Shallow-transfer rule-based machine translation from Czech to Polish

Joanna Ruth
Gdańsk University of Technology
Gdańsk
Poland
joannaruth1@gmail.com

Jimmy O'Regan
Eolaistriu Technologies
Thurles
Ireland
joregan@gmail.com

Abstract

This article describes the development of an Open Source shallow-transfer machine translation system from Czech to Polish in the Apertium platform. It gives details of the methods and resources used in constructing the system. Although the resulting system has quite a high error rate, it is still competetive with other systems.

1 Introduction

Czech and Polish are both Western Slavic languages. Polish is spoken by approximately 50 million, Czech by 12 million. "In the 10th century, Czech and Polish were still basically the same language, which then began to diverge from each other, but even until the 14 century, Czechs and Poles understood each other without problems".[1] (Wikipedie, 2010)

Czech and Polish are typologically similar languages. They are both medium inflected languages with relatively free word order, sharing seven cases and three grammatical genders. In addition, both languages draw a distinction between animate and inimate masculine nouns, while Polish draws a further distinction with masculine animate nouns that refer to people.[2] We chose to treat these animacy differences as separate genders, to simplify concordance operations. Although this greatly reduces the complexity of transfer between the languages, it introduces a number of artificial ambiguities.

Several multilingual corpora are available that feature both Czech and Polish, such as JRC Acquis (Ralf et al., 2006), OPUS (Tiedemann, 2009), etc.; however, in these corpora most, if not all, of the text has been translated from a third language (usually English),[3] and we considered the potential for "translation drift" to be a serious limiting factor in the creation of a statistical system.

Further, as the primary goal of the project was that it be Open Source, we felt more inclined towards investigating the rule-based paradigm. We were also interested in determining if the successes of previous projects based on the Apertium platform for closely-related languages in the Romance and Germanic families would apply to a Slavic language pair.

We have been guided by earlier work on Czech–Polish rule-based machine translation (Dębowski et al., 2002), and used issues raised by that work as a guideline in creating an Open Source system. In addition, we have also sought to follow the example of two previous Apertium-based projects, for Swedish to Danish (Tyers and Nordfalk, 2009) and Norwegian Nynorsk and Bokmål (Unhammer and Trosterud, 2009).

As the project was begun with known time constraints, it was decided early on that we would

[1] "V 10. století byly čeština a polština v podstatě stále jeden jazyk, pak se začaly od sebe rozcházet, ale ještě ve 14. století si Češi a Poláci bez problémů rozuměli." Authors' translation.

[2] The same distinction is present in Czech, but as it serves no role in concordance, we chose not to treat it specially.

[3] Even if there were direct translations in those corpora, there's no indication in the metadata of the original language.

F. Sánchez-Martínez, J.A. Pérez-Ortiz (eds.)
Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 69–76
Barcelona, Spain, January 2011. http://hdl.handle.net/10609/5650

concentrate on *assimilation* (that is, providing an understanding of the source text). We also chose to focus on the Czech to Polish direction initially, as it is the direction we felt more comfortable with.

Also, as there has not yet been a translation system based on Apertium between medium inflected, free word order languages, to a certain degree we are testing the limits of the Apertium platform: part of the high error rate of the resulting system (73%) could be reduced with some extensions to the Apertium platform.

2 Design

The Apertium machine translation platform uses a shallow-transfer model. Originally designed for the Romance languages of Spain, later development has extended the system to enable the development of translators between more divergent language pairs, such as Basque–Spanish (Ginestí-Rosell et al., 2009).

Apertium is constructed in a modular, assembly-line fashion. A source language text is first morphologically analysed using finite-state transducers. It is then disambiguated for part of speech by a bigram HMM part-of-speech tagger, which produces a single disambiguated form.

The disambiguated text is subsequently processed by the syntactic transfer module, which performs both lexical and structural transfer. Lexical units consisting of lemma, part of speech, and morphological information are matched on the basis of fixed-length patterns, upon which operations such as insertions, removals, reorderings, substitutions and concordancing are performed.

Finally, generation is performed by the same module that performs analysis. Figure 1 shows the main modules of a given system built upon the platform. A more complete description of the platform may be found in Armentano-Oller et al. (2006).

Two models of structural transfer are supported by the platform: a single-stage transfer, where only one set of transfer rules is used, and a three-stage transfer where rules are also used to group words into *chunks*, on which later operations can be performed. The Czech–Polish pair uses three-stage transfer as a means of simplifying the problems of concordance and difference in case governed by preposition mentioned in Dębowski et al. (2002).

3 Development

3.1 Resources

Czech and Polish are both well-resourced languages: "no other Slavic language has so many resources for stochastic natural language processing" as Czech (Hajič et al., 2003).

There are several tools available for both languages, many open source, which we were able to use in the creation or validation of resources. In particular, the Open Source Proofreading tool, LanguageTool (Miłkowski, 2010), includes both morphological analysis and disambiguation for both Polish and Czech, though the Czech disambiguator is incomplete.

Aside from external resources, we were able to take advantage of the Open Source nature of the Apertium platform, and to draw upon other work being done within the project. In the "incubator" module (which houses works in various stages of early development) of Apertium's source repository we were able to use pieces being created for English-Polish and Czech-Slovenian for morphological analysis, and Polish-Slovakian for a set of initial transfer rules.

3.2 Morphological analysis and generation

None of the Apertium "incubator" materials were at an advanced stage of development; in addition, both Czech and Polish analysis modules had serious problems that needed to be resolved. Compounding this, the number of distinct paradigms for nouns and verbs[4], along with the number of forms each contained, lead to files that were difficult to edit using standard tools on an average computer, and made fixing inconsistencies impractical.

To overcome this, we abstracted the common parts of the open categories into sub-paradigms and replaced those entries with a reference to the new paradigm. Figures 2 and 3 illustrate this solution. Some of the paradigms from the Czech–Slovenian translator contained a number of errors

[4]Jagodziński (2008) lists 174 verb paradigms and 146 noun paradigms; Bielec (1998) mentions more, and Futrega (2010) contains yet more, mostly proper nouns.

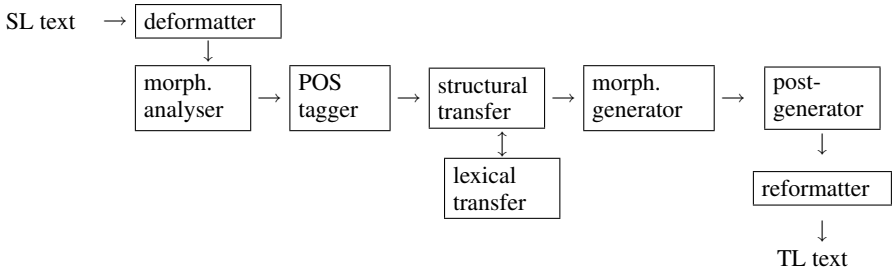

Figure 1: The eight modules of the shallow-transfer machine translation system

```
<pardef n="mat/ka__n">
  <e><p><l>ka</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="nom"/></r></p></e>
  <e><p><l>ki</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="gen"/></r></p></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>kę</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="acc"/></r></p></e>
  <e><p><l>ką</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="ins"/></r></p></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
  <e><p><l>ko</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="voc"/></r></p></e>
</pardef>
<pardef n="dro/ga__n">
  <e><p><l>ga</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="nom"/></r></p></e>
  <e><p><l>gi</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="gen"/></r></p></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>gę</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="acc"/></r></p></e>
  <e><p><l>gą</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="ins"/></r></p></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
  <e><p><l>go</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="voc"/></r></p></e>
</pardef>
```

Figure 2: Paradigms for the singular parts of *matka* and *droga*, written in full form.
Each *e*ntry contains a *p*air, where the **l**eft contains the suffix to be analysed/generated, and the *r*ight contains the suffix of the lemma, along with the set of *s*ymbols to be attached.

```
<pardef n="BASE__matka">
  <e><p><l>a</l><r><s n="f"/><s n="sg"/><s n="nom"/></r></p></e>
  <e><p><l>i</l><r><s n="f"/><s n="sg"/><s n="gen"/></r></p></e>
  <e><p><l>ę</l><r><s n="f"/><s n="sg"/><s n="acc"/></r></p></e>
  <e><p><l>ą</l><r><s n="f"/><s n="sg"/><s n="ins"/></r></p></e>
  <e><p><l>o</l><r><s n="f"/><s n="sg"/><s n="voc"/></r></p></e>
</pardef>
<pardef n="mat/ka__n">
  <e><p><l>k</l><r>ka<s n="n"/></r></p><par n="BASE__matka"/></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>ce</l><r>ka<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
</pardef>
<pardef n="dro/ga__n">
  <e><p><l>g</l><r>ga<s n="n"/></r></p><par n="BASE__matka"/></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="dat"/></r></p></e>
  <e><p><l>dze</l><r>ga<s n="n"/><s n="f"/><s n="sg"/><s n="loc"/></r></p></e>
</pardef>
```

Figure 3: Paradigms for the singular parts of *matka* and *droga*, redefined in terms of a "base paradigm" containing the common parts of both.
The **par** element contains a reference to another, previously defined, paradigm.

and omissions; we used the Czech Free Morphology[5] tool (Hajič, 2004) to validate entries, and to generate missing entries.

To extract entries for the monolingual dictionary, we used a similar process mentioned in Weiss (2005) on the development of Lametyzator: as both Czech and Polish ispell dictionaries were created according to linguistic principles, there was a one-to-one mapping between ispell "flags" with suffix, and paradigms in the morphological dictionaries.

3.3 Corpus collection

Although rule-based machine translation is not corpus based, a corpus is a necessary tool in the empirical evaluation of rule hypotheses. We also wished to collect a set of test sentences to use as a set of "regression tests", to ensure that additions of new rules or changes to existing rules do not accidentally introduce new errors.[6]

Despite the availability of parallel text, we required *direct* translations, and sought our own sources. We built an initial set of sentences from the example sentences given on the Polish edition of Wiktionary. Though there were few sentences (141), they were mostly useful for our purposes.

We also found some material that had been translated from Czech to Polish at an online grammar site.[7] We also found some public domain material from the Polish embassy in Prague, which provided us with a more substantial set of direct translations. The Open Content parts of the collected material has been contributed to the Open-Content Text Corpus (Bański and Wójtowicz, 2010), to contribute to the pool of freely available linguistic resources.

3.4 Disambiguation

For disambiguation we first chose to train a basic unsupervised bigram part-of-speech tagger using the `apertium-tagger` tool. Although there are disambiguators available for both Polish and Czech in LanguageTool, the disambiguators are not integrated with Apertium, and the Czech disambiguator is far from complete.

3.5 Lexical transfer

We were able to, with slight modifications, reuse a set of initial transfer rules from a nascent Polish–Slovakian system[8] under development for Apertium, though the rules were quite basic and many new rules were required.

We used several methods to create a bilingual dictionary. Closed categories (such as pronouns, prepositions, etc.) were added by hand, while semi-automatic methods were used to add the open categories:

- Cognates – we initially used the same process described in Tyers and Nordfalk (2009), where a set of common transformations (such as *v* to *w*, *ovat* to *ować*) are used on a source language list to produce a target language list of cognates, but initial results were not promising. To increase accuracy in the induced lexicon, we limited the search to words which, by their suffixes,[9] seemed likely Latin or Greek derivatives, to adjectives referring to languages.[10] While the yield was low, the accuracy was much higher.

- Wordlists – we used a number of Czech-Polish wordlists, to which part-of-speech and gender information was added. We also used a number of Polish–English and Czech–English wordlists via triangulation, with restrictions based on part-of-speech to filter the resulting lists.

- Wikipedia – we used the process described in Tyers and Pienaar (2008), of collecting translations using Wikipedia *interwiki* links, although modified to remove the following of interwiki links, which, although it provided "close" concepts (such as providing

[5] `http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/index.html`

[6] The set of test sentences is available on the Apertium wiki `http://wiki.apertium.org/wiki/Polish_and_Czech/Pending_tests`

[7] `http://www.finito.zanet.pl/czeski/ath/slawistyka/www.slawistyka.ath.bielsko.pl/czeski/index.html`

[8] Czech and Slovakian are similar enough that word-for-word translation is sufficient (Hajič et al., 2000), which allowed this reuse.

[9] *-ogia, -afia* in Polish to *-ogie, -afie* in Czech

[10] Ending in *-ski* in Polish, with a corresponding *-sku* form, ending in *-sky* in Czech with a corresponding *-ský* form.

diabetes as a translation for *diabetic*) lead to too many incorrect entries.

- poterminology – we used the `poterminology` tool, as described in Unhammer and Trosterud (2009), with the localisation data from the KDE project.

- Probabilistic dictionary – We trained a statistical machine translation system using Moses (Koehn et al., 2007) on the JRC Acquis Corpus (Ralf et al., 2006), extracting the most probable translations.

- "Stupid" word alignment – we passed our parallel text sources though the analysers of both systems, taking sentences with one unknown word each, and selected them as the probable translation.

The bilingual dictionaries were manually checked at several stages of the development process, and bad entries were modified or discarded. In addition, we used the intersection of multiple wordlists, along with output from Google Translate, to partially automate the validation process.

3.6 Syntactic transfer

Czech and Polish are syntactically quite similar, though there are some divergences:

- Preposition changes – Some prepositions take different cases in Czech and Polish, such as *pro* in Czech (accusative) to *dla* in Polish (genitive). We handled these differences as second level transfer operations, to pass the case change to a whole noun phrase chunk at a time.

- Past and conditional tenses – Czech uses the present tense of *být* ("to be") to indicate person in the past and conditional tenses, while in Polish this has become lexicalised, so that an enclitic form of the equivalent *być* is considered part of the conjugation.

- Different modal verbs – Czech uses the conditional particle *by*, the past tense of *mít*, and the infinitive of the verb to express "ought to", while Polish uses the defect modal verb *powinien* with the infinitive.

	Number entries
Monolingual dict. (`pl`)	13,973 lemmas
Bilingual dict.	12,419 lemmas
Monolingual dict. (`cs`)	11,378 lemmas
Transfer rules (`cs` → `pl`)	
Transfer Level 1	49
Transfer Level 2	20
Transfer Level 3	3

Table 1: Status of pair as of SVN revision 27271, 26th November 2010

- Transgressive – The transgressive is a verb form specific to Czech and Slovak, which resembles a nominative-only adjective; the equivalent form in Polish is adverbial, so we need only to discard the extra morphological information.

4 Status

Table 1 gives details of the current status of the system in terms of the number of lemmas in each of the dictionaries and the number of transfer rules. The Polish dictionary contains quite a large amount of duplicate entries, which explains much of the disparity between the numbers.

5 Evaluation

5.1 Coverage

The vocabulary coverage of the system is calculated over an available corpus. Here coverage is defined as *naïve coverage*, that is for any given surface form at least one analysis is returned. This may not be complete. Although we have concentrated on the Czech to Polish direction, we also provide results for the Polish to Czech direction as an indicator of future work. The coverage figures roughly correspond to the first public versions of other Apertium systems. We used the database dumps of the corresponding Wikipedias.[11] The results are presented in table 2.

[11] Czech: `http://cs.wikipedia.org`; Access date: 30th March 2010; Filename: `cswiki-20100330-pages-articles.xml.bz2`. Polish: `http://pl.wikipedia.org`; Access date: 4th January 2010; Filename: `plwiki-20100104-pages-articles.xml.bz2`.

Corpus	Running tokens	Known tokens	Coverage
Polish	39,293,427	27,997,757	71.25%
Czech	17,165,777	10,925,926	63.65%

Table 2: Naïve coverage for both translation directions

Corpus	WER	PWER
News Samples	76 %	62 %
UDHR	47 %	32 %

Table 4: Evaluation results for Google Translate. Free rides and Unknowns did not apply in this case, so were omitted.

5.2 Quantitative

The quantitative evaluation involved the post-edition of 46 human translated sentences (486 words) from the Polish learner's guide to Czech. The sentences were selected from the portion which had been drawn from news sources. As a secondary source, we used the text of the UN Universal Declaration of Human Rights (UDHR), although as the texts are not mutual translations, we can expect a greater degree of divergence.

Both word error rate (WER) and position-independent error rate (PWER) were calculated by counting the number of insertions, substitutions and deletions between the post-editted text and the original translation. The tool used for calculating both WER and PWER was the freely available `apertium-eval-translator`.[12]

The results of this evaluation are shown in table 3.

5.3 Comparison

The only other publicly available Czech to Polish system we are aware of is Google Translate. We did not perform a full contrastive analysis, as in Tyers and Nordfalk (2009), as we did not feel the results would be particularly instructive. For the puposes of a simple comparison, we translated the same test sets using Google Translate. The results of the comparison are shown in table 4.

The disparity in the results is somewhat surprising. It is possible that the UDHR was part of the training set used by Google; it is also possible that Google Translate simply performs better on material that has been translated from English, either directly, through the use of pivot languages (Kumar et al., 2007), or indirectly, through the use of corpora such as JRC Acquis (Ralf et al., 2006) where both Polish and Czech had been translated from a third language.

Although we did not have access to the system mentioned in Dębowski et al. (2002), we found it instructive to compare the performance of our system with theirs, using the two example sentences they provide. The results of the comparison are shown in table 5.

In the first sentence, we have almost the same output, with one difference that can be attributed to an error in their Polish generator. In the second sentence, we have the correct auxiliary verb, though incorrectly inflected, as agreement is being taken from the wrong chunk. We also have doubly incorrect output in the last noun chunk (*pracujący użytkownika*) – firstly, agreement has not been performed; secondly, the singular has been generated instead of the plural, as the result of tagging errors.

6 Shortcomings of the system

Although the high number of unknowns accounts for many of the errors encountered (words immediately following unknown words are treated as having begun the sentence, and are capitalised), most of the errors can be attributed to tagging errors. Discounting unknowns, and non-errors such as different choices of synonyms, 58% of the remaining errors can be attributed directly to mistagging: 24% were due to selecting incorrect morphological forms, 18% due to part of speech errors, and 16% were due to incorrect agreements made with a word with an incorrect morphological form.

Another relatively common source of errors (12%) were word order differences in noun phrases. Adjectives usually preceed the noun in both Czech and Polish, but in certain fixed multiwords in Polish, the adjective follows the noun (normally describing an inherent property). A

[12] Available from Apertium SVN, for details see `http://www.apertium.org/`.

Corpus	WER	PWER	Free rides	Unknowns
News Samples	71 %	60 %	5 %	28 %
UDHR	88 %	68 %	0 %	22 %

Table 3: Evaluation results for the assimilation task. Free rides are those words which are identical in both the source and target language. Thus although they do not cause a degradation in translation quality, it is relevant to take them into account when evaluating the system. Unknown words are included as an indication of naïve coverage over the test sets.

Czech	Požadavky starší třiceti dnů se mažou.
Dębowski	Żądania starszy trzydziestu dzieni się smarują.
Apertium	Żądania starsze trzydziestu dni się smarują.
Corrected	Żądania starsze niż trzydzieści dni są wymazywane.
Czech	Počet dialogových procesů by měl pokrývat pracující uživatele.
Dębowski	Ilość dialogowych procesów miałby pokrywać pracujących użytkowników.
Apertium	Ilość dialogowych procesów powinien pokrywać pracujący użytkownika.
Corrected	Ilość procesów dialogowych powinna pokrywać ilość pracujących użytkowników.

Table 5: A comparison of Dębowski and our system.

separate project to add multiword processing to Apertium was begun at the same time as this work, but did not yield any useful results.

Chunking errors were not a significant source of errors in the test sets we used, though we anticipate that they will be a frequent source of errors. Apertium's chunker uses the left-to-right, longest match strategy in pattern matching, which can lead to over-matching. To address this, we extended the chunking module, to allow backoff to the previous longest match when an exception to the rule currently being processed has been detected, but the change is still being tested,[13] and rules to make use of this backoff mechanism have yet to be written.

Dębowski et al. (2002) mentions the problem of differences with T-V distinction; that is, Polish uses the words *pan*, *pani*, *państwo* (sir, madam, "people") to express polite sentiments, while Czech uses *vy* (you, plural). We made no attempt to address this.

7 Conclusion

We have presented results from the first free-software translator from Czech to Polish. This is also the first translator between free-word order languages developed for the Apertium platform. Although the current results are far from impressive, the problem with disambiguation is not an unsolveable one. As work has already begun on other, similar language pairs, we believe the solution to this problem will become more compelling.

For future work, there are further subcategories of words from which cognates could be induced, as well as further sources of statistical data to be processed. As we made no attempt to use morphological analysis or stemming to maximise the available data, we believe our sources can yield further vocabulary.

8 Acknowledgements

Development was funded as part of the Google Summer of Code[14] programme. We wish to acknowledge the debt we owe to the authors of Tyers and Nordfalk (2009), which we used as a template for the project, and for this article. We also wish to acknowledge the contributions of Petr Homola and Francis Tyers, for their input and contributions to the development of the project; Jernej Vičič, for providing the Czech–Slovenian material; and Marcin Miłkowski, for discussions on the processing of Slavic languages. We also wish to thank the anonymous reviewers for their useful comments.

[13] Available from Apertium's SVN: `https://apertium.svn.sourceforge.net/svnroot/apertium/branches/apertium/jimregan`

[14] `http://code.google.com/soc/`

References

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source Portuguese–Spanish machine translation. In *Computational Processing of the Portuguese Language, Proc. PROPOR 2006*, volume 3960 of *LNCS*, pages 50–59. Springer-Verlag.

Bański, P. and Wójtowicz, B. (2010). Open-Content Text Corpus for African languages. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 9–14.

Bielec, D. (1998). *Polish: An Essential Grammar*. Routledge.

Dębowski, L., Hajič, J., and Kuboň, V. (2002). Testing the limits–adding a new language to an MT system. *The Prague Bulletin of Mathematical Linguistics*, 78:95–102.

Futrega, M. (2010). Słownik SJP.pl. `http://www.sjp.pl/`.

Ginestí-Rosell, M., Ramírez-Sánchez, G., Ortiz-Rojas, S., Tyers, F. M., and Forcada, M. L. (2009). Development of a free Basque to Spanish machine translation system. *Procesamiento del Lenguaje Natural*, (43):187–195.

Hajič, J. (2004). *Disambiguation of Rich Inflection–Computational Morphology of Czech.* Charles University–The Karolinum Press, Prague.

Hajič, J., Homola, P., and Kuboň, V. (2003). A simple multilingual Machine Translation system. In *Proceedings of the MT Summit IX*.

Hajič, J., Kuboň, V., and Hric., J. (2000). Machine translation of very close languages. In *6th ANLP Conference / 1st NAACL Meeting.*, pages 7–12.

Jagodziński, G. (2008). A Grammar of the Polish Language. `http://grzegorj.w.interia.pl/gram/en/gram00.html`.

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. *ACL Demonstration Session*.

Kumar, S., Och, F., and Macherey, W. (2007). Improving word alignment with bridge languages. In *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Miłkowski, M. (2010). Developing an open-source, rule-based proofreading tool. *Softw., Pract. Exper.*, 40(7):543–566.

Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, 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 (LREC'2006).*, Genoa, Italy.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent Advances in Natural Language Processing (vol V).*, pages 237–248, Amsterdam/Philadelphia.

Tyers, F. M. and Nordfalk, J. (2009). Shallow-transfer rule-based machine translation for Swedish to Danish. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 27–33, Alicante.

Tyers, F. M. and Pienaar, J. (2008). Extracting bilingual word pairs from Wikipedia. *Proceedings of the SALTMIL Workshop at the Language Resources and Evaluation Conference, LREC08*, pages 19–22.

Unhammer, K. and Trosterud, T. (2009). Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42, Alicante.

Weiss, D. (2005). A survey of freely available Polish stemmers and evaluation of their applicability in Information Retrieval. In *Proceedings of the 2nd Language and Technology Conference*, pages 216–221, Poznań, Poland.

Wikipedie (2010). Polština — Wikipedie: Otevřená encyklopedie. [Online; navštíveno 20. 11. 2010].