Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 73-80 recibido 20-11-2015 revisado15-01-2016 aceptado 03-02-2016

Tectogrammar-based machine translation for English-Spanish and English-Basque

Traducción automática basada en tectogramática para inglés-español e inglés-euskara

Nora Aranberri, Gorka Labaka, Oneka Jauregi, Arantza Díaz de Ilarraza, Iñaki Alegria, Eneko Agirre
IXA Group
University of the Basque Country UPV/EHU
Paseo de Manuel Lardizabal, 1, 20018 Donostia-San Sebastián
{nora.aranberri, gorka.labaka, ojauregi002, a.diazdeilarraza, i.alegria, e.agirre}@ehu.eus

Resumen: Presentamos los primeros sistemas de traducción automática para inglés-español e inglés-euskara basados en tectogramática. A partir del modelo ya existente inglés-checo, describimos las herramientas para el análisis y síntesis, y los recursos para la trasferencia. La evaluación muestra el potencial de estos sistemas para adaptarse a nuevas lenguas y dominios.
Palabras clave: traducción automática, tectogramática, inglés, español, euskara

Abstract: We present the first attempt to build machine translation systems for the English-Spanish and English-Basque language pairs following the tectogrammar approach. Based on the English-Czech system, we describe the language-specific tools added in the analysis and synthesis steps, and the resources for bilingual transfer. Evaluation shows the potential of these systems for new languages and domains.
Keywords: machine translation, tectogrammar, English, Spanish, Basque

1 Introduction

Phrase-based machine translation (MT) systems prevail in the MT sphere. For minority languages with limited resources, however, they are far from providing quality translations, and these languages tend to look for rule-based alternatives. For languages such as English and Spanish, which have vast quantities of resources, statistical systems produce quality translations, but even in such cases, they often fail to capture linguistic phenomena such as long-distance grammatical cohesion, and domain adaptation is also a challenge.

Syntax-based systems are an alternative to tackle these limitations. While similar languages go for shallow approaches (Brandt et al., 2011), dissimilar language-pairs go deeper (Aranberri et al., 2015). The abstractions of deeper systems aim to strip off language-dependent attributes while preserving their meaning, making abstractions more comparable between languages. Then, a synthesis step provides the correct surface form for each language.

TectoMT (Popel and Žabokrtský, 2010) is an architecture to develop such an approach. It is based on tectogrammar (Hajičová, 2000), which represents language as deep syntactic dependency trees. Transfer works at tecto-level representations, in contrast to other dependency systems such as Matxin (Mayor et al., 2011), which uses transfer to synchronize language-dependent differences. Alternatively to Matxin, TectoMT combines linguistic knowledge encoded in rules, and statistical techniques.

The work presented here is carried out in the context of the QTLeap project[1], which targets a question-and-answer (Q&A) scenario in the information technology (IT) domain. We aim to test if TectoMT can improve state-of-the-art SMT systems for this domain with a relatively low effort.

We have developed a TectoMT system for both directions of English-Spanish (henceforth en-es, es-en) and English-Basque

[1] http://qtleap.eu

ISSN 1135-5948

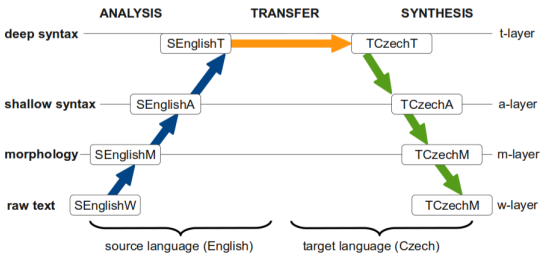

Figure 1: The general TectoMT architecture (from Popel and Žabokrtský (2010:298)).

(henceforth en-eu, eu-en), based on the existing English-Czech TectoMT system.[2] Due to project requirements, we have mainly focused on translation from English. Specifically, we distributed our effort as en-es 50%, en-eu 25%, es-en 15% and eu-en 10%. We estimate a total effort of 12 person/months for the current systems.

The article is structured as follows. In Section 2 we give an overview of the TectoMT architecture and the key linguistic concepts it is based on; in Section 3 we specify the work done to add new language pairs; in Section 4 we show the evaluation of the new prototypes; and finally, in Section 5 we draw some conclusions.

2 *The TectoMT Translation System*

As most rule-based systems, TectoMT consists of analysis, transfer and synthesis stages. It works on different levels of abstraction up to the tectogrammatical level (cf. Figure 1) and uses *blocks* and *scenarios* to process the information across the architecture (see below).

2.1 Tecto layers

TectoMT works on an stratified approach to language, that is, it defines four layers in increasing level of abstraction: raw text (w-layer), morphological layer (m-layer), shallow-syntax layer (a-layer), and deep-syntax layer (t-layer). This strategy is adopted from the Functional Generative Description theory (Sgall, 1967), further elaborated and implemented in the Prague Dependency Treebank (PDT) (Hajič et al., 2006). As explained by Popel and Žabokrtský (2010:296), each layer contains the following representations (see Figure 2):

Morphological layer (m-layer) Each sentence is tokenized and tokens are annotated with a lemma and morphological tag, e.g. *did: do-VBD*.

Analytical layer (a-layer) Each sentence is represented as a shallow-syntax dependency tree (a-tree), with a 1-to-1 correspondence between m-layer tokens and a-layer nodes. Each a-node is annotated with the type of dependency relation to its governing node, e.g. *did* is a dependent of *tell (VB)* with a *AuxV* relation type.

Tectogrammatical layer (t-layer) Each sentence is represented as a deep-syntax dependency tree (t-tree) where lexical words are represented as t-layer nodes, and the meaning conveyed by function words (auxiliary verbs, prepositions and subordinating conjunctions, etc.) is represented in t-node attributes, e.g. *did* is no longer a separate node but part of the lexical verb-node *tell*. The most important attributes of t-nodes are:

tectogrammatical lemma;

functor the semantic value of syntactic dependency relations, e.g. actor, effect, causal adjuncts;

grammatemes semantically oriented counterparts of morphological categories at the highest level of abstraction, e.g. tense, number, verb modality, negation;

formeme the morphosyntactic form of a t-node in the surface sentence. The set of formeme values depends on its semantic part of speech, e.g. noun as subject (n:subj), noun as direct object (n:obj), noun within a prepositional phrase (n:*in*+X) (Dušek et al., 2012).

2.2 TectoMT

TectoMT is integrated in Treex,[3] a modular open-source NLP framework. Blocks are independent components of sequential steps into which NLP tasks can be decomposed. Each block has a well-defined input/output specification and, usually, a linguistically interpretable functionality. Blocks are reusable and can be listed as part of different task sequences. We call these *scenarios*.

TectoMT includes over 1,000 blocks; approximately 224 English-specific blocks,

[2] http://83.240.145.199/WizardQTleap/pilot2

[3] https://ufal.mff.cuni.cz/treex
https://github.com/ufal/treex

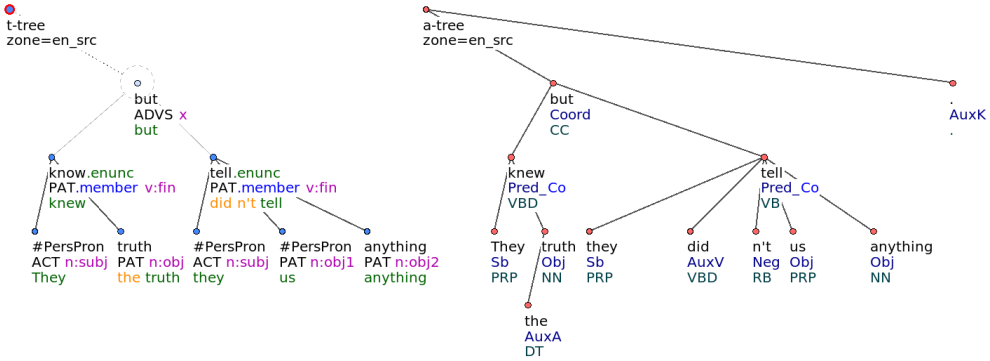

Figure 2: a-level and t-level English analysis of the sentence "They knew the truth but they didn't tell us anything."

237 for Czech, over 57 for English-to-Czech transfer, 129 for other languages and 467 language-independent blocks.[4] Blocks vary in length, as they can consist of a few lines of code or tackle complex linguistic phenomena.

3 *Developing new language pairs*

We set to port the TectoMT system to work for the en-es and en-eu language pairs in both directions. The modules for the English-Czech and Czech-English pairs are divided into language-specific and language independent blocks, thus facilitating the work for new language pairs. As we will see in what follows, a good number of resources were reused, mainly those setting the general architecture and those specific to English; others were adapted, mainly those involving training of new language and translation models; and several new blocks were created to enable language-pair-specific features.

Because the original system covered both directions for the English-Czech pair, English analysis and synthesis were ready to use. Therefore, our work mainly focused on Spanish and Basque analysis and synthesis, and on the transfer stages. In the following subsections we describe the work done on each stage, analysis, transfer and synthesis, for each translation direction.

3.1 Analysis

The analysis stage aims at getting raw input text and analyzing it up to the tectogrammatical level so that transfer can be performed (see Figure 2). The modules needed for English required little effort as they were already developed and running.

For Spanish and Basque, however, new analysis tools were integrated into Treex. For tokenization and sentence splitting, we adapted the modules in Treex. These are based on non-breaking prefixes, and thus required adding a list of Spanish and Basque non-breaking prefixes.

For the remaining tasks, we opted for the `ixa-pipes tools`[5] (Agerri, Bermudez, and Rigau, 2014). These tools consist of a set of modules that perform linguistic analysis from tokenization to parsing, as well as several external tools that have been adapted to interact with them. Our systems include lemmatization and POS tagging (`ixa-pipe-pos` and `ixa-pipe-pos-eu`), and dependency parsing (`ixa-pipe-srl` and `ixa-pipe-dep-eu`).

The tools were already developed, with accurate models for Spanish and Basque. Our efforts focused on their integration within Treex. We used wrapper blocks that, given a set of already tokenized sentences, create the input in the corresponding format and call the relevant tool. Once the tools complete their work, their output is read and loaded in Treex documents.

The analyses generated by the `ixa-pipes tools` follow the AnCora guidelines for Spanish and the Basque Dependency Treebank guidelines for Basque for both morphological tags and dependency tree structures. These mostly equate to the a-layer in the TectoMT stratification but, to fully integrate the analyses into Treex and generate the expected a-tree, the analyses have to be mapped to a universal PoS and dependency tagset. TectoMT currently uses the Interset tagset (Zeman, 2008)

[4] Statistics taken from: `https://github.com/ufal/treex.git(27/08/2015)`

[5] `http://ixa2.si.ehu.es/ixa-pipes/`

and HamleDT guidelines (Zeman et al., 2014). On top of this, and in order to form the t-tree, we used 23 and 22 additional blocks for Spanish and Basque analyses, respectively:

Language-independent blocks Both analyses reuse a similar set of language-independent blocks already available in Treex with 14 blocks for Spanish and an additional tokenization block for Basque. These mainly re-arrange nodes, mark heads (coordinations, clauses, coreference) and set node types.

Adapted blocks We adapted 7 blocks for Spanish and 6 for Basque out of blocks originally used for English or Czech analysis. These include blocks to mark edges and collapse a-nodes into a single t-node, or blocks to annotate function words, sentence mood and grammateme values.

New language-specific blocks We wrote 3 specific blocks for Spanish and 2 for Basque to set the grammatemes and formeme values of t-nodes based on the a-node attributes of function words.

3.2 Transfer

TectoMT's transfer approach assumes that t-tree structures in different languages are shared. Although this is not always true (Popel, 2009), it allows to model translation as a 1-to-1 t-node transfer. The transfer stage combines separate statistical dictionaries for t-lemma and formeme equivalences and a set of manually written rules to address grammateme transfer (Žabokrtský, 2010).

t-lemma and formeme equivalences are obtained by first analyzing parallel corpora (cf. Section 4) up to the t-level in both languages. Next, for each t-lemma and formeme in a source t-tree, we define a dictionary entry and assign a score to all possible translations observed in the training data. This score is a probability estimate of the translation equivalent given the source t-lemma, formeme, and additional contextual information. It is calculated as a linear combination of two main translation models (TM):

Discriminative TM (Mareček, Popel, and Žabokrtský, 2010) It is a set of maximum entropy (MaxEnt) models (Berger, Della Pietra, and Della Pietra, 1996) trained for each specific source t-lemma and formeme, where the prediction is based on features extracted from the source tree (Crouse, Nowak, and Baraniuk, 1998).

Static TM It is a bilingual dictionary that contains a list of possible translation equivalents based on relative frequencies and no contextual features.

The final score assigned to each t-lemma and formeme in the statistical dictionaries is calculated through interpolation. Interpolation weights were defined after a manual optimization. For the t-lemmas, weights of 0.5 and 1 were assigned to the static TM and the discriminative TM, respectively. In the case of formemes, the values were reversed. Using these two TMs, we obtain a weighted n-best list of translation equivalences for each t-lemma and each formeme.

Grammatemes contain linguistically more abstract information, e.g. tense and number, which is usually paralleled in the target language. The grammateme values are assigned by manually written rules which, by default, copy the source values to the target t-nodes. A set of relatively simple exception rules is sufficient to address language-pair-specific differences. So far we have defined exceptions in the systems translating from English, 4 blocks for the en-es direction and another 4 blocks for the en-eu direction. These address the lack of gender in English nouns (necessary in Spanish), differences in definiteness and articles, differences in structures such as *There is...* and relative clauses.

Domain adaptation Lexical adaptation efforts have been done at transfer level for the IT domain. Firstly, we created a new t-lemma dictionary based on the Microsoft Terminology Collection. This collection is freely available[6] and contains 22,475 Spanish entries and 5,845 Basque entries. Secondly, we trained additional discriminative and static TMs using a development corpus of 1,000 IT Q&A pairs (cf. Section 4). These new in-domain models were combined with the generic TMs through interpolation to update the statistical dictionaries. t-lemma equivalents in the terminology collection are given priority over the statistical dictionaries.

[6] `http://www.microsoft.com/Language/en-US/Terminology.aspx`

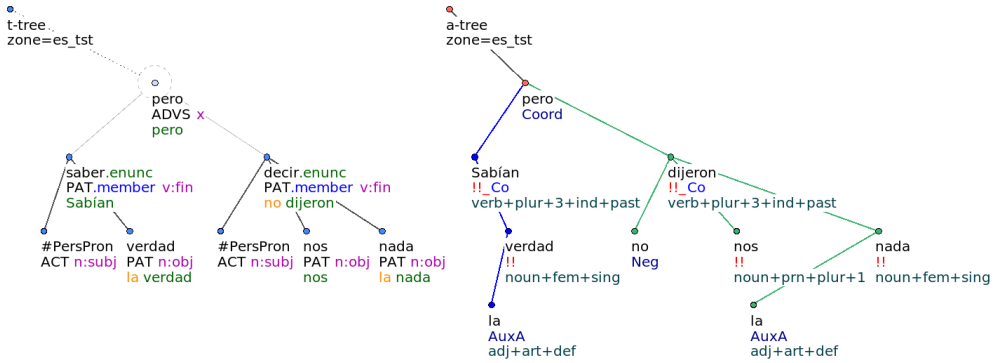

Figure 3: a-level and t-level Spanish synthesis for the final translation `*"Sabían la verdad pero no dijeron nos la nada."`

3.3 Synthesis

Transfer outputs a t-tree. Synthesis first generates its corresponding a-tree and then creates the w-tree, which contains the final surface forms (see Figure 3). This stage was already developed for English and therefore, once again, our work mainly focused on Spanish and Basque.

On a first step, we use a total of 22 blocks for Spanish and 17 blocks for Basque to transform the t-tree into the a-tree.

Language-independent blocks 10 of the blocks applied for Spanish and 11 for Basque were reused from the language-independent set already available in Treex. Among these are blocks to impose subject-predicate and attribute agreements, add separate negation nodes, add specific punctuation for coordinate clauses, or impose capitalization at the beginning of sentence.

Adapted block For Spanish, 9 blocks were adapted from the language-independent, English and Czech synthesis blocks. For example, when creating the a-tree, the morphological categories are filled with values derived from the grammatemes and formemes, and Spanish requires more specific information than that coming from English. This is the case of the imperfect tense (a subcategory of past tense) and imperfect aspect, for instance, which we set on a block. The same issue arises in articles. The definiteness of a noun phrase is not sufficient to decide whether to generate a determiner in the target language. Another example is that of personal pronouns. We remove personal pronoun nodes when acting as subject as pro-drop languages such as Spanish do not require that they appear explicitly because this information is already marked in the verb. For Basque, 6 blocks were adapted. These blocks are responsible for inflectional information needed in Basque, for generating the appropriate verb tenses from the grammatemes or for dropping the object when it is not explicitly needed.

New language-specific blocks 4 blocks were written from scratch to deal with Spanish-specific features. These deal with attribute order, comparatives and verb tenses. Attribute order refers to the position of adjectives with respect to the element they modify. In English, adjectives occur before the noun, but this is the opposite in Spanish, with some exceptions for figurative effect. The block addressing comparatives creates additional nodes for the Spanish structure, which is specially relevant for the cases where no separate comparative word is used in English. Finally, a block was specifically written to address the complex verb tenses in Spanish. This block uses the information about tense, perfectiveness and progressiveness of the English verb to select the appropriate verb form in Spanish. For Basque one specific block to deal with word order differences was written.

Overall, we see that most blocks are used (i) to fill in morphological attributes that will be needed in the second step, (ii) to add function words where necessary, (iii) to remove superfluous nodes, and (iv) to add punctuation nodes.

On a second step, the lemma and morphosyntactic information on the a-tree must be turned into word forms to generate the w-tree. We used Flect (Dušek and Jurčíček, 2013) to do this, by training new models for Spanish and Basque. Flect is

a statistical morphological generation tool that learns inflection patterns from corpora. We trained the tool with subsets of the parallel corpus used to learn the transfer dictionaries (cf. Section 4): a subset of morphologically annotated Europarl corpus (530K tokens) for Spanish and a subset of the translation memories provided by Elhuyar (540K tokens) for Basque. The tool automatically learns how to generate inflected word forms from lemmas and morphological features. Flect allows us to inflect previously unseen words, as it uses lemma suffixes as features and induces edit scripts that describe the difference between lemma and word-form.

On a third step, once we obtain the w-tree with the word forms, a number of blocks are written to polish the final output. For Spanish, for example, we use a block to concatenate the prepositions *a* and *de* with the masculine singular article *el*, which should be presented as the single forms *a+el* → *al* and *de+el* → *del*. For Basque language-independent blocks are reused.

4 Evaluation

We evaluated the all four new TectoMT prototypes in three different scenarios: (i) using language-independent blocks only[7], (ii) adding the blocks written and adapted for Spanish and Basque, and (iii) activating lexical adaption. The transfer components of the English-Spanish prototypes were trained on Europarl (∼2 million sentences). The Basque prototypes were trained on translation memory data containing academic books, software manuals and user interface strings (∼1.1 million sentences), and web-crawled data (∼0.1 million sentences) made available by Elhuyar Fundazioa (see Table 1).[8]

Also, we evaluated the new TectoMT systems against phrased-based statistical systems. These systems were trained on the same corpora used for the TectoMT prototypes. To this end, we built four SMT systems, one per language-pair and direction.

For Spanish, we used tools available in the Moses toolkit for tokenization and truecasing, while mGiza was used for word alignment. For language modeling, we used SRILM to train the language model. We used the target side of the bilingual corpus to train the language models.

For Basque, the systems used language-specific preprocessing tools for tokenization and lowercasing and, in addition, we performed lemmatization. In particular, Stanford CoreNLP was used for the English side and Eustagger for Basque. The length threshold for filtering the training sentences was adjusted to a maximum of 75 words per sentence in order to meet the language-specific length properties. Word alignment was performed using mGiza based on lemmas, which was then projected to lowercased full word-forms for the rest of the training process. After translation, a recasing process was performed based on the tool available in Moses. Note that for the language model, we added the Basque text of Spanish-Basque translation memories of administrative texts (∼7.4 million sentences) to the Basque text of the English-Basque parallel data used in the TectoMT systems.

Our evaluation focuses on a Q&A scenario in the IT domain. Therefore, for tuning, we used a development set of 1,000 in-domain interactions (question-answer pairs) -same set used in the lexical adaptation of the TectoMT systems. The original interactions were in English and they were translated into Spanish and Basque by human translators. We calculated BLEU scores for the systems on a held-out test-set of 1,000 interactions (see Table 1).

We can draw several conclusions from the BLEU scores. First, we observe that the TectoMT prototype beats the statistical system for the en-es system evaluated on the IT test-set (8 points ahead of the baseline).

Because a large portion of the TectoMT systems is based on manual rules, the lower scores of the Basque prototypes was to be expected, given the lower effort put at this stage of development. In addition to this, the scores for the Basque statistical systems are more difficult to beat because a section of their training corpus is in-domain data.

The scores also reflect the difference in development for the TectoMT systems in terms of language direction. As mentioned, priority was given to the en-es system and it is this system that has the highest score.

With regard to the TectoMT systems,

[7] This setup includes `ixa-pipes tools` and Flect models for Spanish and Basque analysis and synthesis, and bilingual transfer dictionaries.

[8] Elhuyar: https://www.elhuyar.eus/en

	English-Spanish	Spanish-English	English-Basque	Basque-English
Moses	16.23	27.53	18.59	11.94
(i) TectoMT – language independent blocks	6.29	10.24	8.20	3.41
(ii) TectoMT – + target language blocks	13.65	15.66	9.16	6.62
(iii) TectoMT – + lexical adaptation	24.32	18.64	10.83	6.79

Table 1: BLEU scores for the English-Spanish and English-Basque TectoMT prototypes

we observe how the BLEU scores increase as we customize the system. The systems with only language-independent blocks score lower than the systems that include language-specific blocks. For the en-es system, BLEU scores almost double. es-en scores also increase although not as much. When activating the lexical adaptation resources, we observe that the BLEU scores increase almost 3 points for the en-es direction and over 1 point for the es-en direction.

In addition to the automatic metrics, we performed a manual error analysis for the best-scoring en-es and en-eu TectoMT systems. Annotators marked 25 sentences using a selection of issue types taken from the Multidimensional Quality Metrics framework[9]. Table 2 summarizes the number of errors annotated per upper-level category. We see that Fluency errors are the most frequent. These include grammatical errors, with function words being the most problematic in both languages.

Error type	English-Spanish	English-Basque
Accuracy	13	10
Fluency	42	72
Terminology	12	15

Table 2: Error type frequencies

A qualitative analysis shows that the improvements of the TectoMT systems over the statistical approach come from better domain adaptation of the former, both in terms of lexical and syntactic coverage. The Q&A test set used for evaluation contains many imperative verbs, distinctive of this domain, which are hardly present in the parallel corpora used for statistical training, but typically included in the verb-type range of the syntax-based approaches. Based on the results of the MQM analysis, it is clear that our priority for the near future is to continue enriching the systems with more sophisticated grammar blocks and, in particular, a better treatment of function words.

[9] `http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics`

5 Conclusions

In this paper we have shown the work done to develop entry-level deep-syntax systems for the English-Spanish and English-Basque language pairs following the tectogrammar MT approach. Thanks to previous work done for the English-Czech pair in the TectoMT system, we have reused most of the English analysis and synthesis modules, and mainly focused on the integration of tools and the development of models and blocks for Spanish and Basque.

In particular, we have integrated the `ixa-pipes tools` for PoS and dependency parsing, and adapted their output to comply with the tecto-level representation of language, which uses universal labels. For transfer, we have trained new statistical models for all four translation directions. For synthesis, we have trained a new morphological model to obtain Spanish and Basque word forms. Substantial effort was also put on writing sets of blocks to address differing linguistic features between the language pairs across all stages.

The es-en system includes 61 reused blocks and 9 new/adapted blocks; the en-es uses 71 reused blocks and 17 new/adapted blocks; the eu-en system has 63 and 8, respectively; and the en-eu 74 and 11. The systems are open-source and they can be downloaded from `https://github.com/ufal/treex`. The evaluation has shown that with some effort, the TectoMT prototypes can surpass the statistical baselines, as it is demonstrated by the en-es system in a domain-specific scenario. Also, we observed that the TectoMT architecture offers flexible customization options. We have shown that the BLEU scores increase considerably as these are integrated and tuned to the working language pair.

Acknowledgements

We thank the anonymous reviewers for their helpful suggestions. Elhuyar Fundazioa is also kindly acknowledged for generously providing us with the en-eu corpus.The research leading to these results has received funding from FP7-ICT-2013-10-610516 (QTLeap project, `qtleap.eu`).

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Conference on Language Resources and Evaluation*, Reykjavik.

Aranberri, N., G. Labaka, A. Díaz de Ilarraza, and K. Sarasola. 2015. Exploiting portability to build an RBMT prototype for a new source language. In *Proceedings of EAMT 2015, Antalya*.

Berger, A., V. Della Pietra, and S. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Brandt, M., H. Loftsson, H. Sigurthórsson, and F. Tyers. 2011. Apertium-icenlp: A rule-based Icelandic to English machine translation system. In *Proceedings of EAMT 2011, Leuven, Belgium*.

Crouse, M., R. Nowak, and R. Baraniuk. 1998. Wavelet-based statistical signal processing using hidden markov models. *Signal Processing, IEEE Transactions*, 46(4):886–902.

Dušek, O. and F. Jurčíček. 2013. Robust multilingual statistical morphological generation models. *ACL 2013*, page 158.

Dušek, O., Z. Žabokrtský, M. Popel, M. Majliš, M. Novák, and D. Mareček. 2012. Formemes in English-Czech deep syntactic MT. In *Proceedings of WMT7*, pages 267–274.

Hajič, J., J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štepánek, J. Havelka, M. Mikulová, Z. Žabokrtský, and M. Ševcıková Razımová. 2006. Prague dependency treebank 2.0. *CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia*, 98.

Hajičová, E. 2000. Dependency-based underlying-structure tagging of a very large Czech corpus. *TAL. Traitement automatique des langues*, 41(1):57–78.

Mareček, D., M. Popel, and Z. Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proceedings of WMT5 and MetricsMATR*, pages 201–206. ACL.

Mayor, A., I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, and K. Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine translation*, 25(1):53–82.

Popel, M. 2009. Ways to improve the quality of English-Czech machine translation. *Master's thesis, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic*.

Popel, M. and Z. Žabokrtský. 2010. TectoMT: modular NLP framework. In *Advances in natural language processing*. Springer, pages 293–304.

Sgall, P. 1967. Functional sentence perspective in a generative description. *Prague studies in mathematical linguistics*, 2(203-225).

Žabokrtský, Z. 2010. From treebanking to machine translation. *Habilitation thesis, Charles University, Prague, Czech Republic*.

Zeman, D. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of LREC*, pages 213–218.

Zeman, D., O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.