United Nations General Assembly Resolutions: A Six-Language Parallel Corpus

Alexandre Rafalovitch[1,2]
[1]United Nations
New York
USA
arafalov@gmail.com

Robert Dale[2]
[2]Centre for Language Technology
Macquarie University
Sydney, Australia
rdale@science.mq.edu.au

Abstract

In this paper we describe a six-ways parallel public-domain corpus consisting of 2100 United Nations General Assembly Resolutions with translations in the six official languages of the United Nations, with an average of around 3 million tokens per language. The corpus is available in a pre-processed, formatting-normalized TMX format with paragraphs aligned across multiple languages. We describe the background to the corpus and its content, the process of its construction, and some of its interesting properties.

1 Introduction

Parallel corpora are a useful resource for a wide variety of purposes including the training of machine translation algorithms (Koehn, 2005), multilingual terminology extraction (Le An Ha and Corpas, 2008) and even bootstrapping algorithms for languages that do not enjoy the research resources of English (Yarowsky et al., 2001). Multi-parallel corpora can be more useful than bilingual corpora when the additional languages may be used to assist text alignment (Simard, 1999) or as translation bridge languages, in the sense of (Kumar et al., 2007).

While many European and Germanic languages already have good parallel research corpus resources, such as JRC-Acquis (Steinberger et al., 2006) and EuroParl (Koehn, 2005), material in the Slavic, Sino-Tibetian and Semitic language families is much rarer. The corpus presented here consists of a collection of documents containing manually translated official resolutions of the General Assembly of the United Nations (UN) in the six official languages of the UN: Arabic, Chinese, English, French, Russian, and Spanish. Since resolutions are legally significant, they pass through multiple levels of human translation and verification, and so the translations can be expected to be of high quality.

While documents of the United Nations are already available online via the Official Document System of the UN[1] and are mostly in the public domain (see (STAI, 1987)), they are not in the most convenient form for machine processing. The documents are typically PDF files with reasonably complex typography consisting of two-column text and extended footnotes, and some are available only as images without a text layer. The last public research-ready corpus of UN documents was produced by the Linguistic Data Consortion (LDC; see (Graff, 1994)) and included English, French and Spanish only. Given the linguistic variety encompassed by the UN's official languages, it is therefore somewhat disappointing that this material is not easily available for research in machine translation.

This paper describes some steps towards addressing this concern. The corpus described here contains the text of 2100 resolutions for each language aligned at the level of paragraphs, with just over 74000 paragraphs in each language. The corpus contains an average of around 3 million tokens for each language.[2] The corpus is encoded in XML

[1]See http://documents.un.org.

[2]Counting tokens in the Chinese corpus is difficult; not including the Chinese data, the average token count across the other five languages is 3.11 million.

using Translation Memory eXchange format, with some of the significant sections and text segments marked to assist future research. TMX format was selected as a storage format as it is a standard used in Computer-Assisted Translation tools and has a structure that, while simple, is nonetheless sufficient for our needs.[3]

In this paper, we first provide in Section 2 some background information on the context in which the documents in this corpus appear, before going on to describe the corpus itself and the motivation for the selection of this subset of UN documents in Section 3. Section 4 discusses a number of interesting properties exhibited by the corpus that may encourage specific research directions. Section 5 provides details regarding access and availability.

2 The UN Document Space

In order to understand the nature of this corpus, it may be useful to first understand the overall structure of the United Nations and the range of documents that the organisation produces.

The United Nations is a large international organisation consisting of six principal organs: the General Assembly, the Security Council, the Economic and Social Council, the Trusteeship Council, the International Court of Justice, and the Secretariat. These, together with a number of other agencies, programmes and bodies (such as UNICEF, UNITAR and UNU), form the United Nations family.

Collectively, these bodies generate a massive quantity of documentation: over the last 15 years, about 14000 documents have been published each year, with around half of these being available in all six languages (these are mostly public documents), and the other half in some subset of the languages (these are generally internal documents).

Documents belong to a number of categories:

1. Official resolutions, decisions, statements and legal instruments: these are the outputs of the work by the bodies that define the official positions and provide internal and external guidance. These documents often carry significant legal weight and are widely published and distributed.
2. Reports: these are documents reporting work done; they often serve as inputs for the deliberations of organs, and provide assistance in decision making. A report may also be delivered by one body or organisational unit to another as a way of summarising the decisions taken by the originating unit; for example, the Secretary-General reports to the General Assembly on the work of the Secretariat.
3. Records of meetings and discussions: these may either be verbatim reports or summaries.
4. Letters and notes from the Member States to the organisations.
5. Internal records such as daily journals, agendas of work and draft resolutions.
6. Sales publications, such as books and key reports (e.g., *World Economic Situation and Prospects 2009*).

For non-repudiation reasons, changes and additions to the documents are recorded as separate corrigenda and addenda documents.

The General Assembly (GA) is the main deliberative organ of the United Nations, with a current membership of 192 Member States. Final deliberations of the General Assembly are made in the plenary sessions, but most of the work is done in one of the six main committees or many subcommittees, boards, commissions, working groups and other bodies.[4]

The official output of the GA is collected together as Official Records, which consist of resolutions, decisions and key reports. The GA meets in sessions described as *regular*, *special* and *emergency special*. Regular sessions start in September and last as long as required, often right until the start of the next session; for example, the 62nd regular session started on September 18, 2007 and lasted until September 15, 2008. Regular sessions are divided into two parts: a *main session*, which lasts until the end of the year, and *a resumed session*, which starts in January. Special and emergency special sessions have more flexibility in the organisation of their work.

[3] We use TMX Version 1.4b: see `http://www.lisa.org`.

[4] The six main committees of the General Assembly are: Disarmament and International Security; Economic and Financial; Social, Humanitarian and Cultural; Special Political and Decolonization; Administrative and Budgetary; and Legal.

Draft resolutions are introduced into the discussion under agenda items and may be amended, merged or withdrawn during the course of discussion. Draft resolutions adopted by a committee are then published in that committee's report for further discussion and approval in one of the GA's plenary meetings. Resolutions can be adopted at a plenary meeting by a vote or by an acclamation. The GA adopts around 300 resolutions in any given session.

While non-binding, resolutions of the General Assembly carry legal weight. Final resolutions and decisions of the General Assembly are issued in three volumes of the Official Records: Volume I contains resolutions from the main part of the session; volume II contains decisions from the main part; and volume III contains both resolutions and decisions from the resumed part of the session.

3 The Resolutions Corpus

Of the various types of documents present in the UN document collection, the resolutions are particularly interesting from a machine translation perspective because of their high quality of translation and strict adherence to editorial conventions; they also cover an extremely broad range of topics, whereas other document subsets are more focussed. We chose, therefore, to begin the construction of an NLP-friendly UN corpus with the resolutions data. At around 300 per year, the final resolutions documents are a relatively small proportion of the total number of documents produced by the UN, but the full document count also includes draft resolutions and a large number of other documents that ultimately lead to the final resolutions.

The corpus described here consists of the resolutions in Volume I of the regular sessions of the General Assembly for sessions 55 through 62, corresponding to the period 2000–2007. Prior to this period, and going back to the UN's first session in 1946, the only electronic versions of the resolutions available are scans with no text layers.

3.1 The Nature of Resolutions

An example of the initial fragment of a resolution is shown in Figure 1. A resolution consists of the following parts:

1. A symbol that identifies the resolution, consisting of a number corresponding to the session, and a number corresponding to the ordinal position of this resolution in the series of resolutions adopted in this session; in the present example, this is *59/34*.
2. Information regarding the adoption of the resolution, which may include a list of the Member States that voted on the resolution.
3. The title of resolution; in the present example this is *Nationality of natural persons in relation to the succession of States*.
4. The name of the organ stating the resolution; in the corpus described here, this is always *The*

RESOLUTION 59/34

Adopted at the 65th plenary meeting, on 2 December 2004, without a vote, on the recommendation of the Committee (A/59/504, para. 7)[1]

59/34. Nationality of natural persons in relation to the succession of States

The General Assembly,

Having examined the item entitled "Nationality of natural persons in relation to the succession of States",

Recalling its resolution 54/112 of 9 December 1999, in which it decided to consider at its fifty-fifth session the draft articles on nationality of natural persons in relation to the succession of States prepared by the International Law Commission,

Recalling also its resolution 55/153 of 12 December 2000, the annex to which contains the articles on nationality of natural persons in relation to the succession of States,

Taking into consideration the comments and observations of Governments[2] and the discussion held in the Sixth Committee at the fifty-ninth session of the General Assembly[3] on the question of nationality of natural persons in relation to the succession of States, in particular, to preventing the occurrence of statelessness as a result of a succession of States,

Taking note, in this regard, of the efforts made at the regional level towards the elaboration of a legal instrument on the avoidance of statelessness in relation to State succession,

1. *Reiterates its invitation* to Governments to take into account, as appropriate, the provisions of the articles contained in the annex to resolution 55/153, in dealing with issues of nationality of natural persons in relation to the succession of States;

2. *Encourages* States to consider, as appropriate, at the regional or subregional levels, the elaboration of legal

Figure 1: Initial fragment of a resolution [English]

РЕЗОЛЮЦИЯ 59/34

Принята без голосования на 65-м пленарном заседании 2 декабря 2004 года по рекомендации Комитета (A/59/504, пункт 7)[1]

59/34. Гражданство физических лиц в связи с правопреемством государств

Генеральная Ассамблея,

рассмотрев пункт, озаглавленный «Гражданство физических лиц в связи с правопреемством государств»,

ссылаясь на свою резолюцию 54/112 от 9 декабря 1999 года, в которой она постановила рассмотреть на своей пятьдесят пятой сессии проекты статей о гражданстве физических лиц в связи с правопреемством государств, подготовленные Комиссией международного права,

ссылаясь также на свою резолюцию 55/153 от 12 декабря 2000 года, в приложении к которой содержатся статьи о гражданстве физических лиц в связи с правопреемством государств,

Figure 2: Initial fragment of a resolution [Russian]

General Assembly.

5. Zero or more *preambulatory paragraphs*; these set the context for the rest of the resolution. By editorial convention, each of these begins with the present participle form of a verb or verb phrase in italics.
6. One or more *operative paragraphs* that make up the essence of the resolution: generally speaking, these are the actions the GA wants to see take place. Each is introduced by a present tense verb or verb phrase in italics; the specific choice of verb has some significance. By editorial convention, the operative paragraphs are numbered when there is more than one.

For comparison, the Russian translation of the initial part of this resolution is shown in Figure 2.

Resolutions have an unconventional syntactic and orthographic structure. In each language, from the name of the organ onwards, the resolution takes the form of a single, extended sentence, where the sentence is broken into a series of distinct paragraphs. Each orthographic paragraph is therefore really what we would normally think of as a complex clause. These characteristics, and a number of other typographic features, are dictated by the resolution editing conventions (United Nations, 1983).

Language	# tokens	# characters (M)
English	3067550	20.7
French	3442254	22.8
Spanish	3581566	22.9
Russian	2748898	22.0
Chinese	–	5.7
Arabic	2721463	17.2

Table 1: Corpus statistics for the six languages. Token count for the Chinese data is omitted because of the difficulty in providing a reliable or meaningful number for comparison purposes.

While simple resolutions contain only the elements listed above, more complex resolutions can contain additional sections, often with their own titles and/or preambles. Furthermore, some resolutions contain annexes and embedded texts that may not follow the editorial conventions; and tables may also appear.[5]

3.2 The Content of the Corpus

For each language, the corpus contains just over 74000 paragraphs of text, and, for English, around 3 million tokens. Table 1 provides statistics on the data for the six languages; Figure 3 shows the 20 most common tokens that appear in five of the languages. The most frequent words in this corpus are consistent with those found in other corpora, with the unsurprising exception of the appearance of the terms *United* and *Nations* and a few other domain specific elements.

Within the UN's own document processing environment, the resolutions that make up Volume I are grouped together in seven large Microsoft Word files: six of these contain resolutions that came through one of the six main committees, and the seventh contains those that were introduced directly to the plenary meeting.

3.3 Building the Corpus

We extracted the individual resolutions from these files and converted them into basic HTML format

[5]These items are relatively rare. In the 2100-resolution corpus described here, 41 (1.95%) contain tables and 71 (3.38%) contain annexes; 27 (1.29%) of these contain both.

Rank	English		Arabic		French		Russian		Spanish	
1.	280459	the	96368	في	207935	de	146087	и	335993	de
2.	187989	of	45628	من	137638	et	105426	в	184648	la
3.	147575	and	39314	-	130461	la	39662	по	140746	y
4.	100091	to	38861	على	105192	des	38916	на	97550	en
5.	69888	in	34724	إلى	89772	les	28157	с	88149	los
6.	37652	on	24234	أن	80308	à	21005	Объединенных	86136	el
7.	32758	for	20629	الأمم	70140	le	19977	Организации	81001	las
8.	25121	United	19450	المتحدة	67401	du	19248	,	79506	a
9.	22717	that	19268	التي	48013	en	18637	о	69536	que
10.	21207	its	18940	وإذ	33464	pour	15677	для	56189	del
11.	20505	with	16509	([1]	30430	que	15095	от	37867	para
12.	20190	a	15331	عن	30288	dans	14599	Наций	28502	con
13.	20040	as	11590	الدول	29612	;	14186	что	23567	por
14.	18949	Nations	11157	و	26524	sur	13868	к	22783	su
15.	18228	by	11059	أو	24242	aux	12807	также	21911	sobre
16.	13994	at	10529	،(	22727	au	12361	года	21448	al
17.	13589	States	10209	المؤرخ	20468	Nations	9487	года,	21330	Naciones
18.	12835	all	9681	القرار	18826	par	9354	декабря	15615	se
19.	12069	international	9633	كانون	17120	qui	8535	их	15211	Estados
20.	11179	their	9493	مع	15048	Unies	8399	призывает	15089	Unidas

Figure 3: The 20 most frequent tokens in five of the languages

using Word's HTML export capability. The HTML output was then run through a cleanup process that would allow only basic paragraph marking and typographic markup (normally italics) at the start of preambulatory and operative paragraphs, indicating lead-in phrases. Additionally, some normalization was performed to account for the fact that formatting that looks continuous within MS Word may actually consist of multiple formatting segments on the HTML code level; for example, a contiguous sequence of italicised words may appear in the source as a sequence of distinct italicisation events. Inconsistencies with the use of quotes and non-breaking spaces were also normalized. Finally, tables were stripped from the text as they mostly contain numbers and form nested paragraph structures, which are difficult to represent in TMX form.

From the HTML format, the multiple language versions for the same resolution symbol (the identification numbers introduced earlier) were aligned, using the assumption that the translations were strict at the level of formatting as well as at the level of content. In a small number of cases, the Word formatting caused problems (typically the introduction of spurious paragraph breaks); these were fixed manually. The aligned resolution texts were then converted into TMX format, while at the same time marking the adoption information section, incorporating and marking footnotes, and converting lead-in phrase marking into standard TMX markup. An example is provided in Figure 4.[6]

Given the availability of a number of existing tokenisers that users might wish to apply to the texts for different purposes, we have not carried out a complete tokenization of the corpus. However, we have marked document symbols, since these tokens might be problematic for standard tokenisers.

4 Interesting Properties of the Corpus

4.1 Document Symbols

Because of their importance, it is essential that the scope for misinterpretation of resolutions be minimised. To this end, these documents generally contain all relevant context and make heavy use of fully-explicit and unambiguous document symbol

[6]Note that the Arabic token ordering displays incorrectly here as a consequence of the XML labels.

```
<?xml version="1.0"?>
<tmx version="1.4">
 <header
  creationtool="ORESAligner" creationtoolversion="1.0"
  datatype="plaintext" segtype="paragraph"
  adminlang="en-us" srclang="EN" o-tmf="ORES"
 >
 </header>
 <body>
  <tu tuid="59-35:5">
   <tuv xml:lang="EN">
    <seg><hi type="lead">Recalling</hi> its resolution <hi type="symbol">56/83</hi> of
12 December 2001, the annex to which contains the text of the articles on
responsibility of States for internationally wrongful acts,</seg>
   </tuv>
   <tuv xml:lang="FR">
    <seg><hi type="lead">Rappelant</hi> sa résolution <hi type="symbol">56/83</hi> du
12 décembre 2001, en annexe à laquelle figure le texte des articles sur la
responsabilité de l'État pour fait internationalement illicite,</seg>
   </tuv>
   <tuv xml:lang="ES">
    <seg><hi type="lead">Recordando</hi> su resolución <hi type="symbol">56/83</hi>,
de 12 de diciembre de 2001, cuyo anexo contiene el texto de los artículos sobre la
responsabilidad del Estado por hechos internacionalmente ilícitos,</seg>
   </tuv>
   <tuv xml:lang="RU">
    <seg><hi type="lead">ссылаясь</hi> на свою резолюцию <hi type="symbol">56/83</hi>
от 12 декабря 2001 года, в приложении к которой содержится текст статей об
ответственности государств за международно-противо-правные деяния,</seg>
   </tuv>
   <tuv xml:lang="ZH">
    <seg><hi type="lead">回顾</hi>其2001年12月12日第<hi
type="symbol">56/83</hi>号决议，其附件载有国家对国际不法行为的责任的条款案文，</seg>
   </tuv>
   <tuv xml:lang="AR">
    <seg><hi type="lead">إذ تشير</hi> إلى قرارها <hi type="symbol">56/83</hi> المؤرخ
12 كانون الأول/ديسمبر 2001 الذي يتضمن مرفقه نص المواد المتعلقة بمسؤولية الدول عن
الأفعال غير المشروعة دوليا،</seg>
   </tuv>
  </tu>
 </body>
</tmx>
```

Figure 4: A six-way aligned resolution paragraph.

Document Description	Symbol Used
A document from the 62nd session of the General Assembly	A/62/100
A resolution from the 52nd session of the General Assembly	52/215 A to D
A Security Council resolution adopted in the year 2000	1325 (2000)
Second addendum to the document of the Commission on Human Rights (CN.4) of the Economic and Social Council	E/CN.4/1998/53/Add.2
A document with a dual symbol, one from the General Assembly and one from the Security Council	A/50/60-S/1995/1
A resolution from the 50th session of the International Atomic Energy Agency	GC(50)/RES/16

Table 2: Document symbols

references. Similar to complex symbols in the biological domain (Proux et al., 1998), these symbols may include slashes, dashes, full stops, brackets and spaces. Some examples are shown in Table 2.

While it is not practical to identify all possible symbol variations,[7] we have developed a set of regular expressions to locate and mark a significant proportion of the symbols in the corpus.

4.2 Lead-in Phrases

As noted above, preambulatory and operative paragraphs begin with specially-marked lead-in phrases based on verbs whose meaning carries some significance. While there are no official guidelines on what constitutes an acceptable phrase, the requirements of reliable translation tend to limit the chosen words and phrase forms to a number of popular choices used in the majority of the cases. Table 3 shows the ten most frequent lead-in phrases across three of the six languages.

4.3 Named Entity Mentions

United Nations documentation in general, and the resolutions of the General Assembly in particular, include a large number of complex named entity mentions referring to a broad variety of entity types. Here are some examples, separated by semi-colons:

Bodies: United Nations; International Atomic Energy Agency; United Nations Educational, Scientific and Cultural Organization.

Organisational units: General Assembly; Economic and Social Council; the Advisory Committee on the United Nations Programme of Assistance in the Teaching, Study, Dissemination and Wider Appreciation of International Law; Open-ended Ad Hoc Working Group on the Causes of Conflict and the Promotion of Durable Peace and Sustainable Development in Africa.

Agents: Secretary-General; United Nations Special Representative for Children and Armed Conflict; the Special Rapporteur of the Commission on Human Rights.

As can be seen from these examples, named entity mentions can be very long and may contain tokens, such as commas, that are normally treated as delimiters.

5 Corpus Availability

The corpus described here is available from `http://www.uncorpora.org` as a 49.4Mb zip file that contains just over 74000 pre-processed, formatting-normalised aligned paragraphs in Translation Memory eXchange (TMX) 1.4b format. As noted earlier, special markup is included for document symbols and the lead-in phrases in preambulatory and operative paragraphs; footnote content is also marked. Tables have been removed. The voting information is also marked specially, as it contains country lists in alphabetic order, and may not particularly useful for alignment purposes.

Basic utilities are provided to manipulate, extract, or delete specially tagged areas as well as to extract specific languages from the six-language set.

[7] (Griffiths, 2005) presents an analysis of 14000 symbols assigned in one year, and identifies a number of flaws that include inconsistent application of editorial conventions.

Rank	English		French		Spanish	
1	3707	Requests	3509	Prie	3905	Pide
2	3325	Recalling	3383	Rappelant	3323	Recordando
3	2177	Calls upon	1973	Demande	1877	Exhorta
4	1927	Welcomes	1870	Décide	1871	Insta
5	1797	Decides	1738	Souligne	1796	Decide
6	1688	Urges	1660	Invite	1587	Alienta
7	1604	Encourages	1446	Réaffirme	1502	Reconociendo
8	1402	Invites	1407	Réaffirmant	1396	Invita
9	1350	Recognizing	1361	Se félicite	1291	Reafirmando
10	1269	Reaffirming	1273	Encourage	1257	Reafirma

Table 3: The 10 most frequent lead-in phrases in the three languages

6 Conclusions

In this paper we have described a unique six-language parallel corpus consisting of 2100 UN resolutions, multiply-aligned and marked-up for a number of constituent phenomena. The variety of language families present is particulary of interest for work based on the use of bridge languages (Kumar et al., 2007).

We see this as the first step in the construction of a constantly growing corpus of aligned documents harvested from the UN's document collection. We encourage wide use of the corpus; our sponsors will be likely to support further extension if there is a perceived value in its availability.

Acknowledgments

We would like to thank the Department for General Assembly and Conference Management of the United Nations Secretariat for providing access to the source documents.

References

D Graff. 1994. UN parallel text (complete). Linguistic Data Consortium, Philadelphia.

D Griffiths. 2005. The united nations classification scheme a critique and recommendations. *Cataloging and Classification Quarterly*, 40(1):19–41.

P Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*, Phuket, Thailand, September.

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint EMNLP-CoNLL Conference*, pages 42–50, Prague, Czech Republic, June. Association for Computational Linguistics.

Ruslan Mitkov Le An Ha, Gabriela Fernandez and Gloria Corpas. 2008. Mutual bilingual terminology extraction. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

D Proux, F Rechenmann, L Julliard, V Pillet, and B Jacq. 1998. Detecting gene symbols and names in biological texts: A first step toward pertinent information extraction. In S. Miyano and T. Takagi, editors, *Genome informatics: Workshop on Genome Informatics*, volume 9, pages 72–80, Tokyo, Japan, December.

Michel Simard. 1999. Text-translation alignment: Three languages are better than two. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 2–11.

STAI. 1987. UN Administrative Instruction ST/AI/189/Add.9/Rev.2.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, Genoa, Italy, May.

United Nations. 1983. United Nations Editorial Manual. Department of Conference Services ST/DCS/2, Sales No. E. 83.I.16.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.