WikiTrans: The English Wikipedia in Esperanto

Eckhard Bick
GrammarSoft ApS & University of Southern Denmark
eckhard.bick@mail.dk

Abstract:
WikiTrans is a translation project and web portal for translated Wikipedias. Using the GrammarSoft's rule-based GramTrans technology, we created a high-quality English-Esperanto machine translation system, and used it to translate the entire English Wikipedia (ca. 3.000.000 articles), at a speed of 17.000 articles a day. The translated articles are searchable both locally (www.wikitrans.net) and in the original Esperanto Wikipedia, where we maintain a revision interface for users who wish to turn our translated articles into new "originals". In this paper, we explain the framework and challenges of the project, and show how translation rules can exploit grammatical information provided by a Constraint Grammar parser.

1 Motivation

In practice, Wikipedia is now the world's main encyclopedic information source, both in terms of size and user base, and although the quality of individual articles may vary, a system of mutual author control, sourcing enforcement and dispute or excellence markers help users to judge the quality and trustworthiness of a given article. However, in spite of being egalitarian and democratic from an authoring point of view, Wikipedia is far from balanced language-wise. Thus, its English information content is considerably larger than that of other languages and completely dwarfs that of minor languages (Fig. 1). The difference is visible not only in the amount of head words covered, but also in the depth and research level of the individual article. In a sense, language barriers are preventing Wikipedia from achieving its primary goal - to make the knowledge of the world accessible to all its citizens..

The Esperanto Wikipedia, although impressive in relative terms, compared to the size of its user base, and as large as e.g. the Danish one, has only 140.000 articles, while the English Wikipedia with its 3.4 million articles (or 2.345.000.000 words) is roughly 24 times as big. In addition, there is a difference in article size[1], with an average of 3.600 letters (~ 600 words) for English and German, and a little over 1500 letters (~ 250 words) in Esperanto, translating into an even bigger factor of difference, 57, when focusing on content volume. In other words, more than 98% of the English language information is not accessible in Esperanto (or Danish). One could argue that the Esperanto articles concentrate on the important and frequently sought-after topics, but it is not least in this kind of major articles that the difference in depth is most palpable, compounded by correspondingly fewer internal links (indirect depth shortage).

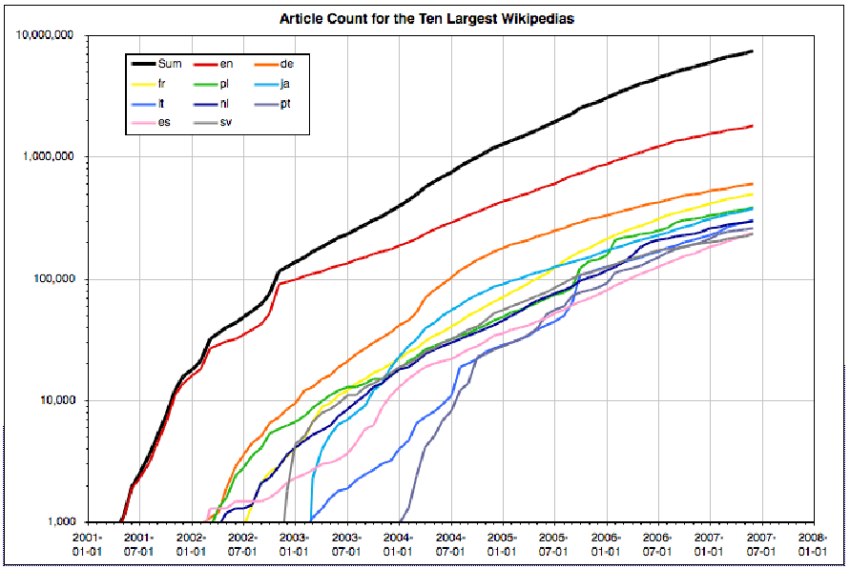

Fig. 1: Chronological language statistics for Wikipedia

Even at the price of some cultural biasing, one obvious solution for this problem is the translation of the English Wikipedia into Esperanto, thus permitting Esperanto readers from different

1 http://stats.wikimedia.org/EN/TablesArticlesBytesPerArticle.htm

countries to access the English ”über-Wikipedia”, and possibly those in other major languages (as defined by size of articles, culture or number of speakers). Manually, at a translation speed of 500 words an hour, such an English-Esperanto translation would cost 4.690.000 man hours. In Denmark, this is equivalent to 3.000 man years, or - at 0.25 EUR/word - ~ 600 million EUR. An unimaginably large sum, beyond any hope of public, let alone private or commercial funding. And even if a one-time funding could be found, it would not be possible to maintain translations in sync with originals, resulting in a rigid system difficult to update.

2 Our solution

The only logical solution to this dilemma, in our view, is the use of machine translation (MT) to save man power, possibly in combination with voluntary linguistic post-revision, for instance concerning major topics, or simply motivated by user interest, professional or private. MT is capable of solving both the quantity and the updating issues, because it allows easy and regular addition of new articles or the management of changes in existing articles. A possible problem for an MT solution is the fact that Wikipedia articles are by no means simple texts, that the lexicon covered is gigantic in its encyclopedic nature, and that any serious user community would demand a fluent and accessible translation without too many errors or untranslated source-language inserts. For the majority of languages there simply is no MT system of sufficient quality, and Esperanto, in particular, is virtually absent from the inventory of the usual commercial MT providers, be it Google, Systran or others.

Technically, MT falls into two technological camps - on the one hand rule based, symbolic systems, on the other statistical machine-learning systems, both having advantages and disadvantages. The traditional solution is the rule-based one, in line with the analytical-structural tradition of general linguistics. The method is, however, very labor-intensive, and too dependent on specialized linguistic skills to be of interest to commercial companies, if the language in question is small in market-economic terms. Statistical MT (SMT) does not need linguists and authors, but only their data, and with a bilingual text collection (a parallel corpus) and preferably as linguistically annotated text data, it is possible to cheaply train a translation model for a new language or domain. In this approach, the problem is that quality is proportional to the amount and quality of training data, and that good SMT therefore needs huge human-translated, i.e. parallel, corpora. Google, for instance, has this in the form of people's bilingual web pages, but not in sufficient quantities for small languages.

GramTrans (Bick 2007-1) is a relatively new approach to MT. Though rule based, the system saves some of the work by exploiting the robustness and depth of existing Constraint Grammar (CG) analyzers (Karlsson 1990). Mature CG parsers offer both better coverage and higher accuracy than most systems, so that GramTrans can build on the linguistic information already available in syntactic-semantic CG analyses of a given sentence (Fig. 2). For instance, the translation module can exploit dependency links between words, as well as their function tags (e.g. 'subject', 'predicative') and semantic classes (e.g. 'tool', 'vehicle', 'food'), in order to craft conditions for the selection of one or other translation alternative in the case of ambiguous constructions, polysemous words, or usage-governed synonym conventions. While CG rules remove, select, add or change linguistic tags (PoS, inflexion, function ...), translations rules simply add yet another layer to this process, targeting translation equivalents and movement operations rather than tags. In operational terms, GramTrans' MT rules are very close to CG proper, since both types of rules work by checking a list of context conditions (e.g. neighboring or dependency related words and their functions or semantic types, valency fillers etc.).

Traditional Constraint Grammar is designed to work on raw, running text, with linguistic analysis and corpus annotation in mind. While most systems do handle sentence separation, tokenization and abbreviations fairly well, and some are robust enough to manage simple corpus markup, they will not automatically handle full xml, documents or the like. In an applicational context, not least when working on heavily layouted text such as Wikipedia, with images, tables, footnotes, links and macros, wrapper solutions are therefore necessary. In order to separate layouting information from grammatical information, we implemented a system where all such information is turned into so-called style tags. This solution permits the wrapper program to reconstitute the exact text attributes and layouting after the CG and translation steps, while at the same time allowing CG rules to make active disambiguation use of such non-linguistic information, for instance in order to recognize titles or links as linguistic units deserving separate syntactic treatment.

3 The WikiTrans project

GramTrans is the motor in the MT technology used by the Danish company GrammarSoft, which offers,

in cooperation with the Norwegian company Kaldera, translations between the Scandinavian languages, and between these and English. GrammarSoft has a close cooperation with the University of Southern Denmark, and a correspondingly strong focus on research, so it was possible to launch WikiTrans, a project without any obvious commercial potential, with the explicit goal of making major language Wikipedias accessible to minor languages, with the English-Esperanto language pair as a proof of concept. Apart from the author, also GrammarSoft's programmer, Tino Didriksen, has been involved in the project.

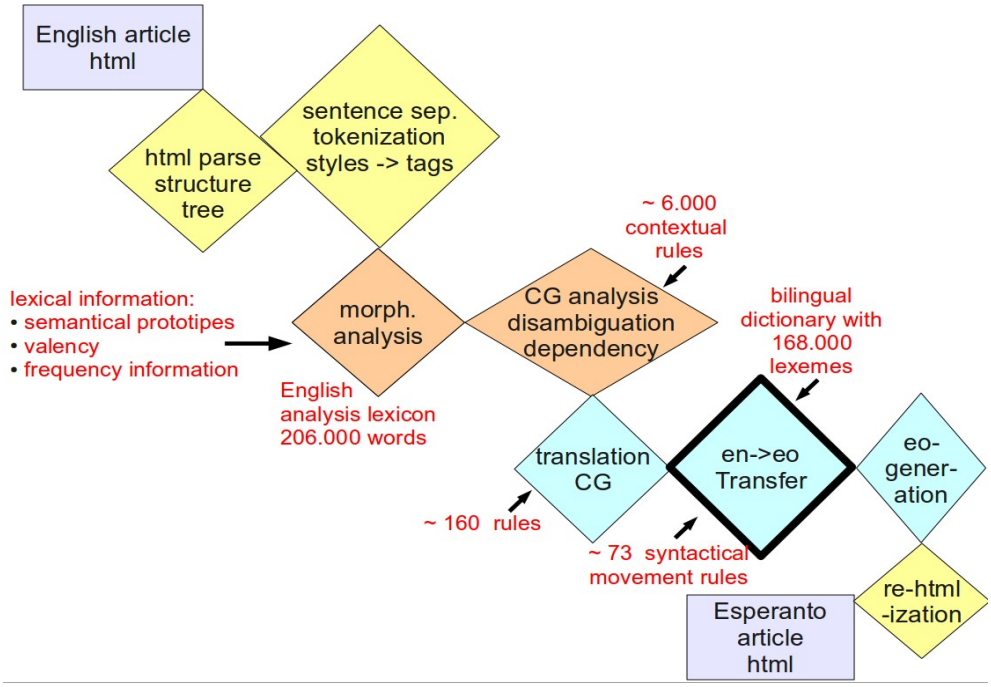

Fig. 2: Flow chart of the WikiTrans modules

The WikiTrans-project was conceived in 2009, and has gone through the following phases:

- preparation phase: 2009 - February 2010: linguistic and lexicographic work
- 1st translation phase (Feb/Mar 2010): 100.000 most frequently read articles
- 2nd translation phase (Mar-Jun 2010): 500.000 longest articles, plus articles with one-word titles (i.e. items more likely to be nouns than names)
- 3rd translation phase (Jun-Dec 2010): the main bulk, ultimately covering all 3 million articles
- use phase: updating, re-translations, human revision

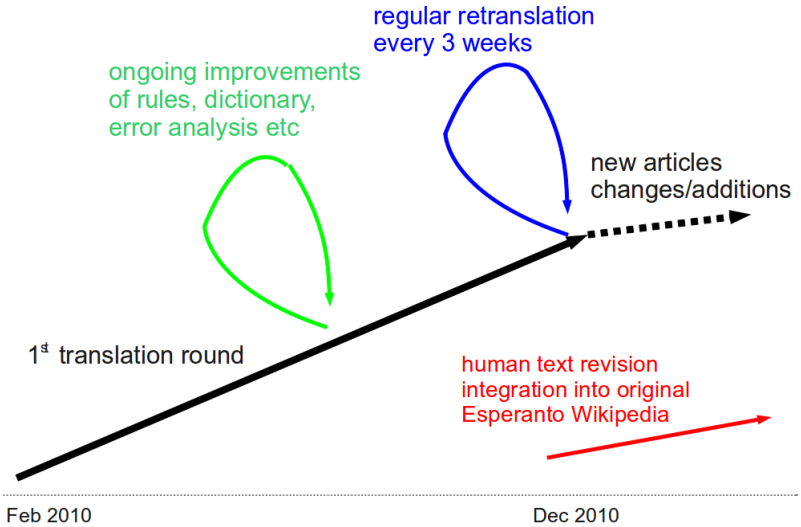

Fig. 3: Project phases of WikiTrans

4 The search interface of WikiTrans

An important reason for translating the whole Wikipedia, rather than simply translate the individual article once a user asks for it, is the possibility to systematically access and search all information. Live translation, though technically possible, would mean either searching in English or translating the search term into English, then choosing between the (English) articles before translating one of them live. Such a service would in reality only serve near-bilingual users preferring to

read in Esperanto rather than English. To really search in another language, the to-be-searched text has to exist in that language, especially considering that many search terms may not even be title words themselves, but still occur several times within the text body of different articles. On the technical side, pretranslated articles allow faster loading, and smoother internal link navigation, and allow a separation, and therefore optimization, of translation infrastructure and server search load.

For WikiTrans, we use the open-source search program *Lucene,* which allows multiple search terms at the same time, and contains an algorithm to order hits according to probability and relevance, based on term frequency and co-occurrence in individual articles. *Lucene* marks this with a probability index between 0 and 1, to which we have added a few further criteria: For instance, a article will be moved to the top of the list if the search term appears as part of the title, or - in the case of a multi-word search expression - if the words appear next to each other, overriding in-article frequency counts. The user is presented with a list of max. 20 hits, providing both title and a short snippet (Fig. 4) to allow quick, but informed selection clicks. The chosen article or articles will be presented with exactly the same layout as the original, with the same pictures, table structure etc., but entirely in Esperanto.

From a technical, programmer's point of view, a very challenging aspect of the search interface was the enormous amount of data - more than 20 GB of text (100 GB with grammatical tags). In order to effectively search a data space of this order, special database optimizations are necessary, and even using cash memory is problematic because at some point searching the growing cash memory becomes less effective than searching the database itself. Unlike a corpus linguist, who is prepared to wait for minutes for the results of a statistical or concordance search, the patience-horizon of the average Wikpedia user is only a few seconds, preferably less than one second. After that, many people may even repress the search button, forcing the server to search for the same information twice, and possibly contributing to server overload.

Fig. 4: From search term to WikiTrans article

In order to allow alphabetic searches, or get an overview over the range of articles, we have also made it possible to simply thumb through the article list from A to Z, using a letter tree ordering system, where the user moves from first to second to third letter and so on, until finally choosing from a one-screen subsection of article names.

5 Links and Bibliography

An important aspect of an electronical encyclopedia, and one of its major advantages over a paper-based one, are internal links. It is such links that combine the readability and fluency of an overview article with the much greater depth of a major background article. Simple back-and-forth clicking will allow everybody to read the article at exactly their individual knowledge level, using or not using internal links to define scientific terms, visualize personal names or explore the thematic context of a given assertion.

Technically, internal links posed several problems: First, during the translation run, there was no guarantee that the linked article had already been translated, so we had to add the (interim) option of live translation, and make sure that sufficient server capacity was available. Second, because the system is handling translations in a semi-intelligent, context-dependent way, the same word chain may receive different translations in different places, with the risk of the translated (in-context) link not matching the (out-of-context) translation of the linked title. We solved this problem by conserving the original English term (or a digital hash representation of it) in the <a href> mark itself, invisible to the user. After the translation and database creation phases, we then in a third step (taking almost a week) matched link translations to title translations.

External links and references are technically more simple, but often full of names, abbreviations and numerical expressions making translation difficult. After first trying to translate as much as possible, we now apply a more cautious policy, not translating a large part of the names, and discussing the option of not translating book and film titles either. Because it is difficult for an automatic system to be reasonably sure what is a work of art, personal name, publisher name or town name, the simplest solution would be not to touch Wikipedia bibliography sections at all, not least considering that the external sources linked will themselves not be in Esperanto, and in a certain sense often serve the function of authenticity proof more than that of background reading.

6 Integration with the monolingual Esperanto Wikipedia

The feedback reactions WikiTrans has received from the Esperanto community, were generally very positive, though many seemed to focus on the publicity aspect more than on the information content aspect. It is difficult for a lay person to appreciate the difficulty of the task, or to compare results with those for other minor languages in Google's translator, Babelfish or the like, and - understandably - the most common critical comment was therefor that translation quality was not good enough, and that the project might ”dilute” the quality of the existing Esperanto Wikipedia. And of course, though good enough for fluent reading, our automatic translations are by no means error-free, nor is a translated article a new original.

Still, this argument can be refuted by pointing out that even without an MT system, it has always been the case that minor-language Wikipedia authors have heavily borrowed from articles in other, major languages by means of translation. In fact, the open-source framework of Wikipedia encourages and supports this flow of text from one language to another. Is it not then better to perform this work more efficiently and rapidly with the help of an automated system? What is needed, is simply marking what's what, and where the user is in a browser clicking chain at any given point in time. Our own proposal is a traffic light colour marking - a red corner mark for a ”virgin” MT-derived WikiTrans article, green for a fully revised article and yellow for a superficially revised article. ”Green” articles could then be moved into the ”true” Wikipedia (while retaining the marker), and red or yellow articles would be searchable both through the WikiTrans portal and - in the case of search failures, or to increase accessible information - in the monolingual Esperanto Wikipedia itself. Fig. 5 shows our scheme for integrating translated and original Wikipedias.

In consultation with Wikipedia administrators, we addressed the practical aspects of this integration between July 2010 and February 2011. The current state of affairs is a solution where user-side javascript programs interact with the GramTrans software at its own server. The user-side software was developed by Marek Blahus (E@I), while Tino Didriksen (GrammarSoft) implemented the necessary GramTrans interface, handling the slightly idiosyncratic internal Wikipedia-syntax, and creating a graphical revision interface. At the time of writing it is already possible for individual registered Wikipedia users to activate the revision-

and-integration module, and parallel WikiTrans searches have been activated for the general public, using WikiTrans as a fall-back solution for search failures.

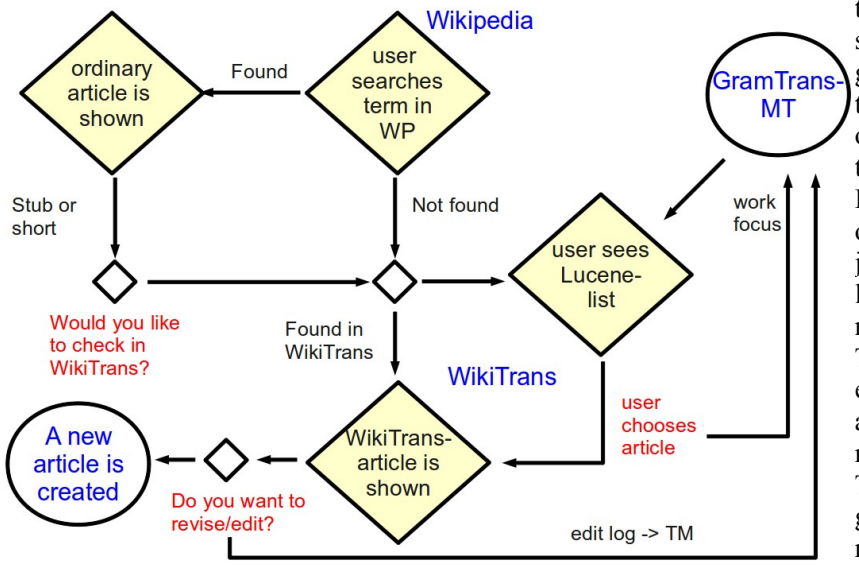

Fig. 5: Integration with the original Wikipedia

7 Linguistic aspects of the translation interface

From a classification point of view, GramTrans is neither a surface MT system nor an interlingua system (Fig. 6). It avoids, of course, the problems of simple word-for-word translations, but does not risk abstraction all the way up to an interlingua level. The ”costs”, in terms of robustness losses, for a full symbolic interlingua are very high, and it is possible to achieve the same with a somewhat ”flatter” transfer from source to target language - simply because most language pairs have more in common, structurally and semantically, than there are differences. This is true also for the English-Esperanto language pair - even more so, because Esperanto with its constructional flexibility is an ideal target language, allowing to mold translations to grammatical patterns found in many different languages without the results sounding unnatural.

As pointed out above, GramTrans relies on comprehensive and robust analysis of the source language, in this case provided by the EngGram parser (http://visl.sdu.dk/visl2/constraint_grammar.html). EngGram is a CG system with more than 6000 rules, a 200.000 word core lexicon, and a dependency style syntactic analysis (Bick 2005). In experiments reported in (Bick 2009), EngGram was evaluated on Wikipedia texts with F-scores of 98.2 and 93.4 for PoS/morphology and syntactic functions, respectively. GramTrans exploits the categories and word links from the EngGram source language analysis in order to create lexical transfer rules designed to resolve semantic ambiguities and choose the correct translation equivalent among several options. The third step, generation, profits heavily from the morphosyntactic flexibility of Esperanto, and from the fact that the generation of an Esperanto morpheme (ending or affix) is almost equivalent to just specifying the desired linguistic category (tense, number, part of speech etc.). The task is made almost embarrassingly simple by the almost perfect regularity and modularity of the language. The only complication in generation is therefor syntax, or rather word order, because in spite of an officially free word order, Esperanto does of course have fairly strong usage conventions with regard to constituent order, and ignoring them - even if not agrammatical as such - would impair fluent reading.

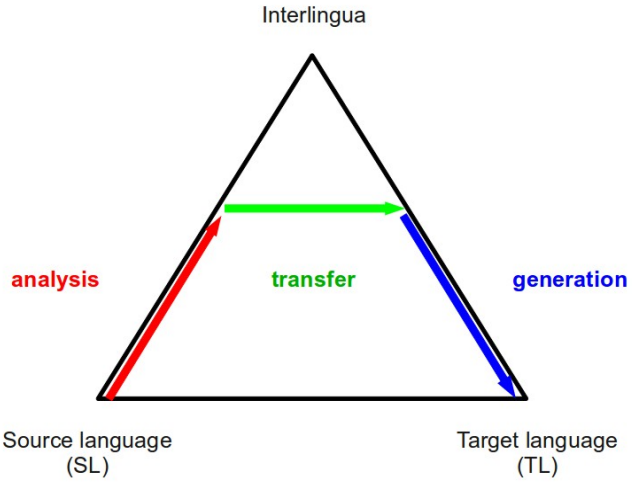

Fig. 6:The translation triangle

7.1. Lexical transfer

The simplest way to exploit Constraint Grammar tags for lexical transfer is one-dimensional in the sense that only local tags (i.e. of the word itself) are used as sense discriminators. This method simply exploits part of speech (1-2) or inflexion (3-4):

1. type_N (noun) **:tipo, :speco**
2. type_V (verb) **:tajpi**
3. force_NS (singular) **:forto**

4. force_NP (plural) **:armeo, :trupo**

In a two-dimensional approach, transfer choices are based on contextual CG information instead, either directly, or indirectly in the form of local tags with relational meaning, such as function tags (5), semantic roles or valency instantiation (e.g. <¤vr> for reflexive verbs where a reflexive pronoun has been found in the context.

5. rather_ADV ... S=(@ADVL) :prefere;
   S=(@>A) :sufiĉe;

Even lexeme-bound traits such as morphological features or semantic class can sometimes be harvested from context, as when nominal agreement features are propagated from head noun to (underspecified) determiner or attribute. An example from the generation task is the fact that Esperanto adjectives have number, while English ones don't, and we use CG propagation rules to add the correct number feature to adjectives. And in the lexical transfer module the ±human is frequently exploited as a translation discriminator, and can be contextually propagated by projecting the feature onto nouns that function as subjects of cognitive or communication verbs, even if the noun itself is sense ambiguous or semantically unmarked due to incomplete lexical information.

6. too_ADV ... S=(@ADVL) :**ankaŭ**;
   S=(@>A) P2?=(INFM)_por :**tro**;
   D=(@>A) :**tro**

Example (6) contains both indirect relational tags (function tags for S=self) and direct relational tags (function tags for D=dependent), as well as positional conditions (P2=second word to the right). All in all, our transfer rules use the following relations:

Dependency: S=self, D=daughter, M=mother,
B=brother, GD=granddaughter, GM=grandmother
Position: right P1, P2 ... Pn, left P-1, P-2 ... P-n

The targeted distinctions do not necessarily reflect conventional dictionary or encyclopedic distinctions. Among other things, metaphors or genre-variation may well be isomorphic in the two languages, making an explicit distinction irrelevant. In more general terms, one can say that one of the biggest secrets of MT (and an important reason for not going all the way to the top of the translation triangle) is the importance of *distinguishing* rather than *defining*. In other words, it is sufficient to have enough context and semantic knowledge in the system to select one or other translation equivalent, but the final *understanding* will only occur in the mind of the target language reader, who has more world knowledge and other background context than any computer could possibly have - so there is no need for the system to explicit everything at an abstract, super-linguistic level. A large part of the semantics is simply transported unaltered from source to target language, without real disambiguation having taken place. For instance, the metaphorical use of containers as units works similar in all languages (2 *glasses of* beer - 2 *glasoj da* biero). On the other hand, it may sometimes be necessary to separate (mainly target language) usage-differences (synonyms, frequency considerations), on top of possible sense distinctions. This problem is less pertinent in Esperanto than in other languages, but it does exist.

Together, the various disambiguation techniques permit quite complex lexicographical work, the most important aspect being the possibility to link verbal information with tags attached to the complements of a given verb (Bick 2007-2). The example below shows how contextual transfer discriminators are able to translate the English verb 'apply' into 9 different verbs in Esperanto. Contexts are tried in successive order, and if no later context conditions apply, the first translation is chosen as the default. It is therefor important to make sure that this translation is robust and maximally ambiguous rather than just the most frequent translation for the word in question.

```
apply_V :uzi;
   D=("for")_pri :peti
   D=(<H> @SUBJ) D=("to"'PRP)_por :kandidatiĝi
   D=(@ACC) D=("to" PRP)_al :apliki
   D!=(@ACC) D=("to" PRP)_por :validi
   D=(<(conv|sem)> @SUBJ) D!=(@ACC) :validi
   D=(<(cm.*|rem)> @ACC) :surŝmiri
   D=("dressing" @ACC)_pansaĵo :surmeti
   <¤vr> D=("to" PRP)_pri :koncentriĝi
   D=("match")_alumeto :malestingi
```

[@SUBJ=subject, @ACC=accusative object, PRP=preposition, <H>=human, <conv>=convention, rule, <sem>=semantical, <cm>=concrete mass word, <rem> remedy, substance, <¤vr>=reflexive]

7.2 Multi-word expressions, translations memory and names

In some cases, it doesn't make sense to translate a word chunk analytically-sequentially - the meaning of the whole is not transparent from the meanings of its parts. GramTrans handles these cases as ”words” with internal spaces. The concept covers complex

nouns *(recovery_position - savpozicio),* a very common category in English, but also prepositional or adverbial phrases with summary translations *(in_violation_of - malobee_al, every_inch_as - tute_same, all_year_round - tutjare),* or simply fixed expressions such as *see_also - vidu_ankaŭ.* Multi-word expressions are not only relevant to the translation module, but also play a role during morphosyntactic analysis, where the concept of complex function words, in particular prepositions and conjunctions, simplifies the assignment of syntactic functions and relations: *each_other (unu_la_alian), instead_of (anstataŭ), other_than (krom).*

A similar simplification can be gained from translations memory (TM) lists, common in many MT systems, and useful to cover special words that are always translated in the same way, i.e that are contextually unaffected and that can be inserted into a translation without any need for transfer rules. One field of TM application are terminology lists, which our systems can turn on or off depending on the to-be-translated domain. But it is also possible to use TM to remedy systematic errors, that can be fixed with a once-and-for-all intervention. In the revision interface we programmed for WikiTrans articles, the system thus remembers all human-made corrections. Besides providing an overview of errors and MT feed back, the change log can be fed into a translation memory, or even used to suggest to the reviewer drop-down translation alternatives for frequently mis-translated expressions.

Independently of the name-recognition (NER) qualities of the EngGram parser, names are hard to translate, and being a very productive category, they have an exceptionally bad lexicon coverage. It isn't eve n possible to trust upper case initials, since uppercasing may occur for other reasons, such as sentence-initially, after a colon, or simply as a means of emphasis. Therefore, it is not possible to 100% sure whether a word is a name or an unknown or compound word from another PoS class. From a purely MT perspective, the question is whether to translate a name, retain the original form or transliterate it with target language phonetics. Here, it is important to distinguish between two main scenarios:

(a) institutions and events, to be translated part for part

European Union - Eŭropa Unio,
Olympics - Olimpikoj,
World War II - Dua Mondmilito

(b) personal names and product names, to be left untranslated

George *Bush - **Georgo Arbusto*

For WikiTrans we also have a compromise solution, where the original is retained, but accompanied by a translation in parentheses, for instance in the case of book, music or film titles that are clearly marked as such in Wikipedia's html structures.

8 Generation and Structural transfer

The last step in the translation chain is morphological and syntactic-structural generation. Again, we exploit CG information inherited by the translation module from the EngGram parser. Basically structural transfer is achieved with the help of movement rules that can change the order of constituents (as defined by the set of all dependency daughters of a target word), using CG tag conditions, and optionally adding or replacing grammatical traits or word forms. One of the structural problems we had to solve was turning genitives into (moved) preposition phrases *(Michel's father - la patro de Michael)*. In some cases, no direct translation exists, and only structural rephrasing can approximate the intended meaning, or it may be necessary to add or remove constructions necessary only in one of the languages, such as English *don't* negation, English *do* questions or Esperanto *ĉu*-questions (yes-no-questions).

As suggested above, the second generative task, morphological generation, is very simple in Esperanto, but in cases where Esperanto is grammatically more explicit than English, context may be needed to add the desired feature. Apart from plural agreement on noun dependents, this is the case for the accusative marker *-n,* which in Esperanto attaches to all nominal word classes and had to be recovered from indirect clues such as CG function tags. Also, the two languages differ in their use of participles (e.g. English *have*-tense), and sometimes there are clashes between semantic and surface number *(wages [pl] - salajro [sg], stools [pl] - feko [sg])*.

9 Conclusions and Perspectives

The language technology project WikiTrans (www.wikitrans.net), succeeded in little more than a year to create an English-Esperanto MT system of sufficient quality to automatically translate Wikipedia texts, and finished in December 2010 the translation of the about 3.000.000 articles in the English Wikipedia, at a speed of ~17.0000 articles a day. The system offers not only target language searches inside translated articles, but also allows integration into Wikipedia proper, through a post-

editing interface.

The perspective for 2011 is the creation of a framework for automatical retranslation and updating. For this purpose the project is setting up a linux cluster consisting of 8 four-core computers to handle fast and parallel MT. The hardware has been sponsored by ESF (Esperanto Studies Foundation), and is hosted at the University of Southern Denmark. Depending on the degree in which the community accepts and uses our post-editing interface, we plan regular treatment of error statistics and corrections suggestions.

A remaining linguistic challenge is terminology: Despite the fact that the WikiTrans dictionary with its 168.000 entries is already the largest English-Esperanto dictionary ever produced, many specialized terms continue to be translated using heuristic methods, e.g. analytical or partial translations, transliterations, Latinisms etc. As a minimal goal, these automatic suggestions should be validated by hand (either by the author, or through a community web portal). Also, existing terminological dictionaries should, if copyright allows, be integrated - which is not as easy as it might seem. First, entries that are assumed to be translations, may in reality be explanations, definitions or terms at a slightly different level in the other language, while what is needed is terms that can directly replace a target language term in the same context, with the same inflexion etc. Second, ambiguity may arise between a specialized term and the same word's meaning in everyday language. If such ambiguities are not spotted and handled with transfer discrimination rules, they will result in a deterioration of the system, with rare translations supplanting common ones. Ideally, new terms should be subjected to a discussion in Esperanto professional and scientific communities, stimulating terminological work proper rather as opposed to mere lexicography, but given the size of the language community, for many domains this is not a likely outcome.

Long term, WikiTrans is to cover further language pairs, the 2011 focus being on English-Danish. From a quantitative point of view, this task is similar to Esperanto, both in terms of article number, article size and size of the bilingual MT lexicon, and we therefor expect a certain synergy, for instance in the identification and translation of ”unknown” English complex nouns, and in the harvesting and classification of name expressions. Another logical step would be the addition of another source language for the same target language - Esperanto, which would allow the user to fill in ”cultural information gaps” - a possible problem immanent to any monolingual Wikipedia. A second source language would also make it possible to compare same-topic articles in areas where information may be biased (e.g. politics, history, religion). GramTrans itself already has a working Danish-Esperanto system, and it would be technically feasible to add translations from further languages using open source systems such as Apertium (http://www.apertium.org/), if and when such a system reaches a sufficient quality level.

Bibliography

Bick, Eckhard. 2005. ”Turning Constraint Grammar Data into Running Dependency Treebanks”. In: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (ed.), *Proceedings of TLT 2005 (4th Workshop on Treebanks and Linguistic Theory, Barcelona, 2005)*, pp.19-27

Bick, Eckhard. 2007-1. ”Dan2eng: Wide-Coverage Danish-English Machine Translation”. In: Bente Maegaard (ed.), *Proceedings of Machine Translation Summit XI, 10-14. Sept. 2007, Copenhagen, Denmark*. pp. 37-43

Bick, Eckhard. 2007-2. ”Fra syntaks til semantik: Polysemiresolution igennem Dependensstrukturer i dansk-engelsk maskinoversættelse”. In: Henrik Jørgensen & Peter Widell (eds.), *Det bedre argument, Festschrift til Ole Togeby på 60-årsdagen* pp.35-52

Bick, Eckhard. 2009. ”Introducing Probabilistic Information in Constraint Grammar Parsing”. In: *Proceedings of Corpus Linguistics 2009, Liverpool, UK*. Electronically published at: ucrel.lancs.ac.uk/publications/cl2009/

Karlsson, Fred. 1990. Constraint Grammar as a Framework for Parsing Running Text. In: Karlgren, Hans (ed.), COLING-90 Helsinki: *Proceedings of the 13th International Conference on Computational Linguistics*, Vol. 3, pp.168-173