quotes themselves have been normalised to the 'directionless' single or double quote, and it is the type attribute that specifies whether this is an opening or closing quote.

A special component of the corpus is the MULTEXT-East English-Slovene '1984'. In addition to alignment segments, the text is also annotated for sentences, of which there can be more than one in a segment. More importantly, the word tokens of this component are marked up for disambiguated lemma and morphosyntactic description, an invaluable annotation for lexical studies, extraction programs and other applications. The Slovene morphosyntactic specifications and the lexicon are further explained in (Erjavec, 1998); below we give the first translation unit from the `orwl` text:

```
<tu lang="en-sl" id="orwl.1">
<seg lang="en">
<s id="Oen.1.1.1.1"><w>It</w> <w>was</w>
<w>a</w> <w>bright</w> <w>cold</w> <w>day</w>
<w>in</w> <w>April</w><c>,</c> <w>and</w>
<w>the</w> <w>clocks</w> <w>were</w>
<w>striking</w> <w>thirteen</w><c>.</c></s>
</seg>
<seg lang="sl">
<s id="Osl.1.2.2.1">
<w lemma="biti" function="Vcps-sma">Bil</w>
<w lemma="biti" function="Vcip3s--n">je</w>
<w lemma="jasen" function="Afpmsnn">jasen</w>
<c>,</c>
<w lemma="mrzel" function="Afpmsnn">mrzel</w>
<w lemma="aprilski" function="Aopmsn">aprilsk
<w lemma="dan" function="Ncmsn">dan</w>
<w lemma="in" function="Ccs">in</w>
<w lemma="ura" function="Ncfpn">ure</w>
<w lemma="biti" function="Vcip3p--n">so</w>
<w lemma="biti" function="Vmps-pfa">bile</w>
<w lemma="trinajst" function="Mcnpnl">trinajs
<c>.</c>
</s>
</seg>
```

The annotation of the corpus is readable directly in the TEI format, but hardly pleasing to the eye. One of the benefits of SGML encoding is easy down-translation for the required application. We have implemented a conversion to HTML for headers and texts. The IJS-ELAN corpus headers and text sample in this rendition are available on the WWW page of the project.

5 Availability of the corpus

The question of reusability has for long been a key issue of digital language resources. It is well known that making such resources is a lengthy process, yet the work is all too often done again and again, because ready-made resources were not available in a usable format or not available to others. Reusability suffers where resources are stored in proprietary, diverse and poorly documented encodings; for this reason we used TEI.

The other obstacle to reusability is that resources are not available for distribution. This is less due to the lack of distribution mechanisms, then to the unwillingness of the corpus owners/producers to distribute them further; copyright restriction can be exercised on the corpus annotation, i.e., on the corpus as a whole, was well as on the component texts.

In line with the idea of the ELAN project, namely to make language resources available to the language community, and our local GNU orientation, we aimed at a very simple distribution mechanism, namely to make the IJS-ELAN deliverables available for downloading via the WWW.

To respect copyright on the original texts, we chose source texts that were available either (1) under the very liberal GNU license, or (2) our institution signed a contract with the text providers, or (3) the texts were available on the WWW, and are copyrighted to a publicly funded source, mostly Offices of the Government of the Republic of Slovenia. Furthermore, the encoding we use impoverishes the original texts so that they are not suitable for quality printing, yet they remain useful for target applications: the corpus is in effect usable over its translation units, but over the complete source documents.

The availability statement of the complete corpus can thus be very liberal, requesting only the acknowledgement of the resource and its sources.

5.1 The distribution

The corpus distribution is available on the IJS-ELAN WWW page, packed as a 3.6 MB `.tar.gz` file, which extracts 22 MB of corpus files into the `IJS-ELAN` directory. The corpus proper consists of $3 + 2 \times 15$ files.

The first three files are the SGML declaration, `ijs-elan.decl`, the second the one-file SGML DTD, `ijs-elan.dtd`, and the third, `ijs-elan.sgml`, the SGML corpus document with the corpus header and references to the corpus components. Each of component is stored in two files, one with the text header, *ID*-`hdr.tei`, and the other with the aligned bi-text itself, *ID*-`txt.tei`.

Further information about the corpus, including headers and samples in HTML, the distribution, and on-line concordancing can be found at the IJS-ELAN WWW page *http://nl.ijs.si/elan/*.

6 Conclusions

The paper presented the IJS-ELAN 1 million word Slovene / English parallel aligned corpus. Parallel aligned corpora are an infrastructure resource for development of multilingual technologies, translation and terminology studies, and this corpus is the first such resource for the Slovene language. Special attention has been given to enabling further distribution of the corpus, by making it available via simple dowloading and by encoding it in a standard format.

While the corpus is TEI conformant, we have not implemented the TEI/CES suggestions for encoding parallel aligned corpora, but rather chose an encoding closer to that of translation memories. This has the effect of simplifying up-translation, of protecting the copyright of the text owners, and of making the structure of the distributable corpus suitable for processing with simple (say, standard Unix) tools. Our translation memory oriented approach to encoding thus minimises the cost of corpus acquisition and processing, and maximises the transparency of the corpus distribution.

At this point it is important for the corpus to be, on the one hand, used, and, on the other, further developed. Initial exploitation has focused on making the corpus available for on-line concordancing and on lexical analysis of the corpus (Špela Vintar, 1999). Further work would involve enriching the annotation of the corpus and, as is always the case with corpora, making it more representative, as regards composition and size.

Currently, the most pressing need and the most interesting task seems to be the lemmatisation and morphosyntactic tagging of the corpus. Such annotation opens opportunity for further computational exploitation, as lemmatised words and simple syntactic patterns can be used in the processing of the corpus. This enables work on shallow syntactic parsing (e.g., bracketing of NPs), term recognition and translation, named entity extraction etc.

Automatic part-of-speech and lemma annotation of the English half should be relatively simple, as there exist publicly available taggers for the language, although it is likely to take up some time. The Slovene part presents significantly greater problems; quality tagging *a la* '1984' means either hand tagging the corpus or having a substantial hand-annotated corpus, with which to train a stochastic tagger and preferably the environment and labour to correct the results.

While we have trained and tested a few taggers on '1984', with seemingly good results (Džeroski et al., 1999), the task becomes much harder in dealing with texts that are lexically and syntactically different from the training set. How to best approach this problem is the topic of further research, most likely in cooperation with partners in the FIDA project (Krek et al., 1998), which aims to build a monolingual reference corpus of the Slovene language.

Acknowledgements

Making the IJS-ELAN corpus would not have been possible without the work of Roman Maurer, Andrej Skubic and Špela Vintar. Thanks is due to the Offices of the Government of Slovenia, especially the Office for European Affairs, to the Linux Users Group of Slovenia, LUGOS, and Lek d.d., OTC Division for providing the source texts for the corpus. The work presented in this paper was in part supported by subcontract to MLIS-ELAN 121 project, Institut für deutsche Sprache, and the grant MZT L2-0461-0106 from the Ministry of Science and Technology of Slovenia.

References

Lars Ahrenberg, Magnus Merkel, Daniel Ridings, Anna Sågvall Hein, and Jörg Tiedemann. 1999. Automatic processing of parallell corpora: A swedish perspective. http://numerus.ling.uu.se/~corpora/plug/.

Susan Armstrong, Masja Kempen, David McKelvie, Dominic Petitpierre, Reinhardt Rapp, and Henry Thompson. 1998. Multilingual corpora for cooperation. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 579–980, Granada. ELRA.

Philippe Di Cristo. 1996. Mtseg: The multext multilingual segmenter tools. MULTEXT Deliverable MSG 1, Version 1.3.1, CNRS, Aix-en-Provence. http://www.lpl.univ-aix.fr/projects/multext/MtSeg/.

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevič, and Dan Tufiş. 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL '98*, pages 315–319, Montréal, Québec, Canada.

Sašo Džeroski, Tomaž Erjavec, and Jakub Zavrel. 1999. Morphosyntactic tagging of slovene: Evaluating pos taggers and tagsets. Research Report IJS-DP 8018, Jožef Stefan Institute, Ljubljana.

Tomaž Erjavec and Nancy Ide. 1998. The MULTEXT-East corpus. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 971–974, Granada. ELRA.

Tomaž Erjavec, Ann Lawson, and Laurent Romary. 1998. East meets West: Producing Multilingual Resources in a European Context. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 233–240, Granada. ELRA. http://www.ids-mannheim.de/telri/cdrom.html.

Tomaž Erjavec. 1998. The Multext-East Slovene Lexicon. In *Proceedings of the 7th Slovene Electrotechnical Conference, ERK '98*, pages 189–192, Portorož, Slovenia. http://nl.ijs.si/et/Bib/ERK98/.

Tomaž Erjavec. 1999. A TEI encoding of aligned corpora as translation memories. In *Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99)*, Bergen. ACL.

Nancy Ide. 1998. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 463–470, Granada. ELRA. http://www.cs.vassar.edu/CES/.

Stig Johansson, Jarle Ebeling, and Knut Hofland. 1996. Coding and aligning the english-norwegian parallel corpus. In K. Aijmer, B. Altenberg, and M. Johansson, editors, *Languages in Contrast*, pages 87–112. Lund University Press. http://www.hit.uib.no/enpc/.

Simon Krek, Marko Stabej, Vojko Gorjanc, Tomaž Erjavec, Miro Romih, and Peter Holozan. 1998. FIDA: korpus slovenskega jezika. http://www.fida.net.

Tony McEnery, Andrew Wilson, Fernando Sánchez-León, and Amalio Nieto-Serrano. 1997. Multilingual Resources in European Languages: Contributions of the CRATER Project. *Literary and Linguistic Computing*, 12(4).

Alan Melby. 1998. Data exchange standards from the OSCAR and MARTIF projects. In *First International Conference on Language Resources and Evaluation, LREC'98*, pages 3–8, Granada. ELRA. http://www.lisa.unige.ch/tmx/.

C. M. Sperberg-McQueen and Lou Burnard, editors. 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford.

Henry Thompson and David McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. In *SGML Europe'97*. http://www.ltg.ed.ac.uk/~ht/sgmleu97.html.

Jörg Tiedemann. 1998. Parallel corpora in Linköping, Uppsala and Göteborg (PLUG). Work package 1., Department of Linguistics, Uppsala University. http://numerus.ling.uu.se/~corpora/plug/.

Špela Vintar. 1999. A Lexical Analysis of the ELAN Slovene-English Corpus. In *Proceedings of the Workshop on Language Technologies - Multilingual Aspects*, pages 63–70, Ljubljana, Slovenia. University of Ljubljana.

Dekai Wu and Xuanyin Xia. 1995. Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon. *Machine Translation*, 9(3-4):285–313.