The United Nations Parallel Corpus v1.0

Michał Ziemski[⋆], **Marcin Junczys-Dowmunt**[†‡], **Bruno Pouliquen**[‡]
[∗]United Nations, DGACM, New York, United States of America
[†]Adam Mickiewicz University, Poznań, Poland
[‡]World Intellectual Property Organization, Geneva, Switzerland
mziemski@unog.ch, junczys@amu.edu.pl, bruno.pouliquen@wipo.int

Abstract

This paper describes the creation process and statistics of the official United Nations Parallel Corpus, the first parallel corpus composed from United Nations documents published by the original data creator. The parallel corpus presented consists of manually translated UN documents from the last 25 years (1990 to 2014) for the six official UN languages, Arabic, Chinese, English, French, Russian, and Spanish. The corpus is freely available for download under a liberal license. Apart from the pairwise aligned documents, a fully aligned subcorpus for the six official UN languages is distributed. We provide baseline BLEU scores of our Moses-based SMT systems trained with the full data of language pairs involving English and for all possible translation directions of the six-way subcorpus.

Keywords: United Nations, parallel corpus, statistical machine translation

1. Motivation

The United Nations[1] (UN) is mandated to publish documents in six official languages and built up a considerable archive of parallel documents from its own translation operations. Multilingualism is a strategic priority for the United Nations, as an essential factor in harmonious communication among peoples.

The official publication of this corpus is a reaction to the growing importance of statistical machine translation (SMT) within the UN Department for General Assembly and Conference Management (DGACM) translation services. In 2011, a research project — in cooperation with the World Intellectual Property Organization (WIPO) — to explore a prototype SMT system based on the TAPTA system used at WIPO (Pouliquen et al., 2011) for the language pair English-Spanish (Pouliquen et al., 2012) was spearheaded by the Spanish Translation Service (STS) in New York and quickly noticed by other United Nations language services. Further development underlined the good performance of the SMT approach and its applicability to UN translation services. The system was expanded (Pouliquen et al., 2013) to a total of 10 language pairs, resulting in a production-grade cloud-based SMT service called TAPTA4UN. Especially since its integration with the in-house computer assisted translation (CAT) tool eLUNa, TAPTA4UN has become a critical global tool in the UN translation toolkit.

DGACM publishes documents in the six official UN languages and additionally in German[2] and is running major translation operations in various locations. The global DGACM translation output for 2014 alone was 231 million words. The translated documents are hosted on the Official Document System[3] (ODS) and are publicly available. Historically, this parallel data has been a major resource for SMT and NLP research, and has resulted in various (unofficial) corpora, most of them incomplete due to resorting to scraping ODS (Rafalovitch and Dale, 2009; Eisele and Chen, 2010; Chen and Eisele, 2012). Other resources are also available from the Linguistic Data Consortium (LDC)[4]. Depending on the language pair, the present corpus is between two (e.g. en-fr) to four times (e.g. en-ru) larger than data published by (Chen and Eisele, 2012); half of the documents are available for all six languages.

The scope of documents used for the SMT models has continuously expanded as additional United Nations documents have become available. The present corpus is the result of this going collection process. The sharing of technology, expertise, and data has proven to be a crucial factor in enabling the adoption of machine translation (MT) at the UN. In the past, DGACM has successfully shared its translation models with other organizations such as WIPO, IMO (Pouliquen et al., 2015), FAO and ILO. Consequently, in order to facilitate research into and the adoption and development of SMT, DGACM is making available a more complete corpus of its parallel documents in a reusable format, including sentence level alignments.

2. License and Availability

The UN parallel corpus is composed of official records and other parliamentary documents of the United Nations that are in the public domain. The UN corpus will be made available for download at http://conferences.unite.un.org/UNCorpus.

The following disclaimer[5], an integral part of the corpus, shall be respected with regard to the United Nations Parallel Corpus v1.0 (no other restrictions apply):

- The UN corpus is made available without warranty of any kind, explicit or implied. The United Nations

[1] Referring to the Department for General Assembly and Conference Management which is responsible for the document processing chain, including translation, of the UN Secretariat

[2] Only some documents are translated by the German Translation Section in New York.

[3] http://ods.un.org

[4] See e.g. Franz, Alex, Shankar Kumar, and Thorsten Brants. 1993-2007 United Nations Parallel Text LDC2013T06. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

[5] Drafted with the advice of the General Legal Division, Office of Legal Affairs, United Nations.

```
<TEI.2>
  <teiHeader>
    <fileDesc>
      <publicationStmt>
        <date>20100713</date>
        <idno type="symbol">CD/1890</idno>
        <idno type="jobno">G1061646</idno>
        [...]
        <keywords>
          <term>ARMS RACE</term>
          <term>OUTER SPACE</term>
          <term>INTERNATIONAL SECURITY</term>
        </keywords>
        [...]
  </teiHeader>
  <text>
    <body>
      <p id="1">
        <s id="1:1" lang="en">CD/1890</s>
      </p>
      [...]
      <p id="6">
        <s id="6:1" lang="en">The permanent Mission of C
        <s id="6:2" lang="en">The conference took place
```

(a) English sample document (some elements were omitted)

```
<linkGrp fromDoc="Xml/fr/2010/cd/1890.xml" toDoc="Xml/en
/2010/cd/1890.xml" score="0.352899">
<link type="1-1" xtargets="1:1;1:1" score="1"/>
<link type="1-1" xtargets="2:1;2:1" score="1"/>
<link type="1-1" xtargets="3:1;3:1" score="1"/>
<link type="0-1" xtargets=";4:1" score="0"/>
<link type="1-1" xtargets="4:1;5:1" score="0.733075"/>
<link type="1-1" xtargets="5:1;6:1" score="0.613475"/>
<link type="1-1" xtargets="6:1;7:1" score="0.648559"/>
<link type="1-1" xtargets="6:2;7:2" score="0.662173"/>
<link type="1-1" xtargets="7:1;8:1" score="0.416193"/>
<link type="1-1" xtargets="8:1;9:1" score="0.428882"/>
<link type="1-1" xtargets="9:1;10:1" score="1"/>
<link type="1-1" xtargets="10:1;11:1" score="1"/>
<link type="1-1" xtargets="11:1;12:1" score="0.738796"/>
<link type="0-1" xtargets=";13:1" score="0"/>
<link type="1-1" xtargets="12:1;14:1" score="0.638055"/>
<link type="1-1" xtargets="13:1;15:1" score="0.317246"/>
<link type="1-1" xtargets="13:2;15:2" score="0.565939"/>
<link type="1-1" xtargets="14:1;16:1" score="0.164868"/>
<link type="1-1" xtargets="14:2;16:2" score="0.35008"/>
<link type="1-1" xtargets="14:2;16:2" score="0.35008" />
<link type="1-1" xtargets="14:3;16:3" score="0.285692" /
<link type="1-1" xtargets="14:4;16:4" score="0.41574" />
```

(b) Sentence alignment information for two documents

Figure 1: TEI-based XML format of raw corpus files

specifically makes no warranties or representations as to the accuracy or completeness of the information contained in the UN corpus.

- Under no circumstances shall the United Nations be liable for any loss, liability, injury or damage incurred or suffered that is claimed to have resulted from the use of the UN corpus. The use of the UN corpus is at the user's sole risk. The user specifically acknowledges and agrees that the United Nations is not liable for any conduct of any user. If the user is dissatisfied with any of the material provided in the UN corpus, the user's sole and exclusive remedy is to discontinue using the UN corpus.
- When using the UN corpus, the user must acknowledge the United Nations as the source of the information. For references, please use this very publication.
- Nothing herein shall constitute or be considered to be a limitation upon or waiver, express or implied, of the privileges and immunities of the United Nations, which are specifically reserved.

3. File Organization and Format

All documents are organized into folders by language, publication year, and publication symbols. Corresponding documents are placed in parallel folder structures, and a document's translation in any of the official languages (if it exists) can be found by inspecting the same file path in the required language subfolder.

For individual documents, it was decided to follow the TEI-based format of the JRC-Aquis parallel corpus (Steinberger et al., 2006). Documents retain the original paragraph structure and sentence splits have been added automatically (see Figure 1a; details on the processing steps are given in Section 5.). Documents for which multiple language versions exist have corresponding link files (Figure 1b) for each of the maximum 15 language pairs. They contain information about the alignment link type, ids of linked sentences (`xtargets`) and the alignment quality score.

We also make available plain-text bitexts that span all documents for a specific language pair and can be used more readily with SMT training pipelines.

4. Document Meta-Information

Every document in XML file format has embedded meta-information:

Symbol Each UN document has a unique symbol[6] which is common for all language versions.

Translation job number A unique language-specific document identifier.

Publication date The original publication date for a document by symbol, which applies to all language versions. This date does not necessarily correspond to the release date of each individual document.

Processing place Possible locations are New York, Geneva and Vienna.

Keywords Any number of subjects covered by the document, according to the ODS subject lexicon, which is based on the UNBIS Thesaurus[7].

5. Creating the Parallel Corpus

During processing, we differentiate between primary and secondary language pairs. Primary language pairs consist of one non-English language and English. Secondary language pairs are formed from non-English language pairs.

Figure 2 illustrates all the processing steps for creating the sentence alignment link file from two parallel documents

[6] A detailed description of these symbols can be found at http://research.un.org/en/docs/symbols

[7] http://lib-thesaurus.un.org/

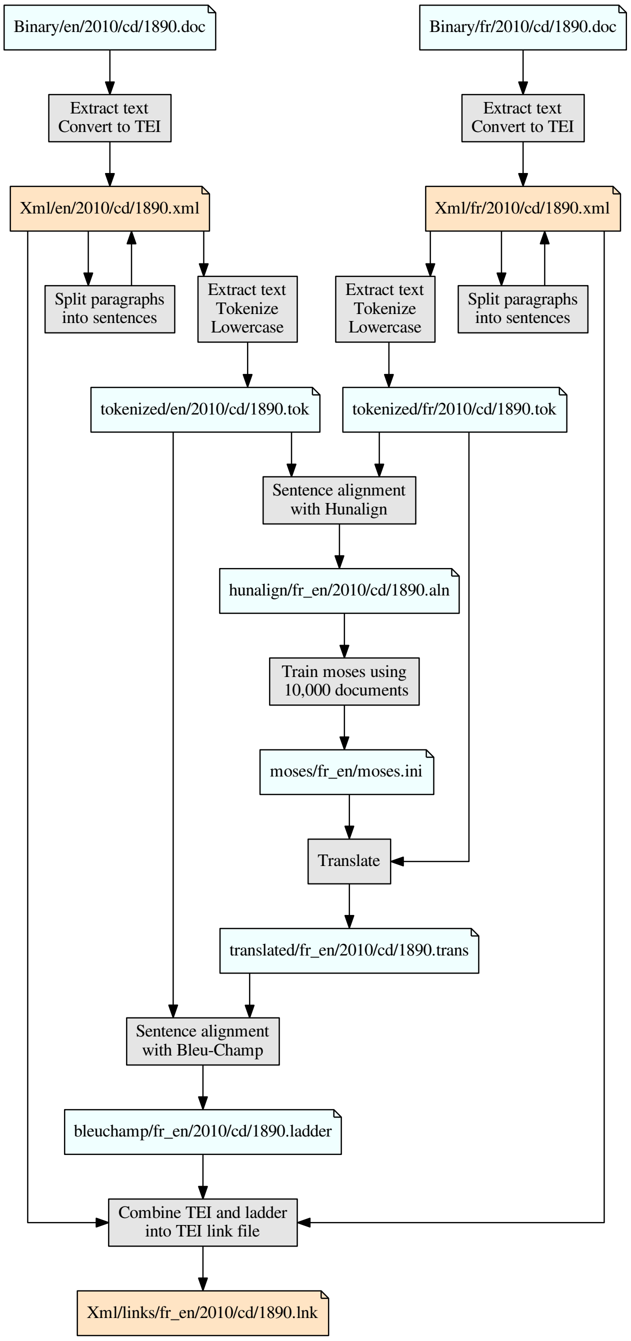

Figure 2: Sentence alignment dependency graph

for a primary language pair, here English-French. The dependency graph featured is modeled very closely after our pipeline based on GNU Make.

After converting binary formats (MS Word, WordPerfect) to the presented TEI-XML format, sentence splitting[8] is applied to the XML file, retaining the original paragraph structure as shown in Figure 1a.

To ensure a high-quality sentence alignment, we rely on a two-step approach similar to Sennrich and Volk (2011). French documents are translated into English first. We randomly select a subset of 10,000 document pairs and align them using Hunalign (Varga et al., 2005), selecting only 1-1 alignments that are themselves surrounded by 1-1 alignments. This small lower-quality parallel corpus is used to train an SMT system with Moses (Koehn et al., 2007). Following Sennrich and Volk (2011) we use significance pruning (Johnson et al., 2007) to filter out noise resulting from alignment errors.

Next, our monolingual sentence aligner BLEU-Champ[9] is applied. BLEU-Champ relies on smoothed sentence level BLEU-2 as a similarity metric between sentences and uses the Champollion algorithm (Ma, 2006) with that metric. In order to avoid computational bottlenecks for long documents, first a path consisting only of 0-1, 1-0, 1-1 alignments is calculated. In a second step, the search is restricted to a 10-sentence-wide corridor around the best path allowing for all alignment combinations up to 4-4 alignments. This procedure avoids search errors and is fast enough to use the Champollion algorithm with documents consisting of thousands of sentences. Given the English tokenized text and the translated French text, BLEU-Champ produces a ladder file (Hunalign's numeric alignment format) which eventually is combined with the two TEI documents to form the final TEI sentence alignment file (see Figure 1b).

The XML and link files in Figure 2 are distributed as part of the corpus. Since the link files contain pointers to the original XML documents, any set of link files can be used to produce plain-text parallel corpora.

In the case of secondary language pairs, the same steps are followed, except that both documents are translated into English and sentence alignment is performed on the English translation results of both files.

6. Statistics

Statistics for all language pairs are presented in Table 1a. We also make available a fully aligned subcorpus (Table 1c). This subcorpus consists of sentences that are consistently aligned across all languages with the English primary documents. We believe this might be one of the largest resources of this kind and of particular value for comparative linguistic research.

7. Test and Development Data

Documents released in 2015 (excluded from the current corpus) were used to create official development and test sets for machine translation tasks. Development data was randomly selected from documents that were released in the first quarter of 2015 and test data was selected from the second quarter. To avoid repetitions, we only chose translation tuples for which the English sentence was unique. We also skewed the distribution of sentence lengths slightly by requiring that half of the sentences not be chosen if their length was below 50 characters and not imposing any restrictions on the other half. This was done to reduce the occurrence of formulaic and less informative sentences.

Both sets comprise 4,000 sentences that are 1-1 alignments across all official languages. As in the case of the fully aligned subcorpus, any translation direction can be evaluated (see Table 2b).

[8] Using Eserix, an SRX-based sentence splitter `https://github/emjotde/eserix`. The algorithm and rules have been extracted from Psi-Toolkit (Graliński et al., 2012).

[9] `https://github/emjotde/bleu-champ`

	ar	en	es	fr	ru	zh
ar	–	111,241 18,539,207	113,065 18,578,118	112,605 18,281,635	111,896 18,863,363	91,345 15,595,948
en	456,552,223 512,087,009	–	123,844 21,911,121	149,741 25,805,088	133,089 23,239,280	91,028 15,886,041
es	459,383,823 593,671,507	590,672,799 678,778,068	–	125,098 21,915,504	115,921 19,993,922	91,704 15,428,381
fr	452,833,187 597,651,233	668,518,779 782,912,487	674,477,239 688,418,806	–	133,510 22,381,416	91,613 15,206,689
ru	462,021,954 491,166,055	601,002,317 569,888,234	623,230,646 513,100,827	691,062,370 557,143,420	–	92,337 16,038,721
zh	387,968,412 387,931,939	425,562,909 381,371,583	493,338,256 382,052,741	498,007,502 377,884,885	417,366,738 392,372,764	–

(a) Statistics for pair-wise aligned documents. Cells above the diagonal contain the number of documents and lines per language pair. Cells below the diagonal contain tokens numbers in a language pair — the upper number refers to the language in the column title, the lower to the language in the row title. Tokens were counted after processing with the Moses tokenizer. For Chinese, Jieba was used before applying the Moses tokenizer with default settings.

Total documents	Aligned document pairs
799,276	1,727,539

(b) Document statistics

Documents	Lines	English Tokens
86,307	11,365,709	334,953,817

(c) Statistics for fully aligned subcorpus

Table 1: Statistics for the United Nations Corpus v1.0 (1990 – 2014)

8. Machine Translation Baselines

Based on the described test sets we also provide baseline results for our in-house Moses (Koehn et al., 2007) systems that were trained on the described data.

Sentences longer than 100 words were discarded. To speed up the word alignment procedure, we split the training corpora into four equally sized parts that are aligned with MGIZA++ (Gao and Vogel, 2008), running 5 iterations of Model 1 and the HMM model on each part.[10] We use a 5-gram language model trained from the target parallel data, with 3-grams or higher order being pruned if they occur only once. Apart from the default configuration with a lexical reordering model, we add a 5-gram operation sequence model (Durrani et al., 2013) (all n-grams pruned if they occur only once) and a 9-gram word-class language model with word-classes produced by word2vec (Mikolov et al., 2013) (3-grams and 4-grams are pruned if they occur only once, 5-grams and 6-grams if they occur only twice, etc.), both trained using KenLM (Heafield et al., 2013). To reduce the phrase-table size, we apply significance pruning (Johnson et al., 2007) and use the compact phrase-table and reordering data structures (Junczys-Dowmunt, 2012). During decoding, we use the cube-pruning algorithm with stack size and cube-pruning pop limits of 1,000.

All scores are provided for lowercased data; the data was tokenized with the Moses tokenizer. For Chinese segmentation we used Jieba[11] before applying the Moses tokenizer.

Full Data into and from English At DGACM, translation is mainly done between English and the remaining languages. Hence, we have in-house translation systems for these language pairs that are being used in production[12]. Table 2a contains the most recent results for these systems trained with all the available data for a language pair.

	ar	es	fr	ru	zh
en→	42.04	61.35	50.33	43.89	37.68
en←	54.01	60.38	52.58	53.53	43.68

(a) BLEU scores from and into English for all available data

→	ar	en	es	fr	ru	zh
ar	–	53.07	49.77	42.80	36.00	31.58
en	41.96	–	61.26	50.09	43.25	37.84
es	38.13	59.89	–	49.76	39.69	31.27
fr	34.43	52.22	52.44	–	36.48	29.98
ru	34.43	52.59	49.61	43.37	–	32.63
zh	28.02	42.97	39.64	34.42	29.57	–

(b) BLEU scores matrix of the fully aligned subcorpus

Table 2: BLEU scores for baseline systems

Fully aligned Subcorpus In Table 2b, we provide BLEU scores for the entire translation matrix for all official languages from the fully aligned subcorpus. These systems are not used as in-house translation systems and were produced as an academic exercise.

The results do not differ significantly for common translation directions in both settings despite the differences in absolute data sizes. We speculate that this may be caused

[10] We confirmed that there seemed to be no quality loss due to splitting and limiting the iterations to simpler alignment models.

[11] https://github.com/fxsjy/jieba

[12] The data collection efforts for this publication also resulted in a considerably larger set of training data for our own systems.

by the fact that the fully aligned corpus covers 80–90% of the documents even if sometimes only 50% of the segments are present. Optimizer instability during parameter tuning or a certain degree of saturation might be other factors.

9. Conclusions

The publication of the United Nations Parallel Corpus v1.0 makes a more complete resource of UN documents available to the general public. It is the result of the continuous effort and dedication to multilingualism.

The alignment links provided allow for experiments with language pairs, for instance Arabic-Chinese, that have not been widely investigated. Our baselines and test sets can serve as reference data for future publications and we would like researchers to explore machine translation techniques beyond the phrase-based approach that was used to produce them. The fully aligned subcorpus in particular may prove a valuable resource for studying pivoting techniques and multi-source or multi-target approaches. The meta-information and preserved document structure provided can help to advance recent work in document-level translation. We are keen to test the most promising results in our own systems.

In the future, we hope to publish updated versions of the presented parallel corpus, expanding forward and backwards in time.

10. Acknowledgements

The authors wish to thank Ms Catherine Pollard, Under-Secretary-General for General Assembly Affairs and Conference Management, and Ms Cecilia Elizalde, Director, Documentation Division, DGACM, for making the publication of the UN parallel corpus possible. The authors further wish to thank Mr Michael Rudolph Ten-Pow, Senior Adviser on Multilingualism, Michelle Keating, Chief, Languages Service, UNOG, and Ms Danielle Henripin, UNHQ, for all their support.

11. References

Chen, Y. and Eisele, A. (2012). MultiUN v2: UN documents with multilingual alignments. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Durrani, N., Fraser, A., Schmid, H., Hoang, H., and Koehn, P. (2013). Can Markov models over minimal translation units help phrase-based SMT? In *ACL*, pages 399–405. The Association for Computer Linguistics.

Eisele, A. and Chen, Y. (2010). MultiUN: A multilingual corpus from United Nation documents. In *Language Resources and Evaluation*.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. ACL.

Graliński, F., Jassem, K., and Junczys-Dowmunt, M. (2012). PSI-Toolkit: Natural Language Processing Pipeline. *Computational Linguistics - Applications*, pages 27–39.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 690–696.

Johnson, J. H., Martin, J., Forst, G., and Kuhn, R. (2007). Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL'07*, pages 967–975.

Junczys-Dowmunt, M. (2012). Phrasal Rank-Encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *Prague Bull. Math. Linguistics*, 98:63–74.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. ACL.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC-2006*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Pouliquen, B., Mazenc, C., and Iorio, A. (2011). TAPTA: A user-driven translation system for patent documents based on domain-aware statistical machine translation. In *Proceedings of the 15th Annual Conference of the EAMT*, pages 5–12, Trento, Italy.

Pouliquen, B., Mazenc, C., Elizalde, C., and Garcia-Verdugo, J. (2012). Statistical machine translation prototype using UN parallel documents. In *16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 12–19, Trento, Italy.

Pouliquen, B., Elizalde, C., Junczys-Dowmunt, M., Mazenc, C., and García-Verdugo, J. (2013). Large-scale multiple language translation accelerator at the United Nations. In *MT-Summit XIV*, pages 345–352.

Pouliquen, B., Junczys-Dowmunt, M., Pinero, B., and Ziemski, M. (2015). SMT at the International Maritime Organization: experiences with combining in-house corpora with out-of-domain corpora. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 202–205, Antalya, Turkey.

Rafalovitch, A. and Dale, R. (2009). United Nations general assembly resolutions: A six-language parallel corpus. In *MT Summit XII*, pages 292–299. International Association of Machine Translation.

Sennrich, R. and Volk, M. (2011). Iterative, MT-based sentence alignment of parallel texts. *18th Nordic Conference of Computational Linguistics, NODALIDA*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., and Tufiş, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.