An extraction of a Polish-Basque dictionary from parallel corpus

Justyna Pietrzak

Euskal Herriko Unibertsitatea/Universidad del País Vasco
(The University of the Basque Country)
justyna.o.pietrzak@gmail.com

"Hizkuntzaren azterketa eta prozesamendua" masterra
("Analysis and Processing of Language" postgraduate course)

September 2009

Abstract

This project carries out the automatic, statistics-based extraction of Polish-Basque dictionary from a parallel corpus. The dictionary obtained in this way complements the existing conventional Basque-Polish dictionary, compiled manually by the team of lexicographers lead by the author of the present work. The objective is to get reasonable results while using only "off-the-shelf", free (or free for research purposes) software tools at every stage of the process and employing as input source a bilingual, relatively small (433,393 and 424,192 tokens at Basque and Polish part, respectively), untagged, parallel corpus, aligned at a paragraph level. The obtained output is the probabilistic macro-structure of the future dictionary, comprising a list of potential Polish headwords with their translation equivalent(s). The general workflow is drawn and tools used at every stage of the work are enumerated, as well as all necessary pre-process of the raw data. As the effectiveness of the process depends mainly on the accurate word alignment, the related works carried out in this field are examined, with special attention paid to projects for languages of rich morphology and applied to small corpora. The obtained as output Polish-Basque probabilistic dictionary is considered to be half-finished product, the basis for the full-fledged dictionary, to be revised and enhanced afterwards by lexicographers. The interface was prepared to allow following the work on-line. Finally, the possibility of extending input data and improvement of the dictionary-extracting process are also dealt with.

1 Introduction

In recent years, as observed by many scholars (e.g. Wilks 2003, McEnery 2003, Piasecki 2007), the empirical methods are gaining ground in the field of linguistics. Especially the automatic analysis of big corpora is used in constructing of models of language based on empirical data. What is more, we can also observed the soaring popularity of automatic construction of linguistic resources, collecting of corpuses from the web, constructing of grammars, lexicons and dictionaries complement the work performed till now manually by linguists.

In particular, the usage of parallel corpora opens up the possibilities of creation of new recourses. Extraction of lexical data by means of word alignment (WA) from bitext (parallel corpora) is broadly used in many fields of computational linguistics, such as machine translation, cross-language information retrieval, and also bilingual thesaurus or dictionary creation, which is of interest in this project.

Most part of word alignment systems was developed and tested in languages of analytic type of morphology, mainly English. The methodology adopted while working with languages of high grade of morphological synthesis (like Polish and Basque) should be different in order to cope with data dispersion, uncommon in English. For that reason any attempt of efficient WA for languages of this type will be especially challenging.

1.1 The goal and scope of the work

The main goal of this project is the semi-automatic construction of Polish-Basque dictionary based on WA of parallel corpora, investigating in this way the possibility of getting reasonable results while using only "off-the-shelf", free (or free for research purposes) software tools at every stage of the process. The corpus employed as input is a bilingual, relatively small (433,393 and 424,192 tokens at Basque and Polish part, respectively), untagged, parallel corpus, aligned at a paragraph level. More precisely, the project aims towards two targets: (i) construction of a dictionary that could be of provisional and tentative, yet valuable help for students of Basque and Polish and (ii) building a base for further work of lexicographers, which will boost the Polish-Basque dictionary creation. These apparently consecutive goals can be realized simultaneously, operating on the same database, after uploading it on Internet, to be managed at the same time by two different interfaces.

The output of this work is the probabilistic macro-structure of the dictionary, comprising a list of Polish potential headwords with their Basque translation equivalent(s), where the degree of correspondence will be expressed in terms of alignment probabilities. This Polish-Basque probabilistic dictionary is considered to be half-finished product, the basis for the full-fledged dictionary, revised and enhanced by linguists.

The Internet interface for the dictionary was also prepared, extending the existing Basque-Polish dictionary interface. This new tool now offers two new modes of searching: a possibility of inverted search in the existing dictionary and an access to new database with the probabilistic dictionary. This part of interface, as a starting-point for the revised, final version of the dictionary, has the option of text editing, accessible only for authorized users.

1.2 General overview of the workflow

The general workflow and the tools used at every stage are as follows: (i) sentence segmentation of *Lagun*[1] corpus performed according to grammar constructed by the author, based on observation of a corpus and described as a string of regular expression transformations, (ii) morphological analysis of Polish (reduction of word-forms to lemmas and grammatical categories) carried out with *Morfeusz*, (iii) disambiguation of Polish lemmas with *TaKIPI*. Lemmatization was an especially important step, as in languages of rich morphology (as are Polish and Basque) the high ratio *lemma: word-form* causes the data dispersion. (iv) Morphological analysis of Basque (reduction of word-forms to lemmas), performed with *Hunspell* and suitable Basque dictionaries. The next step (v) was word alignment. For WA task I selected *GIZA++*, which can be used free of charge. In fact, *GIZA++* is widely used in various projects, and can be considered "state-of-the-art" in the field of WA software. Resulting word pairs list was purged (vi), eliminating pairs of law probability and low token frequency in the corpus Some intents of measuring the effectiveness of the WA are done afterwards (vii). Accuracy is calculated automatically using an existing Basque-Polish dictionary and some rough manual estimation. Finally (viii), the translation equivalents list was exported to MySQL database and incorporated into online interface of unidirectional Basque-Polish dictionary, which converts itself in bidirectional (ix).

1.3 Paper organization overview

The paper is structured as follows: as the effectiveness of the project depends mainly on the accurate word alignment, the analysis of related works carried out in this field is presented in *Section 2*, with special attention paid to languages of rich morphology and applied to small corpora.

In *Section 3* I briefly present the existing Basque-Polish dictionary, which will be complemented with the dictionary complied in this project. Description of morphological characteristic of Polish and Basque is given in *Section 4*. In *Section 5* various computational tools for morphological analysis are presented, with special attention paid to those used in the project. *Section 6* introduces some information about the input data, provided by parallel corpus used. *Section 7* describes step by step the whole process of dictionary creation. *Section 7.1* deals with the first step of pre-processing of the data, namely sentence alignment. *Section 7.2* treats of the tokenization of raw data. The morphological analysis of Polish and Basque is presented in *Section 7.3.1* and *Section 7.3.2*, respectively. Some differences in morphological treatment of Basque and Polish are mention in *Section 7.3.3*. Word alignment process is described in *Section 8* and *Section 9* presents its results. The Internet interfaces are briefly described in *Section 10*, and, finally, *Section 11* concludes with discussion of the necessary post-processing of the output and the possibilities of improvement, e.g. by extending the input data.

2 Related works

A common intuition underlying automatic extraction of translation equivalents from bitexts is simple: words that are translations of each other are more likely to

[1] Biographic references of every tool are given in corresponding sections.

appear in corresponding bitext sentences then other pairs of words (Melamed 2000). The intuition is followed by two different approaches to the task of statistics-based WA in bilingual corpora: the *hypotheses testing approach* and the *estimating approach* (Hiemstra 1997, Tiedemann 2003). The first one (called also "association approach" or "heuristic approach") relies on a generative device that produces a list of translation equivalence candidates, each of them being subjected to an independent statistical test. Estimation approaches ("statistical alignment") use probabilistic translation models and build a statistical bitext model which allows for global maximization of the translation equivalence relation, considering not individual translation equivalents but sets of such equivalents.

Many competitions (e.g. HLT-NAACL (Mihalcea & Pedersen, 2003)) and projects (e.g. ARCADE I and II (Chiao *et al.* 2006)) organized in the field of WA have demonstrated that the estimation approaches give better results with big corpora, and that some additional linguistic information (like, e.g. POS tagging) will efficiently increase the accuracy of the WA systems when small corpora are available, especially when using cascade systems, with various types of algorithms, data pre-processing, and combing heuristics (Tufiş *et al.* 2004, Han 2001) .

The scarceness of corpora, as proved by various authors (e.g. Al-Onaizan *et al.* 2000, Niessen & Ney 2004) is not necessarily the insurmountable obstacle. In (Al-Onaizan *et al.* 2000) human decoders were asked to align a small (about 1000 sentence) Tetun-English corpus, no one of them knowing Tetun. The alignment was quite a success, which conducted the authors to analyze the strategies taken by humans and to try to reproduce the results while using the machine learning algorithms based on human strategies.

Another approach, although similar to some extent, is taken in (Niessen & Ney 2004). The authors proposed methods of incorporating morphological and syntactic information into systems for statistical machine translation. They constructed hierarchical lexicon models on the basis of equivalence classes of words and introduced sentence-level restructuring transformations. Finally, they were able to reduce the amount of bilingual training data to less than 10% of the original corpus, while losing only 1.6% in translation quality.

There are some interesting projects that use *GIZA++* to align Slavonic languages, e.g. Czech-English (Bojar & Prokopová 2006), and Serbian-English with a small corpus (Popović *et al*. 2005). *GIZA++* was also used with Basque-Spanish bitext (Agirre *et al*. 2006). All of these works experimented with morphosyntactic tagging of data at a pre-process level, and obtained considerable results. They proved the necessity of lemmatization, and experimented with tagging lemmas with other types of morphologic and syntactic information. For Czech and Basque, specialized lemmatizer/taggers were used. In (Bojar & Prokopová 2006) corpus of approx. 20,000 sentences (404,000 tokens without punctuation marks) were used, a quantity similar to the *Lagun* corpus. The authors proved the crucial role of lemmatization of Czech, and pointed out that 38% of erroneous word alignments made by *GIZA++* were difficult to human experts as well. The results obtained were 75% recall (15.0 AER) for intersection and 89.8% recall (17.2 AER) for union of the dictionaries created by *GIZA++*.

In (Popović *et al*. 2005) even a smaller corpus of about 3,000 sentences (20,000 tokens) was used. Employing so small corpus allowed performing the lemmatization and POS tagging semi-manually, getting results quite similar to Czech ones.

In the field of dictionary extraction the main task is not to force the full alignment, that is, alignment of all occurrences of lexical tokens to their translations.

What is expected is the constructing of translational equivalence at a level of lemmatized word-forms. Great part of the morphosyntactic information can be lost during the lemmatization, if it is no longer necessary in WA process.

A project reported in (Gómez & Sacau 2004), made for Galician-English parallel corpus, was elaborated with yet another WA tools – *NATools*. The authors demonstrated that it could be of great interest for the statistics-based dictionary extraction. The corpus used in that project is twice the size of *Lagun*, and various experiments are made with morphosyntactic tagging, which slightly improved the final results. Some important post-process was proposed: automatic filters that starting with the output of *NATools* generate bilingual dictionary, eliminating alignments of poor quality. The filters combine the lemma frequency (>4) with the probability of its translation (≥0.3 but ≠0.5) giving precision = 91.4% and recall = 54.7% for base corpus and precision = 93.9% and recall = 50.2% for corpus tagged at morphosyntactic level.

3. Basque-Polish dictionary

The existing unidirectional Basque-Polish dictionary, announced in (Pietrzak 2002), is available at www.baskijski.net. It is the first Basque-Polish dictionary of that size, going beyond the simple word-list, which can be found on internet. At present it has *c.* 30,000 Basque headwords, with respective 10,000 Polish translation equivalents. The dictionary has been made by a team of lexicographers, and it has taken 10 years to reach the present, acceptable but not yet polished, state. Only the headword list was obtained automatically by compiling various publicly available Basque dictionaries and creating the frequency list based on available corpora. Taking it into account it becomes quite obvious that applying the same method when compiling Polish-Basque dictionary has poor chances of yielding satisfactory result within reasonable period of time. What makes the situation even worse is the fact that this kind of dictionary has scanty number of potential users, hence null possibilities of being commercial product.

Nevertheless, there exists a small niche audience that use the dictionary on regular bases. The course of Basque language started at Adam Mickiewicz University (Poznań, Poland) in 1992. Since then about 200 Polish students has had a chance to study Basque, about 20 of them reaching the EGA (proficiency) level. In the last years similar courses were organized on an occasional basis in other Polish universities, usually in the departments of Spanish language (in Warsaw, Cracow and Wrocław). On the other hand the course of Polish at the Basque Country University started in 2000. All these students are potential or already real users of *baskijski.net* dictionary. For this reason it appears interesting and useful the following up with the work and the creation of Polish-Basque dictionary.

An idea, appealing at the first sight, of simple inverting the existing dictionary has to be rejected for various reasons. First, the Polish translations use to have more general sense that their Basque equivalents. In particular, all Basque headwords tagged as “dialectal”, “vulgar”, “obsolete”, etc. have been translated with the more popular Polish equivalents of neutral sense. For that reason many Polish lemmas are simultaneous equivalents of series of their Basque counterparts. For example the word “uderzenie” (*impact*) appears as much as 48 times as a translation of different Basque words: “danbateko”, “pultsazio”, “kolpe”, “makilada”, “makilazo”, “kolpeka”, “joaldi”, “topeka”, “inarrosaldi”, “danbada”, “kiska”, “sastako”, “zaplateko”, “palu”, “kaska”, “kaskako”, “tanpa”, “brastada”, “zafra”, “zaztada”, “zirti-zarta”, “zaflako”,

“zanpateko”, “zartateko”, “eskukaldi”, “zartako”, “takateko”, “dangada”, “ipurdikada”, “ola”, “dangateko”, “zaplazteko”, “ukaldi”, “zaflada”, “ostia”, “danga”, “kroska”, “jo”, “jotze”, “zafla”, “panp”, “zaplada”, “bizkarreko”, “zart”, “zartada”, “zartada”, “koska” and “ukabilko”.

There are also a numerous group of Basque multi-word units, that do not appear as headwords, but their Polish equivalents should appear in the dictionary (e.g. “janaria prestatu” is not a headword in the Basque-Polish dictionary, which will prevent from appearing as a headword a Polish word “gotować” (*to cook*).

In general, the list of Polish translations existing in the dictionary is not equal to the possible list of Polish headwords, which is caused for example by difference in the treatment of morphological derivation in Basque and Polish (e.g. “indarka”: “wysilając się” (*forcing*), “wysilając się” is an inflected form of verb “wysilać się” (to *force*), not a headword, and should not appear in the dictionary as such).

Although, even rejecting inverting of the dictionary, it is still possible to perform inverted (Polish->Basque) searches inside Basque-Polish dictionary with quite simple computational tools. Such a solution has been implemented in *baskijski.net*, but it should still be considered a half measure and one keeps longing for a full-fledged Polish-Basque dictionary.

4. Morphological description of Polish and Basque

All the tasks in the WA field are highly language-dependent. Questions like word order, inflectional paradigms, morphological complexity, alphabets used, etc. are of vital importance for the final success. Although all regular relation between lexical units can be systematized, even without any overt linguistic reflection, detailed analysis of morphological paradigms of languages in question can help in anticipating WA problems and suggest some pre-processing tasks, which will facilitate the alignment. In this section only the general sketches of Polish and Basque morphology are given. Some detailed question concerning morphology will be treated in the subsequent sections as well.

4.1. Polish language

Polish is highly inflectional language, with seven cases for nouns, adjectives and pronouns, with (depending on the classification) tens declension paradigms. High grade of inflection means that one morphological ending cumulates various category meanings. For example, Polish nouns decline by number (2) and case (7) and adjectives by number (2), case (7), grade (3) and gender (3/9). Combination of all attributes of these categories produces a considerable theoretical number of unique morphological affixes, which in practice is reduced, as a result of a substantial syncretism of endings (homography among inflectional forms of the same lexeme, i.e. one ending is an exponent of more then one combination of morphological categories).

Polish gender is a lexical category (which lacks universal formal exponents, but inflects by different paradigms), with three classes: masculine, feminine and neuter. Nevertheless, in the morphosyntactic analysis as much as 9 gender classes can be identified, governing different verb paradigms and adjective-noun agreement. This is so because two other categories enter the gender system: personhood (personal *vs.* non-personal) and animacy (animate *vs.* inanimate). The exact number of genders detected by morphological analyser depends strictly on the underlying linguistic theory.

Polish verbs conjugate according to person, number, gender and tense. Three moods and three voices can be also formally marked. Polish aspect, although apparently derivated on regular bases, has very complex systems of prefixes, infixes and suffixes, and can be treated as lexical (derivational) or grammatical category, in accordance with the NLP task. Verbs, depending on aspect affixes, can take the same verbal paradigm with a present tense or a future tense meaning. There are only one tense (compound future) that use auxiliary verb forms (based on "być" (*to be*) weak verb), hence is inflected in analytic way. All other tense paradigms use only synthetic verbal forms. Verb agrees only with a subject of a sentence. Other inflected parts of the speech are numerals and pronouns. Adverbs, propositions, conjunctions and particles are uninflected.

At the same time Polish language abounds with relicts of grammatical forms, no longer productive, e.g. dual number, declension patterns proper for pronouns, not infrequent *pluraria tantum* and *singularia tantum*, etc., which form huge number of exceptions for any rule that could be formulated for Polish grammar.

Word roots often undergo morphophonological changes, producing complete or partial suppletion. In extreme cases two forms of the same lexeme's root hardly share a single letter, e.g. "ćma" (ćm-a, *nominativus:sg:*moth) but "ciem" (ciem-Ø, *genetivus:pl:*moth). All these morphophonological changes are historically motivated and can be foreseen to certain, yet not fail-safe, extent.

Inflexion in Polish is predominantly suffix-based. The exceptions are e.g.: the superior grade of adjectives, which can be formed by preffix "naj", like in: wysoki-wyższy-najwyższy (high – higher - highest) and the negation prefixes.

All these characteristics have to be coped with by any efficient morphological analyser of Polish.

4.2 Basque language

Basque is an agglutinative language of rich but quite regular morphology. It builds word-forms by combining in lineal order all morphemes which convey unique morphologic, syntactic or (in few cases) pragmatic meanings.

A Basque noun phrase can be inflected in 17 different cases, modified additionally by the category of number and definiteness. The generative power of such a system is enhance by two exponents of genitive case, which enable the all system to be used recursively, in theory *ad infinitum*. Contrary to Polish, in Basque there exists only one paradigm of declension.

The vast majority of verbs is inflected in a periphrastic (analytic) way, with main and auxiliary verb. The main verb codifies aspect, created on regular basis. Three aspectual categories are: perfect, habitual and potential. The auxiliary verb represents other morphological categories and syntactic relations. Basque has multiple verb agreement: the auxiliary verb agrees with subject, direct and indirect object (if they exist). In addition, auxiliary verbs can be inflected by tense and mood. Subject drop is frequently observed (as well as in Polish). There exists a small set of about 15 verbs that sometimes can take synthetic inflection paradigm and not to use auxiliary verbs. Some relics of other synthetic verbs exist, with fragmentary paradigms, or being a part of fossilised idioms.

"Hika", informal second person singular allocutive verb forms, different for masculine and feminine interlocutor, adds a pragmatics component to the morphological analysis.

Few highly regular and predictable morphophonological changes take place on morpheme borders.

Inflexion in Basque is predominantly suffix-based, although some productive prefixes also exist.

5 Morphological analysis: available tools

Polish as well as Basque are languages that count with variable and modern tools used in NLP. In this section I scrutinize all mayor computer programs performing a morphological analysis of languages in question. The special attention is paid to these available online, either on free software license or free for non-commercial use, which have already been proven in different projects. The tools used in this project are presented in more detailed way.

5.1 Polish language: morphological analysers

We owe the first computational description of Polish to Jan Tokarski, who in 1951 wrote about teaching computer the inflection (Tokarski 1951). During years he was collecting Polish endings and lemmas together with the rules governing their combinations. To speak more precisely, endings in the mean of Tokarski were not morphological endings, but strings of letters subjected to modification in morphological process of inflexion. Tokarski did not finish his work, which was undertaken by Saloni, and published in (Tokarski 1993). The book provides information on virtually all possible endings of Polish words, and the way they combine with lemmas (base forms), constructing the algorithm of automatic morphological analysis of Polish. First experiments with the use of Tokarski's data in practice revealed the need of the incorporation of the base of lemmas without which the overgeneration of interpretations caused the usage of Tokarski's data ineffective.

Many of the now existing morphological analysers of Polish are based on Tokarski's index. Nowadays there are about 10 lemmatizers/taggers for Polish language, at least 3 of which are free (some restricted to scientific research) and can be used in this project: *SAM* (Szafran 1997), *Morfologik* (based on *ispell* engine (Weiss 2005)) and *Morfeusz* (Woliński 2006).

SAM (Sistem of the Morphological Analysys) by Krzysztof Szafran (Szafran 1993, Szafran 1996) was the first morphological analyser based on Tokarski's data which carries out the *a tergo* (from behind) analysis. The program was prepared as a PhD project of the author and later on extended into SAM95. SAM uses rules of Tokarski's Index, and applies Doroszewski dictionary (120,000 lemmas) to prune overgeneration. It is capable to guess and analyze unknown word-forms venturing a probable lemma.

*Morfologik***,** which started as a stand-alone stemmer called *Lametyzator* (Weiss 2005), has undergone transformation and incorporation of different tools (e.g. *Stempler* –another algorithmic stemmer, converting itself in a hybrid stemmer) and is now an open source project of morphological analysis that can be used as a spell checker. In fact, it is now a morphological analyser connected with morphological dictionary and spell-checker. It uses the inflection rules provided with a Polish dictionary of *ispell*. These rules describe how to convert a base form of a term (normally equal to lemma) into a set of inflected word-forms. In *ispell* this information is used mainly for spell checking, but *Morfologik* applies it for generating mappings *token:lemma.* It is also used as a spell checker module e.g. in OpenOffice or Mozilla Firefox.

Morfeusz, presented in (Woliński 2006) is yet another morphological analyser based on Tokarski's index of morphological endings combined with *Doroszewski* dictionary of Polish. As stated by the authors (Woliński 2006) the analyser recognises 95.7% of running words and 69% of word types of the IPI PAN Corpus (almost 85 millions of words). *Morfeusz* provides lemmatized forms of given tokens, as well as their morphosyntactic description. In case of ambiguity, the result of morphological analysis includes all possible interpretations, that is to say, it does not perform the contextual disambiguation. As a stand-alone program, it has no capability to guess unknown forms. In 2006 the *Morfeusz* dictionary consisted of about 4,750,000 word-forms, which provides for recognising about 1,700,000 different Polish lemmas.

The tagset used by *Morfeusz* is the IPI PAN Tagset (Przepiórkowski&Woliński 2003) which combines morphologic and morphosyntactic tagging. It groups the tags in 32 grammatical classes and add 12 grammatical categories (attributes of a given class). Grammatical classes are the new classification of concepts traditionally expressed as parts of the speech. In IPI PAN Tagset the grammatical classes are morphosyntacticly motivated, strongly disjunctive sets of lexemes, including punctuation marks and unrecognized items classes. The categories are: number (with 2 possible values: singular or plural), case (7), gender (9), person (3), grade (3), aspect (2), negation (2), accentability (2), post-prepositionality (2), accommodability (2), agglutination (2), vocability (2). The last five categories add pragmatics and morphophonology to traditional morphosyntactic categories.

There exists one available tool for Polish that performs disambiguation: *TaKIPI* ("Tager Korpusu IPI PAN", *Tagger of the IPI PAN corpus*). It needs a text of already analysed tokens to perform disambiguation, and can be combined with any morphological analyser which uses the IPI PAN Tagset. *TaKIPI* is an example of the praiseworthy tendency present in last years, under the philosophy of cooperation between different programmers and scientist who decided to create programs with free software licence (GNU GPL).

TaKIPI tagger, using *Morfeusz* as a base for morphosyntactic description, performs contextual disambiguation with 86% effectiveness (Piasecki 2007). It is based on a little set of hand-made rules and some thousands of rules obtained automatically. *TaKIPI* uses the subprogram *Odgadywacz* which allows the analyzing of unknown word-forms.

After some preliminary trials and manual comparison of the results, I decided to use *Morfeusz*[2], as a best suited for the project and offering the best results. The desambiguator *TaKIPI*[3]*,* based on *Morfeusz* data format*,* was also used.

5.2 Basque language: morphological analysers

There exist a few morphological analysers that can be used for Basque, e.g. *Eustagger* (Aduriz & Díaz de Ilarraza 2003) that can be used under licence. *Eustagger* produces morphosyntactic description and lemmatization. It was used e.g. in (Agirre *et al.* 2006) where it proved its capacity to perform an intricate tagging and lemmatization when performing an efficient alignment of Spanish and Basque bitext. Nevertheless, there exists only one freely

[2] The program can be downloaded from: http://nlp.ipipan.waw.pl/~wolinski/morfeusz/ Morfeusz is available free of charge for non-commercial use and scientific research, nevertheless it is not free software.

[3] Available on GNU GPL licence, at: http://www.plwordnet.pwr.wroc.pl/g419/tagger/

available morphological analyser - *Hunspell*[4]. *Hunspell* is an open source spell checking, stemming, morphological analysis and generation tool available under GPL licenses. *Hunspell* combined with appropriate Basque dictionaries[5] can served as a perfect morphological analyser, capable to perform lemmatization of a given text. It uses two dictionaries, the first one (*.dic) contains a list of words for the language. Each word may optionally be followed by a slash and one or more flags, which represent affixes or special attributes. The second dictionary (*.aff) defines the meaning of special flags.

5.3 Word alignment tools

Among some free tools available to perform WA I initially selected three, taking into account their availability and performance in other projects. There are: *PWA* (Tiedeman 2003), *NATools* (Hiemstra 1998, Simões & Almeia 2003) and *GIZA++* (Och & Ney 2000, 2003),.

PLUG Word Aligner (PWA) comprises two word alignment systems, the Linköping Word Aligner (LWA) and the Uppsala Word Aligner (UWA), which were first used to align Swedish-English texts, but can be used for any other language pairs. The system integrates a set of modules for knowledge-lite approaches to word alignment, with various possibilities of changing configuration and adapting the system to other language pairs and text types. The system requires sentence aligned bitext as its input and produces a list of word and phrase correspondences in the text (token links) and an additional bilingual lexicon from these instances (word-form links). Few project used that software, so there is not much information about its performance with different languages, but the versatility of the tools incorporated into the toolkit is promising.

NATools is specially designed to create bilingual dictionaries using statistical methods. It includes a sentence aligner, a probabilistic translation dictionary extractor, a word aligner and a set of other tools to study the aligned parallel corpora. The aligner tool is based on Hiemstra's *Twente* aligner (Hiemstra 1998). The alignment process creates two dictionaries, mapping words from one language to a set of words in the other language. This set includes for each translation its probability of being a correct translation. Both dictionaries can be used independently. This additional data can be of special help at a post-process level, when a lexicographer will have to group potential translations of one word into different homonyms or different sense of one entry.

GIZA++ is an open source implementation of the IBM word alignment models. It is now the most broadly used tool for WA, which makes it easier to compare different alignment results. *GIZA++* is used twice to obtain alignments in both directions. This makes it especially useful for dictionary extraction, getting links of high precision while using an intersection of both dictionaries, and better recall when using the union. The latter, although it gives less probable translation equivalents, can be very useful for the post-process work of lexicographers, when decision has to be made as to e.g. grouping potential translations of the word into different homonyms or make different sense of one entry.

---

[4] Available at: http://hunspell.sourceforge.net/

[5] Basque dictionaries to be used with *Hunspell* available at: http://www.euskara.euskadi.net/r59-20660/eu/contenidos/informacion/euskarazko_softwarea/eu_9567/xuxen.html

6 LAGUN Corpus

A corpus which will be used in this project is the only one, publicly available (and it wouldn't be risky to state: the only existing one), parallel corpus containing Basque and Polish texts. *Lagun*[6] is aligned at a paragraph level. As for now, there are 16,280 paragraphs at each side. It includes also English, French and Spanish texts, partially aligned (not every source text is represented in all 5 languages), but they will not be used in here.

It has 433,393 and 424,192 tokens in Basque and Polish part, respectively (including punctuation and numeric tokens). The data comes from 19 different texts, mostly XX century literature. There are also present in the corpus, to less extent, some juridical and scientific texts, e.g. "Treaty establishing a Constitution for Europe" and other European Union official documents or "On the Origin of Species" by Ch. Darwin. The Bible (Old and New Testament) is also represented.

Almost in its totality the texts are translations from languages others than Basque and Polish, which may cause some extra problems of data noisiness. The only exception is the collection of Basque short stories translated and edited in Polish in 2000.

There are 84,521 word-forms (types) in Basque part of the corpus (compared to 83,414 Polish ones). Taking into account Zipf's law, it is not surprising that more than a half of word-forms has the frequency = 1, which makes the statistic alignment quite difficult. There are 58,876 such word-forms in Basque and 54,860 in Polish. The difference may be due to the considerable syncretism of Polish declension paradigms.

Version of the corpus used in this project is dated for December 2008, and was kindly provided to the author of the present work to be used locally.

7 Dictionary creation

Basque and Polish, being both highly inflectional languages with significant differences in word order and morphology, are expected to be hard to align. Therefore I concentrate on methods of increasing the WA precision by some pre-process: alignment at a sentence level, tokenization and lemmatization.

7.1 Pre-process: sentence alignment

The sentence alignment is of crucial importance for word alignment, due to the fact that the latter can be considered as a refinement of the former. As a *Lagun* corpus is aligned at a paragraph level, the first step is to divide the paragraphs into aligned sentences. The sentence segmentation was performed according to grammar constructed by the author, based on observation of a corpus and described as a string of regular expression transformations. This primitive alignment was obtained by dividing the paragraphs using as sentence separators punctuation marks (".", "!", "?" or "...") followed by a capital letter). This general rule was later refined by some language-specific procedures. For example, ordinal numbers in Basque are followed by period sign, which is not always a case in Polish.

The segmentation produced almost 97% of theoretical alignment (estimated by manual checking of 100 randomly chosen paragraphs), close to the today state-of-the-

[6] *Lagun* corpus at: http://korpus.hiztegia.org.

art (Véronis 2000), so, after manually scrutinizing that alignments are correct in the waste part, I decided not to use any specialized tools for sentence alignment. Any further analysis pointed towards refining the alignment would need of language-specific tools, performing analyses at syntactic or/and semantic levels, tools which are not freely available by now.

The paragraphs in which the number of sentences in Polish did not correspond with Basque are left intact. All the operation were carried out with PERL scripts giving (from the initial 16,280 paragraphs) 23,637 aligned items (sentences and paragraphs). *Table 1* shows one of the paragraphs divided into aligned sentences.

Paper-mutur batean idatzita neraman helbidea eta paretaren kontra ikusi nuen kartel urdina, rue Mouffetard.	Na kawałku papieru miałem zapisany adres, a na murze dostrzegłem niebieski szyld, ulica Mouffetard.
Hura zen.	To tu.
Zenbakia baieztatu eta, gehiagorik pentsatu gabe, pentsioko txirrina jo nuen.	Sprawdziłem numer i nie zastanawiając się dłużej, nacisnąłem dzwonek.

Table 1. Example of aligned sentences inside one paragraph.

7.2 Pre-process: tokenization

No tokenization was carried out prior to morphological analysis after testing that the tools, which will be used afterwards, performe kind of rough tokenization, which consists of separation of punctuation marks. Additionally, the authors of Polish analyser *Morfeusz* declare that the program is capable of recognize period mark being the sentence separator from ones used in abbreviations, based on incorporated abbreviation dictionary.

The only exception was made to hyphen and dash marks, which were used in very inconsistent way along the corpus, and the morphological analysers was incapable of their correct separation and analysis. Hyphen and dash usage was corrected by regular expression transformations.

The possible alignment at an unlemmatized token level is presented in *Fig. 1*. Though, as has been already said, I do not search for total alignment at token level, this first approximation shows the existing problem of multi-word units.

Fig. 1. WA at a token level. Tokens marked in red do not have their direct translation in this sentence. Group of tokens underlined are multi-token units, which in the full alignment should be treated as one.

It is important here to pay some attention to this problematic issue in tokenization: the treatment given to multi-word units (MWU). The rule that treats white spaces as a token separation is a simplification, which leaves multi-words out of analysis. A problem especially important in Basque, which abounds in verbal multi-words constructed with weak verbs, e.g. “egin” (*to make*): “lo egin” (*to sleep*), “lan egin” (*to work*), etc.). Polish does not use weak verbs in derivation, and translation equivalents of the above examples should be uni-word lexemes: “spać” and “pracować”.

On the other hand, Polish reflexive aspect is expressed by independent particle “się” (*itself*, similar to Spanish “se” in “reirse”). So the expected headword will be a multi-word unit “śmiać się” (*to laugh*). Although prototypically particle “się” immediately follows the inflected verb, it can also precede it. What is more, the verb and the particle can be separated by other tokens, which make the usage of syntactic parser indispensable if one wants to tokenize verb and corresponding particle “się” into one token. Compare the sentences, slightly different in style:

(1) Zrobiło się już bardzo późno. (It had grown very late.)
(2) Już się bardzo późno zrobiło. (It had grown very late.)

Polish dictionaries practically lack of other multi-word units with the status of headword, and the reflexive aspect is normally conceptually close to non-reflexive one (if exists). That is to say, pair of reflexive and non-reflexive verbs can share the headword, with the only difference being aspectual meaning. Nevertheless, the opposite situation is also frequently the case: reflexive headword and its apparently unreflexive counterpart differ in more than aspectual meaning and need separated headword entries. The particle “się”, apart from reflexive meaning can take a reciprocal meaning as well:

(3) Lubimy się. (We like each other.)
(4) Lubimy się. (We like ourselves.)

All these cases have to be disambiguated in some way. From the 10,208 appearances of the word-form “się”, only about 7,080 could be tokenized together with preceding main verb. In order to not increase data dispersion I decided not to tokenize the corpus in this way.

7.3 Pre-process: morphological analysis

At this point, viewing the sparseness of used data, it turns out to be quite obvious that the lemmatization is an indispensable step. Firstly, because the inflectional (Polish) and agglutinative (Basque) morphologies cause the very high *word-form: lemma* ratio, and consequently great number of word-forms has frequency = 1, which makes effective alignment nearly impossible. Secondly, lemmatization is a necessary step because I am searching for dictionary headwords, which normally are equal to lemmas. As stated in (Bojar & Prokopová 2006) lemmatization of Czech (relatively similar morphologically to Polish) can reduce alignment error to a half.

Lemmatization was conducted with specialized programs, different for each language.

7.3.1 Morphological analysis of Polish

After some preliminary trials I decided to use *Morfeusz* for morphological analysis of Polish text. Apart from slightly better performance, the deciding factor was its integration with the only available desambiguator for Polish: *TaKIPI*. By default it works with *Morfeusz*, and it is how *TaKIPI* was used in this project.

As shown in *Table 2*, every analysed token is put in one line with coma separating token, proposed lemma, and its morphosyntactic analysis. The pipe ("|") mark separates possible interpretation of the same lemma, and semicolon (";") separates possible interpretation belonging to different lemmas. The "?" mark is used to tag unrecognized items.

[To,ten,adj:sg:nom.acc:n1.n2:pos; To,to,conj; To,to,pred; To,to,qub; To,to,subst:sg:nom.acc:n2]
[tu,tu,qub]
[.,.,interp]
[Sprawdził,sprawdzić,praet:sg:m1.m2.m3:perf]
[em,być,aglt:sg:pri:imperf:wok]
[numer,numer,subst:sg:nom.acc:m3]
[i,i,conj]
[nie,nie,qub; nie,on,ppron3:sg:acc:n1.n2:ter:_:praep|ppron3:pl:acc:m2.m3.f.n1.n2.p2.p3:ter:_:praep]
[zastanawiając,zastanawiać,pcon:imperf]
[się,się,qub]
[dłużej,długo,adv:comp; dłużej,dłużeć,impt:sg:sec:imperf]
[nacisnął,nacisnąć,praet:sg:m1.m2.m3:perf]
[em,być,aglt:sg:pri:imperf:wok]
[dzwonek,dzwonek,subst:sg:nom.acc:m3; dzwonek,dzwonko,subst:pl:gen:n1.n2]
[.,.,interp]

Table 2. Fragment of the output of morphological analysis performed by Morfeusz Original text is as follows: "To tu. Sprawdziłem numer i nie zastanawiając się dłużej, nacisnąłem dzwonek." (*That was it. I checked the number and without further ado, rang the bell*.)

Morfeusz does not recognize multi-word units (lexemes), and white spaces always separate different tokens. On the other hand, it differentiates multi-lexeme words, that is to say, it recognizes words compounded by two or more lexemes (according to methodology used by the authors). That allows, among others, to cope with the problem of free inflection endings (historically weak forms of the verb "być" (*to be*) which can stick not only to verbs, but to virtually any other part of speech). According to rules applied by *Morfeusz* agglutinative past verb forms are split from the main verb as separated tokens. For example one lexem "sprawdziłem" (*I checked*), being the form of the first person singular, past tense, is separated in two lemmas: sprawdzić (*to check*) and "być" (auxiliary verb *to be*) as can be seen in *Table 2*.

TaKIPI outputs the results in *xml* format, producing the morphosyntactic analysis of every token of the given text, performs sentence segmentation and indicates which is the most probable of all proposed interpretations. As the sentence segmentation was already done, and the Polish sentences were aligned with Basque ones, the *TaKIPI* segmentation was overridden in this project.

Table 3 presents fragment of an xml output file produced by *TaKIPI*. As can be seen, the program marks with tag <lex disamb="1"> the most probable interpretation between all possible interpretations detected by morphological analysis (tagged by <lex>).

```
<tok>
<orth>To</orth>
<lex><base>to</base><ctag>subst:sg:nom:n</ctag></lex>
<lex disamb="1"><base>to</base><ctag>subst:sg:acc:n</ctag></lex>
<lex><base>ten</base><ctag>adj:sg:nom:n:pos</ctag></lex>
<lex><base>ten</base><ctag>adj:sg:acc:n:pos</ctag></lex>
<lex><base>to</base><ctag>pred</ctag></lex>
<lex><base>to</base><ctag>conj</ctag></lex>
<lex><base>to</base><ctag>qub</ctag></lex>
</tok>
<tok>
<orth>tu</orth>
<lex disamb="1"><base>tu</base><ctag>qub</ctag></lex>
</tok>
<ns/>
<tok>
<orth>.</orth>
<lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
</tok>
<tok>
<orth>ForcedSentenceSeparator</orth>
<lex disamb="1"><base> ForcedSentenceSeparator </base><ctag>tsym</ctag></lex>
</tok>
<tok>
<orth>Sprawdził</orth>
<lex disamb="1"><base>sprawdzić</base><ctag>praet:sg:m1:perf</ctag></lex>
<lex><base>sprawdzić</base><ctag>praet:sg:m2:perf</ctag></lex>
<lex><base>sprawdzić</base><ctag>praet:sg:m3:perf</ctag></lex>
</tok>
<ns/>
<tok>
<orth>em</orth>
<lex disamb="1"><base>być</base><ctag>aglt:sg:pri:imperf:wok</ctag></lex>
</tok>
<tok>
<orth>numer</orth>
<lex><base>numer</base><ctag>subst:sg:nom:m3</ctag></lex>
<lex disamb="1"><base>numer</base><ctag>subst:sg:acc:m3</ctag></lex>
</tok>
<tok>
<orth>i</orth>
<lex disamb="1"><base>i</base><ctag>conj</ctag></lex>
</tok>
<tok>
```

```
<orth>nie</orth>
<lex><base>on</base><ctag>ppron3:sg:acc:n:ter:akc:praep</ctag></lex>
<lex><base>on</base><ctag>ppron3:sg:acc:n:ter:nakc:praep</ctag></lex>
<lex><base>on</base><ctag>ppron3:pl:acc:m2:ter:akc:praep</ctag></lex>
<lex><base>on</base><ctag>ppron3:pl:acc:m2:ter:nakc:praep</ctag></lex>
<lex><base>on</base><ctag>ppron3:pl:acc:m3:ter:akc:praep</ctag></lex>
<lex><base>on</base><ctag>ppron3:pl:acc:m3:ter:nakc:praep</ctag></lex>
<lex><base>on</base><ctag>ppron3:pl:acc:f:ter:akc:praep</ctag></lex>
<lex><base>on</base><ctag>ppron3:pl:acc:f:ter:nakc:praep</ctag></lex>
<lex disamb="1"><base>nie</base><ctag>qub</ctag></lex>
</tok>
<tok>
<orth>zastanawiając</orth>
<lex disamb="1"><base>zastanawiać</base><ctag>pcon:imperf</ctag></lex>
</tok>
<tok>
<orth>się</orth>
<lex disamb="1"><base>się</base><ctag>qub</ctag></lex>
</tok>
<tok>
<orth>dłużej</orth>
<lex disamb="1"><base>długo</base><ctag>adv:comp</ctag></lex>
<lex><base>dłużeć</base><ctag>impt:sg:sec:imperf</ctag></lex>
</tok>
<ns/>
<tok>
<orth>,</orth>
<lex disamb="1"><base>,</base><ctag>interp</ctag></lex>
</tok>
<tok>
<orth>nacisnął</orth>
<lex disamb="1"><base>nacisnąć</base><ctag>praet:sg:m1:perf</ctag></lex>
<lex><base>nacisnąć</base><ctag>praet:sg:m2:perf</ctag></lex>
<lex><base>nacisnąć</base><ctag>praet:sg:m3:perf</ctag></lex>
</tok>
<ns/>
<tok>
<orth>em</orth>
<lex disamb="1"><base>być</base><ctag>aglt:sg:pri:imperf:wok</ctag></lex>
</tok>
<tok>
<orth>dzwonek</orth>
<lex><base>dzwonko</base><ctag>subst:pl:gen:n</ctag></lex>
<lex><base>dzwonek</base><ctag>subst:sg:nom:m3</ctag></lex>
<lex disamb="1"><base>dzwonek</base><ctag>subst:sg:acc:m3</ctag></lex>
</tok>
```

Table 3. Fragment of the output of disambiguation performed by *TaKIPI*. Original sentence: “To tu. Sprawdziłem numer i nie zastanawiając się dłużej, nacisnąłem dzwonek.”

TaKIPI distinguished in *LAGUN* corpus 21,656 different lemmas, that is to say, possible candidates for being headwords of the Polish-Basque dictionary.

The corpus contains 60,807 Polish word-forms (excluding punctuation marks), which gives *lemma:word-form* density in corpus as low as 1:2 - 1:3. The morphological density in Polish is larger (*c.* 1:18) but in a small corpus it is improbable that lemmas reach saturation (compare: (Świdziński 2002), and similar data for Hungarian in: (Kornai 1992)). For example, the lemma "sprawdzić" (*to check* - "baieztatu" from the *fig. 1*) only in active inflection of verb for person and tense can take as much as 25 different word-forms. Meanwhile, it has no more then 5 attested word-forms in *Lagun* corpus, with 32 token representations.

Table 4 presents a fragment of the lemmatized Polish text.

Original:	*Lemmatized:*
Na kawałku papieru miałem zapisany adres, a na murze dostrzegłem niebieski szyld, ulica Mouffetard.	na kawałek papier miał zapisany adres a na mur dostrzec być niebieski szyld ulica mouffetard
To tu.	to tu
Sprawdziłem numer i nie zastanawiając się dłużej, nacisnąłem dzwonek.	sprawdzić być numer i nie zastanawiać się długo nacisnąć być dzwonek

Table 4. Output of morphological analysis and disambiguation performed by TaKIPI: lemmas

At this stage the first problems in the analysis can be detected, problems caused by the lemmatization methods or simple errors.

TaKIPI marked as unrecognized only 550 tokens, which can be gathered in 137 different word-forms. Almost all of them are proper names or foreign words. Only 4 real Polish words were marked as unrecognized. TaKIPI often failed in recognizing roman numerals and abbreviation (in contrary to what stated by its authors). Some geographic names, even those assimilated in Polish language and perfectly suited in Polish phonological system, were not recognized in some cases ("Paryż" (*Paris*), "Londyn" (*London*), "Dunaj" (*Danube*))

For the further analysis two new texts has been prepared, the first one containing only lemmas (see *Table 4*), the second containing new tokens compound by lemma and simplified version of morphologic information (*Table 5*). The morphological tagging is a simplification of the KIPI tagging system used by TaKIPI. The majority of grammatical classes, and some lexical attributes (grammatical categories: gender and aspect) are preserved, omitting all the information concerning grammatical information proper to token, not lemma.

Original:	*Lemmatized:*
Na kawałku papieru miałem zapisany adres, a na murze dostrzegłem niebieski szyld, ulica Mouffetard.	na@prep kawałek@subst:m papier@subst:m miał@subst:m zapisany@adj adres@subst:m a@conj na@prep mur@subst:m dostrzec@praet:perf być@aglt:imperf niebieski@adj szyld@subst:m ulica@subst:f mouffetard@subst:m
To tu.	to@subst:n tu@qub
Sprawdziłem numer i nie zastanawiając się dłużej, nacisnąłem dzwonek.	sprawdzić@praet:perf być@aglt:imperf numer@subst:m i@conj nie@qub

	zastanawiać@pcon:imperf się@qub długo@adv nacisnąć@praet:perf być@aglt:imperf dzwonek@subst:m

Table 5. Output of morphological analysis and disambiguation performed by TaKIPI: lemmas + morphological information

The morphological information was used afterwards as complement data accompanying Polish headwords.

7.3.2 Morphological analysis of Basque

The output of the analysis performed with *Hunspell* is the list of lemmas of all tokens of the given text. As can be seen in *Table 6* the first item of every line is a given token, and the second item is a proposed lemma (if lacking, we understand that *Hunspell* was not able to lemmatize the given word). Consecutive lines with the analysis of the same token are its various interpretations, as proposed by *Hunspell.*

```
Hura hura

zen. zen

ForcedSentenceSeparator

Zenbakia zenbaki

baieztatu baieztatu
baieztatu baiezta

eta eta

gehiagorik gehiago
gehiagorik gehi

pentsatu pentsatu
pentsatu pentsa

gabe gabe

pentsioko pentsio

txirrina txirrina
txirrina txirrin

jo jo

nuen. nuen
nuen. nu
```

Table 6. Fragment of the output of morphological analysis performed by *Hunspell* Original text is as follows: “Hura zen. Zenbakia baieztatu eta, gehiagorik pentsatu gabe, pentsioko txirrina jo nuen.”

Hunspell do not perform disambiguation, so, in the cases when *Hunspell* proposed various interpretations of the same token I arbitrary opted for choosing the first one from the list. *Table 7* contains the effect of the morphological analysis performed with *Hunspell* and combined into sentences, which can be aligned with their Polish counterparts.

Original:	*Lemmatized:*
Paper-mutur batean idatzita neraman helbidea eta paretaren kontra ikusi nuen kartel urdina, rue Mouffetard.	paper-mutur batean idatz neraman helbide eta pareta kontra ikusi nuen kartel urdin rue Mouffetard
Hura zen.	hura zen
Zenbakia baieztatu eta, gehiagorik pentsatu gabe, pentsioko txirrina jo nuen.	zenbaki baieztatu eta gehiago pentsatu gabe pentsio txirrina jo nuen

Table 7. Output of morphological analysis performed by *Hunspel*l, combined into sentences

As can be seen, *Hunspell* does not offer the lemma of auxiliary nor synthetic verbs, leaving them in the original form. Nevertheless, it should not cause serious problem in the task of word alignment, as Polish lacks auxiliary verbs of similar type. The only negative impact of that will be the increase of data noisiness.

The total of 26,177 word-forms (and at this stage of work word-forms equal to lemmas) were identified.

There were 11,920 tokens which *Hunspell* did not manage to lemmatize. So high number of analysis failures deserves some attention here. First, as *Hunspell* does not intend to lemmatize unknown tokens, many of them in fact form part of one single lemma, as is the case of inflected forms of the unrecognized names: “Emma”, “Emmak”, “Emmaren”, etc. All unknown words which do not exist in the dictionary used with *Hunspell* were marked as unrecognized and left intact. These comprise a long list of proper names, foreign words, and abbreviations. Polish analyser it this cases intended to assign the morphological category to unknown word, and if succeed, the word was marked as recognized, even for unknown lemma. Only when *TaKIPI* was not able to assign the category based on syntactic analysis of a sentence the token was marked as unrecognized. Obviously, there exists no reason that prevents the alignment of tokens marked as unrecognized in morphological analysis; nevertheless, the Basque unlemmatized counterparts of Polish lemmatized tokens have less possibility to be aligned correctly. The same can be said of all other unknown to *Hunspell*, hence marked as unrecognized, words.

Hunspell was not able to recognize reduplications, as in “poliki-poliki”, “punttu-punttuan”, “epel-epel”, nor hyphenated compounds, e.g.: “gizarte-arazoa”, “sufrimendu-mota”, “sardina-latak”. In fact, almost 4,000 of unrecognized tokens contain a hyphen.

The idiosyncratic way of the lemmatization of verbs performed by *Hunspell* caused that the verbs are not lemmatized to the perfect form, traditionally treated as headword in dictionaries, but to the root form. Half of the users of the dictionary, which by definition does not know well the Basque language, needs perfect verb form in order to correctly inflect and use it in a sentence.

7.3.3 Basque *versus* Polish morphological analysis: differences and similarities

This section intends to summarize potential limitation of word alignments between Polish and Basque, based on output of morphological analysis.

None of the analysers performed multi-word unit detection. As has been previously said, the only Polish multi-word units that could be headword candidates are the reflexive verbs, which co-occur with reflexive particle “się”. Given the relatively free position of particle in relation to main verb only syntactic parsing would help to tokenize these items as a unique one.

Lack of tokenization of multi-word in Basque will prevent a correct translation of pairs 1:n, like "uzgodnić” : “konforme jarri” (*to agree*), “odejść” : “alde egin” (*to go away*), “wyliniały” : “ile gabe(ko)” (*hairless*), “trzydzieści” : “hogeita hamar” (*thirty*), etc.

Hunspell has left intact all the auxiliary verbs, *Morfeusz*, on the other hand, split the majority of the past verb tokens into two parts: main verb and auxiliary verb “być” (*to be*). This methodology, though well-founded and justifiable in the field of NLP, contributed to increase of data dispersion, as a great number of unlemmatized Basque auxiliary verbs are confronted with an even higher number of a new tokens of very high frequency.

The syntactic verbs will also have no chance of proper alignment. For example verb “chodzić” (*to walk*) is not recognized as a translation equivalent of “ibili” (*to walk*), as only a part of appearances of lemma “chodzić” will co-occur with the lemma “ibili”, the rest having as translation equivalent lemmas “dabil”, “nabil”, etc.

Prepositions, being in Polish an independent part of the speech, which meaning in Basque is conveyed by declension, will be left orphaned as well, just like some particles and subordinate conjunctions.

8 Word alignment

Two decisions ware to be made: (i) which model proposed by *GIZA++* can give the best result (ii) what kind of filters should be applied to the obtained results. These decisions were based on some tests performed with every combination and later manual checking on a randomly sampled translation pairs.

From between all statistic alignment models offered by *GIZA++* Model 1 proved to be the most efficient one in our corpus, which can be due to its small size. Output of Model 4 was the one of poorer quality. It proposed a huge number of alignments with probability = 1 (that is, supposedly sure ones), which in the vast majority resulted incorrect. The differences between the remaining models were not significant, but Model 1, as offering the best alignments, was used as the final option. *GIZA++* was applied using its default parameters, with minor amendments, oriented toward getting the most of a small corpus (see *Appendix C* for details).

Some decision had to be made about the filters applied in the post-process. The goal was to eliminate alignments of poor quality, but without excessive pruning of the wordlist. Some tests were done and finally the filters combining the lemma frequency ($\geq$2) with the probability of its translation ($\geq$0.2) were applied, which left about 26% of the total number of lemmas present in the corpus. The filters can seemed too lenient, but they allowed removing rather improbable alignments, and leaving these of more chances of success. Lenient filters caused the loss in precision, which was compensated

in increase in recall. The decision was dictated by practical considerations: to make the future lexicographers' work easier.

9 Results

The result given by the WA alignment performed by *GIZA++* is a list of 5,732 lemmas of Polish language together with 6,432 translation equivalents. 2,102 of them are attested in Basque-Polish dictionary, hence can be considered to be sure alignments. The generous threshold filters were applied, but manual checking shows that approximately half of the proposed equivalents are correct. I consider these results as a totally satisfactory, as the created probabilistic dictionary is only a half-product, and does not pretend to be final and definitive. Boosting applied filters will considerably reduce the wordlist giving more trustworthy results, higher recall and precision. Nevertheless, the wordlist itself is a valuable by-product, necessary and essential in dictionary creation. *Table 10* presents the estimation of accuracy of results depending on frequency and probability estimated by *GIZA++*. In the estimation only pairs of real translation equivalents, i.e. valuable for a dictionary, are counted, excluding pairs correctly aligned but worthless for this project like: "yes" : "yes", "Hilter" : "Hilter", "III" – "III", etc. As can be seen boosting up of applied filters will result in dropping down of the recall, and drastic cut-off of the wordlist.

Table 8 and *Table 9* recapitulate the numeric characteristics of input and output data at various stage of the project. As has been said in the *Section 3*, although the idea of inverting the Basque-Polish dictionary is appealing at the first sight, especially viewing the disproportion between data available through automatic WA and inverse dictionary, the inversion can not assure reliable results, treated as a stand-alone resource.

	Polish:	*Basque:*
sentences	23,623	
tokens	424,192	433,393
word-forms	84,521	83,414
lemmatized tokens (without punctuation marks)	416,875	395,128
lemmas	21,658	26,177

Table 8. Corpus characteristics.

	Probabilistic dictionary:	*Existing in both*	*Inverse Basque-Polish dictionary*
Headwords (lemmas)	5,732	4,002	29,745
Translation pairs	6,432	2,102	57,730

Table 9. Characteristics of probabilistic dictionary (filtered) with comparison to inverse Basque-Polish dictionary.

probability:	***freq.:***	***num. of items***	***precision of alignment:***
>0.9	≥57	18	100%
0.8-0.9	≥13	128	98%
0.7-0.8	≥7	203	97%
0.6-0.7	≥5	296	97%
0.5-0.6	≥5	444	95%
	2-4	6	66%
0.4-0.5	≥5	722	75%
	2-4	57	35%
0.3-0.4	≥5	1,079	64%
	2-4	322	28%
0.2-0.3	≥5	1,865	49%
	2-4	1,297	20%

Table 10. Estimation of accuracy of results depending on frequency and probability estimated by *GIZA++*.

As can be forseen, all aligned pairs are word-to-word alignments. Nevertheless, many of the items of the same MWU appeared as probabilistic translations of the correct lexemes. For example, the lemma "trzydzieści" (*thirty*) was paired with "hogeita" (with probability 0.45) and "hamar" (0.41), "prawdopodobnie" (*probably*) with "seguru" (0.30) and "asko" (0.25), "podziękować" (to thank) with "eskerrak" (0.26) and "eman" (0.22). Not being perfect translation, this kind of result are of valuable help for language learners, as long as accompanied with examples. Although, this methods left to be trustworthy with other pairs of words that usually go together, e.g. as a translation of "Ceuta" *GIZA++* proposed "Ceuta" (0.22) and "Melilla" (0.22).

Difference in the derivational borders of morphological analysis caused many partially correct alignments. For example, this is the case of Basque genitive case which was normally reduced to nominative case by *Hunspell,* while treated as adjectives formed from nouns by *Morfeusz*. *Table 11* presents some similar cases.

Polish lemma	*Basque lemma as proposed by Hunspell*	*Correct Basque lemma*
pustynny	basamortu	basamortuko
psychologicznie	psikologiko	psikologikoki
tamtejszy	han	hango
jutrzejszy	bihar	biharko
piąty	bost	bostgarren

Table 11. Some examples of the partially correct translation, miscategorized as a result of different derivational borders applied by *Hunspell* and *Morfeusz*.

Polish preposition of very high frequency, as lacking their counterparts in Basque, were typically aligned with other Basque words of high frequency. For example the preposition "w" (*in*) (frequency: 11278), "z" (*with*) (7087),"na" (*on*) (6024)

and “do” (*to*) (5000) were proposed as translation equivalents of Basque conjunction “eta” (*and*), all with the probability close to 0.2.

As a anecdote can be mentioned that the word “ni” (*me*) was proposed with 0.22 probability as a translation of the word “optymista” (*optimist*) which, far from being the correct translation, summarized the spirit of the project.

10 Internet interface

As has been already said this project expects to offer its results for two groups of users: ordinary users (language learners), seeking for word translations, and expert users, lexicographers authorized to work on the raw output of WA, that is, on half-finished product, to convert it in full-fledged dictionary, which will offer fully reliable translations.

In this way, the dictionary can be consulted by ordinary users from the first moment, although, always with the caveat of offering not totally trustworthy information. At the same time lexicographers can keep working in order to improve the quality of the product day by day. Nowadays Internet offers such possibility.

The output texts were uploaded to *MySQL* database as three tables (a list of Polish words, a list of translation equivalents, a list of sentences (that is, the part of the corpus itself). Data was complemented with frequencies of lemma occurrences in corpus and probabilities of being correct translation, as estimated by *GIZA++*. This information can help an ordinary user, indicating the trustworthiness of the translations.

Existing PHP interface was enhanced to manage new kinds of inquiries: inverted search in existing Basque-Polish dictionary, based on a strings search, (as shown in *Fig. 2)* and the new probabilistic dictionary (*Fig. 3*). Data from three new tables are combined offering all translation equivalences of a given word, its frequencies, probability of being correct and a list of aligned sentences in which the given pair appears, offering some kind of aligned KWIC (key word in context) list. These two kinds of information are presented simultaneously.

Fig. 2. Example of an inverted search in Basque-Polish dictionary: the word "dom" appears as a translation of the following Basque headwords: *egoitza, txalet, etxe, etxepe, bizitegi, bizietxe bizitoki, bizigu.*

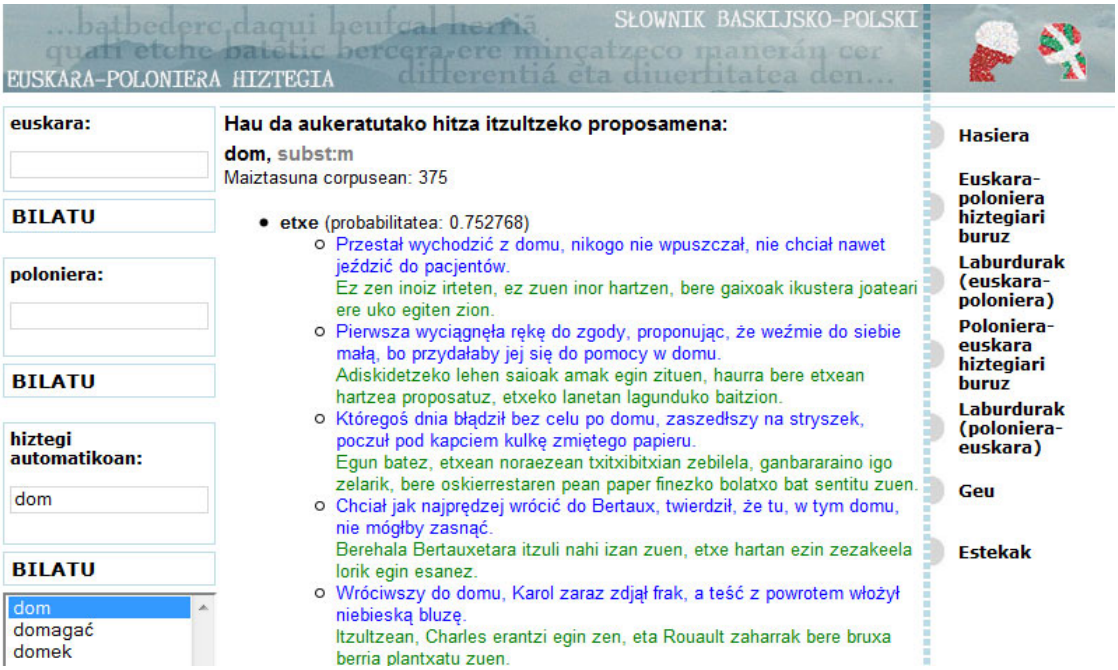

Fig. 3. Example of a search in Polish-Basque probabilistic dictionary: the word "etxe" with the probability of 0.75 is the translation of the word "dom".

The interface for authorized users offers access to the same data with the possibility of editing every item. For now the editor is simple, but will be modified

according to particular needs, which are expected to be coming out during the progress of work.[7] *Figure 4* presents one of the screens of the lexicographer's interface.

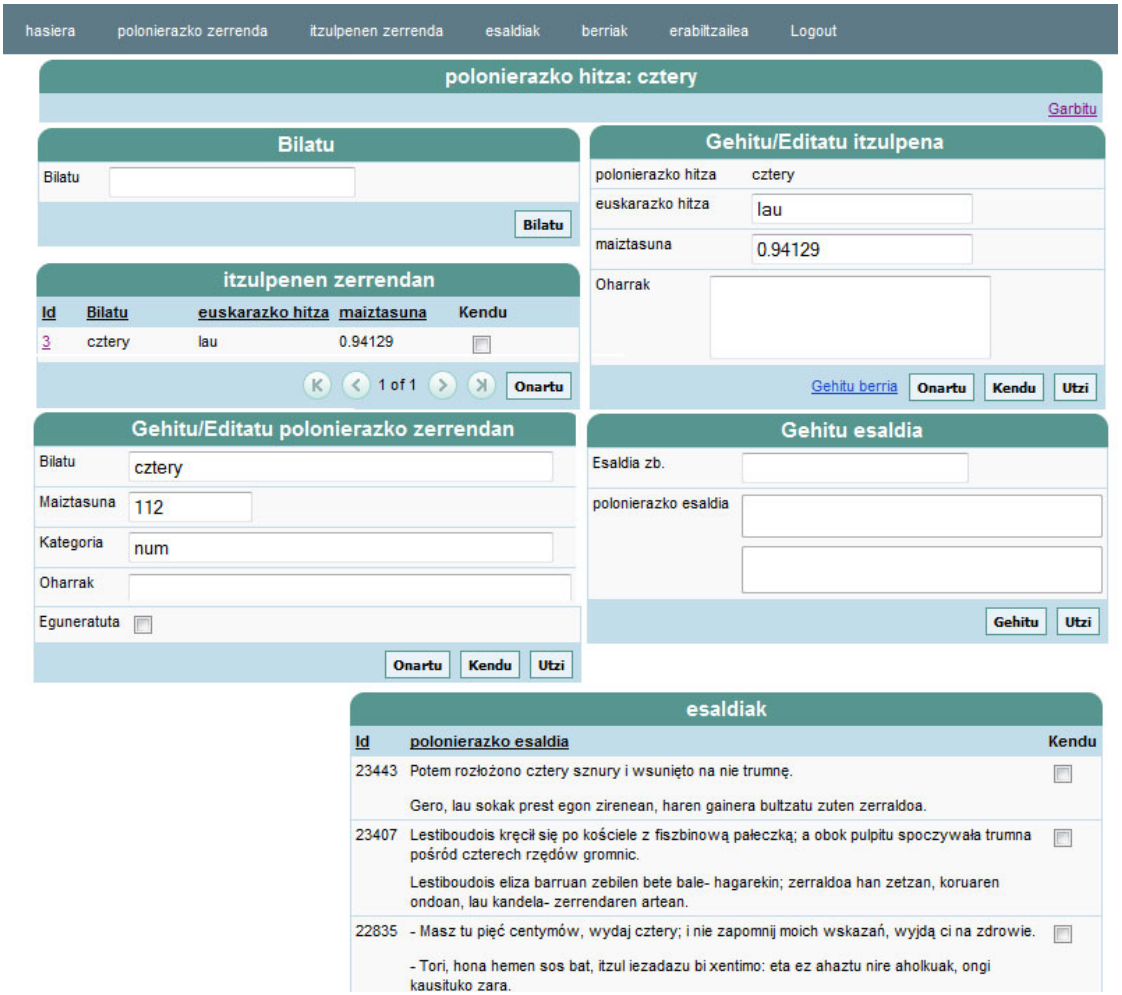

Fig. 4. Sample screen of the lexicographer's interface

11 Conclusions, further work

The aim of this study was to scrutinize the possibility of extracting a Polish-Basque dictionary from existing Polish-Basque corpus by creating such dictionary and evaluating the output obtained. As shown in *Section 8,* the results are satisfactory, and the probabilistic dictionary has been already uploaded to Internet and made available to broad audience.

The next step and main follow-up project is to manually scrutinize the probabilistic output of the dictionary. The editing interface will allow performing it in relatively short period of time.

[7] Available at: http://www.baskijski.net (learner's interface) and http://www.baskijski.net/admin/login.php (lexicographer's interface)

Nevertheless, there are much room to further improvements and enhancing of the probabilistic dictionary. Two possible ways can be considered: (i) enlarging the corpus and (ii) improving the analysis of the existing one.

The first way, i.e. enlarging the corpus, could be done e.g. by mining the Internet for texts in Basque and Polish which could be aligned. *Wikipedia* could be such a source, although there is a large disproportion between the Polish and Basque *Wikipedia*. The Polish *Wikipedia* project is the fourth largest (after English, German and French) counting the number of articles (approx. 630,000) and the Basque has forty sixth place with approx. 42,000 articles (as for August 2009). Another resources which can come in handy, and overcome the scariness of aligned data are external dictionaries combined with using bridge-languages, as proposed in (Borin 2000).

The other interesting way of enhancing the dictionary will be the incorporation of new morphological analysis and WA tools. Especially handling of MWU in Basque will allow for better alignments. MWU could be identified by means of n-gram statistics in the pre-process step. Observed in the last years a growing tendency of offering basic, general-use tools on GNU licence, or even freeing the till know commercial software is encouraging and promises possibilities in the future.

References

Aduriz, I., A. Díaz de Ilarraza. 2003. *Morphosyntactic disambiguation and shallow parsing in Computational Processing of Basque*. In: Anuario del Seminario de Filología Vasca Julio de Urquijo: International journal of basque linguistics and philology, Nº. Extra 46, 2003, pags. 1-23.

Agirre E., A. Díaz de Ilarraza ,G. Labaka, K. Sarasola. 2006. *Uso de información morfológica en el alineamiento Español-Euskara*. XXII Congreso de la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural)

Al-Onaizan, Y., U. Germann, U. Hermjakob, K. Knight, Ph. Koehn, D.l Marcu, K. Yamada. 2000. *Translating with scarce resources*. In: *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pp. 672–678, Austin, TX.

Bojar, O., M. Prokopová. 2006. *Czech-English word alignment*. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, ELRA, pp. 1236-1239.

Borin, L. 2000. *You'll Take the High Road and I'll Take the Low Road: Using a Third Language to improve Bilingual Word Alignment*. In: *Proceedings of the 18th COLING*.

Chiao, Y.C., O. Kraif, D. Laurent, T. M. Huyen Nguyen, N. Semmar, F. Stuck, J. Véronis, W. Zaghouani. 2006. *Evaluation of multilingual text alignment systems: the ARCADE II project.*

Gómez Guinovart, X., E. Sacau Fontenla. 2004. *Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos.* Procesamiento del Lenguaje Natural, 33, pp. 133-140.

Grzegorczykowa R., R. Laskowski, H. Wróbel (ed.). 1999. *Gramatyka współczesnego języka polskiego. Morfologia*. Wydawnictwo Naukowe PWN. Warszawa.

Han, B. 2001. *Building a Bilingual Dictionary with Scarce Resources: A Genetic Algorithm Approach*. In: *The Student Research Workshop, the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, Pittsburgh.

Hiemstra, D. 1997. *Deriving a bilingual lexicon for cross language information retrieval.* In: *Proceedings of Gronics 21-26*.
Hiemstra, D. 1998. *Multilingual domain modeling in twenty-one: automatic creation of a bi-directional lexicon from a parallel corpus. Technical report*. University of Twente, Parlevink Group.
Hualde J.I., J. Ortiz de Urbina. A grammar of Basque. 2003. Mouton de Gruyter. Berlin.
Kornai, A. 1992. *Frequency in morphology*. In: I. Kenesei (ed). *Approaches to Hungarian*. Vol 4 (1992) 246-268.
McEnery, T. 2003. Corpus Linguistics, c. 24, pp. 448–463. In: R. Mitkov, (ed). *The Oxford Handbook of Computational Linguistics.* Oxford University Press.
Melamed, D. 2000. *Word-to-Word Models of Translational Equivalence*. In: *Computational Linguistics*, 26.
Mihalcea, R., T. Pedersen. 2003. *An Evaluation Exercise for Word Alignment*. In: *HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pp. 1–10, Edmonton, Canada.
Niessen, S, H. Ney. 2004. *Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information.* Computational Linguistics, 30(2):181-204.
Och, F. J., H. Ney. 2000. *Improved Statistical Alignment Models*. In: *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447.
Och, F. J., H. Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, volume 29, number 1, pp. 19-51.
Piasecki, M. *Cele i zadania lingwistyki informatycznej*.
Pietrzak, J. 2002. *Polish-Basque Dictionary (Project Report)*. Investigationes Linguisticae. Volume VII, p. 23, Poznań.
Popović, M., D. Vilar, H. Ney, S. Jovićčić, Z. Šarić. 2005. *Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian-English Statistical Machine Translation.* ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, pp 41-48, Ann Arbor, Michigan.
Simões, A. M., J. J. Almeida. 2003. *NATools - A Statistical Word Aligner Workbench.* In: *Procesamiento del lenguaje natural*. ISSN 1135-5948, Nº. 31, 2003 , pp. 217-224.
Szafran, K. 1997. *SAM-96 — The Morfological Analyser for Polish*, In: A.S. Narin'yani (ed.): *Proceedings of International Workshop DIALOGUE'97* Computational Linguistics and its Applications, Yasnaya Polyana, Russia, June, 10-15, 1997, pp. 304–308.
Świdziński, M., M. Derwojedowa, M. Rudolf. 2002. *Dehomonimizacja i desynkretyzacja w procesie automatycznego przetwarzania wielkich korpusów tekstów polskich*. Bulletin De La Société Polonaise De Linguistique, fasc. LVIII, Warszawa.
Tiedemann, J. 2003. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Studia Linguistica Upsaliensia 1.
Tufiş, D, A. M. Barbu, R. Ion. 2004. *Extracting multilingual lexicons from parallel corpora.* Springer, Netherlands.
Véronis, J. 2000. *From the Rosetta stone to the information society: a survey of parallel text processing*. In: Jean Véronis (ed.) *Parallel text processing: alignment and use of translation corpora*. Dordrecht: Kluwer, pp. 1-24.
Tokarski, J. 1951. *Czasowniki polskie. Formy, typy, wyjątki*. Słownik. Warszawa.
Tokarski, J. 1993. *Schematyczny indeks a tergo polskich form wyrazowych*. Ed. Zygmunt Saloni. Wydawnictwo Naukowe PWN, Warszawa.

Weiss, D. 2005. Stempelator: A Hybrid Stemmer for the Polish Language. Institute of Computing Science, Poznań University of Technology, Poland, Research Report RA-002/05.

Wilks, Y. 2003. *Computational linguistics: what comes, what goes* . . . . In: G. Willée, B. Schröder, H.-C. Schmitz, (ed)., *Computerlinguistik Was geht,was kommt?* Computational Linguistics Achievements and Perspectives. Gardez!-Verlag, Sankt Augustin.

Woliński, M. 2006. *Morfeusz - a Practical Tool for the Morphological Analysis of Polish*. Springer, Berlin / Heidelberg.

Appendix A Polish part of Lagun corpus, lemmatized (output of *Morfeusz* program)
Appendix B Polish part of Lagun corpus, disambiguated (output of *TaKIPI* program)
Appendix C Basque part of Lagun corpus, lemmatized (output of *Hunspell* program)
Appendix D Polish input for *GIZA++* (lemmatized sentences*)*
Appendix E Basque input for *GIZA++* (lemmatized sentences
Appendix F gizacfg – *GIZA++* configuration file used in the final step of this project
Appendix G List of all Polish headwords proposed by *GIZA++*, with their respective frequencies and grammatical categories, filtered (file compiled on the bases of various files outputted by *GIZA++* and *TaKIPI*)
Appendix H List of all translation equivalents proposed by *GIZA++*, with their respective probabilities, unique ids and sentence representations, filtered (file compiled on the bases of various files outputted by *GIZA++*)