CLIoS: Cross-lingual Induction of Speech Recognition Grammars

Nadine Perera[1], Michael Pitz[2], Manfred Pinkal[1]

[1]Department of Computational Linguistics, Saarland University, Saarbrücken
[2]Department of Human Machine Interaction, BMW Group, Munich
nadine.perera@gmx.de, michael.pitz@bmw.de, pinkal@coli.uni-saarland.de

Abstract

We present an approach for the cross-lingual induction of speech recognition grammars that separates the task of translation from the task of grammar generation. The source speech recognition grammar is used to generate phrases, which are translated by a common translation service. The target recognition grammar is induced by using the production rules of the source language, manually translated sentences and a statistical word alignment tool. The coverage of the resulting grammars (for Spanish and Japanese) is evaluated on two corpora and compared quantitatively and qualitatively to a grammar induced with unsupervised monolingual grammar induction.

1. Introduction

The localization of spoken dialogue systems is currently gaining interest because of the commercial demand to apply those systems to many different languages. We report project work that starts out from the EU-project TALK[1], which focused on the development of new technologies for adaptive multimodal and multilingual human-computer dialogue systems and produced (amongst others) the SAMMIE system (Becker et al., 2006), a flexible spoken language MP3-player interface for in-car application.

Given a well designed system architecture, the relevant language-dependent modules of a spoken dialogue system are the speech recognition and the language generation component. We focus on the localization of the speech recognition and interpretation component, at the example of the rule-based grammar of the SAMMIE system. Even though this work was developed focusing a specific application, we expect that it is relevant for many possible dialogue system applications.

To construct a speech recognition grammar for a specific language and domain, it is necessary to collect a lot of data to estimate which expressions the user is likely to use, typically in a Wizard-of-Oz experiment. Then, the evaluated data has to be incorporated into the speech recognition grammar. This is time consuming and expensive, and it would be beneficial if the result of data gathering and grammar building in one language could be transferred to other languages (semi-)automatically.

In this paper, we focus on the construction of a rule-based grammar for a new target language by porting a grammar from a source language. There are three options to consider:

1. Data collection and construction of target language grammar for every language separately.
2. No data collection, but direct translation of the grammar by a human expert.
3. No data collection and semi-automatical translation of the source grammar.

Option 1 is very labour-intensive, Option 2 has high error potential and requires human expertise in many fields. Therefore we chose Option 3. The key feature is to separate the task of *translation* from the task of *grammar writing*. The translation is done by a human translator whereas the grammar construction is done by an automatic induction algorithm.

In contrast to traditional grammar induction algorithms, which try to find bracketings for a corpus of sentences by judging similarities and differences (van Zaanen, 2000; Kuhn, 2004; Kuhn, 2005), we exploit the syntactic rules and semantic information of the source grammar *in addition* to the sentence corpus.

Grammar induction may not result in a perfect target language grammar. Nevertheless, it saves work. A small test corpus of speech data can be gathered to evaluate the grammar's coverage and add expressions that did not emerge through the translation approach, to improve the quality of the target language grammar.

Furthermore, this approach provides us with a parallel corpus for spoken dialogue which may be relevant for other applications as well and enables us to profit from findings in the fields of machine translation and cross-lingual knowledge induction, cf. (Kuhn, 2004; Kuhn, 2005).

This paper is organized as follows: the next section explains the grammar induction in detail: sentences and syntax trees are generated, the sentences are translated and the resulting bi-corpus is word-aligned. Using information from the word-alignment, the terminals in the source syntax trees are substituted by the target language terminals, the linear order in the resulting syntax tree is adapted to form a valid target language syntax tree, which is split into production rules. The production rules are merged to obtain the target grammar. Preliminary evaluation results are included at the corresponding positions within Section 2. In Section 3, the final grammar evaluation results are discussed. Section 4 summarizes the results and gives an outlook to further work.

1.1. Related Work

Making grammars re-usable for new languages is a goal also followed by (Ranta, 2004) via a "Grammatical Frame-

[1] www.talk-project.org

work" (GF), a type-theoretic grammar formalism that addresses four aspects of grammars: multilinguality, semantics, modularity and grammar engineering, and re-use of grammars in different formats and as software components. In (Johannson, 2006) GF was used to globalize and localize a Swedish grammar for dialogue system utterances to obtain a set of grammars for Swedish, Spanish and English.

In (Perera and Ranta, 2007) GF was used for spoken dialogue system grammar localization, porting the English SAMMIE speech recognition grammar[2] to GF and then introducing multilinguality, thus creating GF grammars for English, Finnish, French, German, Spanish and Swedish. This produced a second German SAMMIE grammar that was compared to the original one, however, the German GF SAMMIE grammar did not match the coverage of the original German SAMMIE grammar.

MedSLT (Buillon et al., 2007) is a grammar-based medical speech translation system. The system supports simple medical examination dialogues about throat pain between an English-speaking physician and a Spanish-speaking patient. General feature grammar resources from the REGULUS toolkit (Rayner et al., 2006) are compiled into flatter, domain specific grammars, translation is realized via an interlingua.

Alignment Based Learning (ABL) by (van Zaanen, 2000) is an unsupervised grammar induction system based on the idea of substitutability. It can be applied to an untagged corpus of natural language sentences and produces a bracketed version of that corpus. By clustering and selecting the bracketing hypotheses, a grammar is induced which covers the original corpus of sentences plus more similar sentences.

Our approach is different from both GF and MedSLT in the respect that we do not use resource grammars. Even though resource grammars and the idea of re-using grammars is attractive, we wanted to implement a simple solution to the localization problem that does not rely on the introduction of a framework that requires ample resources in turn.

Compared to ABL, our approach requires more resources - the generated sentences, their translations, and the source grammar production rules compared to a monolingual text corpus of the target language only. On the other hand, the grammar which we induce is used for natural language interpretation, whereas ABL can so far only create a grammar that determines if a given sentence is covered by the grammar or not.

2. Cross-lingual Grammar Induction

The CLIoS (**C**ross-**L**ingual **I**nduction **o**f **S**peech Recognition Grammars) system consists of four steps to successfully induce a target grammar: first, the source grammar must be converted to a corpus of sentences which can be translated by human experts. In our approach, the source language speech recognition grammar is used as a generation grammar which generates the source sentences and their derivations., i.e. the syntax tree[3] containing information on the production rules that were used to generate it. Second, this corpus of sentences has to be translated. Third, a statistical word alignment for the parallel corpus is estimated and manually corrected where required. Fourth, using the bi-corpus of source and target phrases and the original syntactic and semantic rules of the source grammar (in the form of single syntax tree derivations of the source grammar), a new target grammar is induced. This is done by combining the rules with the translated words via alignment projection and adapting the grammar rules to the target language.

This is an overview of the tasks necessary for the grammar induction:

1. Generate corpus of source sentences and set of source syntax trees.
2. Translate sentence corpus to target language.
3. Obtain word alignment for the parallel corpus.
4. Induce target grammar by:
   (a) Substituting source terminals in source syntax trees by target terminals through alignment projection.
   (b) Adapting syntax trees to mirror the correct target sentences.
   (c) Splitting syntax trees into production rules.
   (d) Merging production rules to form the target grammar.

Tasks 3 and 4 comprise the grammar induction phases which are shown in more detail in Figure 1. Five phases are distinguished: in Phase (I), a statistical word alignment is provided by the GIZA++ tool (Och, 2003) and manually corrected where necessary. In Phase (II), the alignments are used to substitute source language terminals in the syntax tree by target language terminals, for details see Section 2.4. The resulting unordered target syntax tree is adapted in Phase (III) to comply with the linear order given in the correct translated target sentence. This is done by the `swivel` operation (cf. Section 2.4.). In Phase (IV), the re-ordered syntax tree is split into subtrees of depth 1, which can be interpreted as the production rules that define a trivial local grammar, see Section 2.5. This is done for all the sentences in the bi-corpus and results in a large set of production rules. Those rules are merged in Phase (V) to form the target grammar.

2.1. Source Grammar Format and Phrase Generation

Our designated source grammar is a context free[4] grammar in Nuance GSL format[5] (Nuance Communications, 2003) that does not contain recursive rules. Grammars with comparably flat syntax and semantics can be modeled without recursion and common speech recognizers (e.g. Nuance Communications, Microsoft) do not allow direct or indirect left recursion in particular; right and middle recursion are used infrequently.

[2] a manual translation from the German SAMMIE grammar which we discuss in this paper

[3] technically a *"generation"* or "*derivation* tree".

[4] Due to the lack of recursive rules, the *language* that the grammar describes is *finite* and therefore *regular*.

[5] a form of EBNF

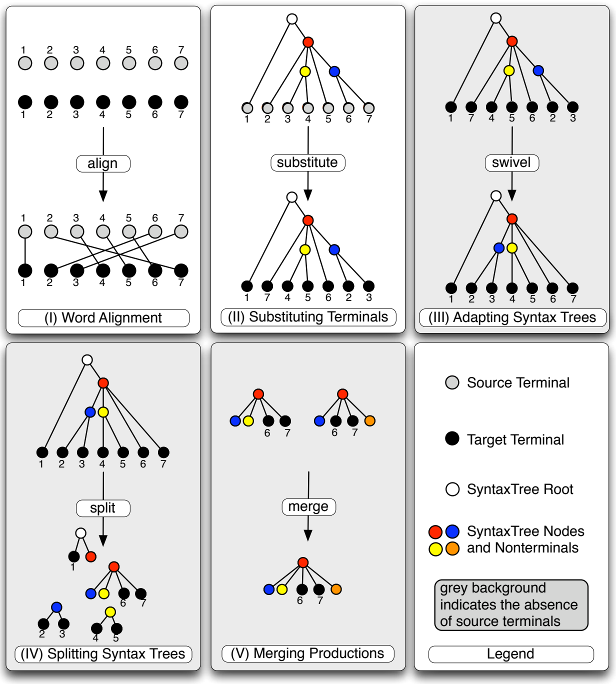

Figure 1: Induction Phases. Phase (I) shows the statistical word alignment that is used in Phase (II) to substitute source terminals by target terminals in the syntax tree. The syntax tree is re-ordered in Phase (III) and split into local structures which define production rules in Phase (IV). In Phase (V), those rules are merged to form the target grammar.

In our source grammar, NPs are modeled "semantically distinguishable": there is no generic NP production rule, but the NPs also carry their semantic value in their name, e.g. NP_{ALBUM}, NP_{SONG} etc. On the right-hand side of its rules three operators are used, namely concatenation, alternatives, and optionality.

Semantic interpretation is realized via slot filling, whereas slots or interpretation tags are not limited to contain semantic information only; meta data and dialogue state information is also transported via slots. Sentences are produced by generating a syntax tree (with interpretation tags and production derivations) and then projecting the syntax tree to a flat sentence.

2.1.1. Sample Selection

As exhaustive generation with the SAMMIE grammar is impossible due to practical restrictions (the grammar is able to generate about 10^{17} different sentences), an optimal generation strategy was chosen that provides much fewer sentences (3856) by leaving out redundant rule expansions (retaining the first expansion data and re-using it at the following expansions), without losing information. The resulting 3856 sentences are close to a minimal expansion of the grammar, which produces a corpus of 1840 sentences, but at the cost of ignoring optional and alternative constructs and omitting terminal symbols.

2.2. Phrase Translation

We use the speech recognition grammar to generate phrases, which can be translated easily by a common translation service. This general approach has the advantage that no language pairing is preferred, and that the approach is feasible as long as there is a human translator available for the language pair in question. It also makes it possible to specify the grammar and behaviour of the spoken dialogue system in one language and expect that the specification will be implemented consistently for the different, localized versions of the spoken dialogue system.

The benefits of using human translators instead of statistical machine translation include accurateness and the incorporation of context knowledge: if there is a whole sentence to be translated, many of the ambiguities that occur with the translation of single words cease to exist. Also, human translators will produce grammatically correct sentences, which is important to our approach and cannot be guaranteed by machine translation systems.

Through the translation of the 3856 sentences, we have obtained a bilingual corpus that is sentence-aligned. In addition to the sentence alignment, the semantic interpretation corresponding to the sentences is known from the interpretation tags.

A central prerequisite for the success of our strategy is the cross-lingual validity of our semantic tag categories. The existence of the parallel corpus gives us the opportunity to confirm this prerequisite. Knowledge from the source language can be re-used in a straightforward manner if these semantic tag categories can be re-used (cf. next section).

We tested the hypothesis that the interpretation tags are consistent or at least transferable over languages: For the language pairs German-Spanish and German-Japanese, analysis results show that there is either a direct translation for the interpretation feature in the target languages or a corresponding mapping that a human translator will use automatically when translating the source sentence to the target language. The semantic transfer happens automatically via the implicit information that the translator uses.

Note that some interpretation categories may not be relevant for certain languages, e.g. the addressing style for English (cf. examples below). The cross-lingual grammar induction approach described here requires that all interpretation categories relevant for the designated target languages are contained in the source grammar even though they are not directly needed there. In our case, porting grammars from German to Spanish and Japanese, this was not necessary, which is one of the reasons why we chose to start out from a German grammar.

Examples:

1. `phrase_mood`: phrase moods like "indicative", "imperative", "interrogative" exist in most European languages (Bodmer, 1997). Japanese has similar features and with the translation of phrases, the interpretation tag on phrase moods is meaningful and correct for the target languages considered here.

2. `addressing_style`: "duzen" and "siezen", a German way of addressing a person familiarly or formally by using different grammatical persons, non-existent

in English: In Spanish, this distinction is realized similarly as in German, only that the third person singular instead of the third person plural is used to be formal. In Japanese, there are two styles that are mainly used; the simple form that is used when speaking with family and friends, and a more formal style (called desu-masu-style), that is used for polite conversation with colleagues. These two styles were automatically chosen by the human translators to express the corresponding German sentences. If we considered English as a target language for the grammar induction, we would generate a grammar with production rules that distinguish between the formal and the informal addressing style, although the generated and accepted sentences would be identical for both styles. However, the dialogue manager could still make us of this information, for instance to decide whether to call the user by his or her given name or last name. If that distinction is not wanted, there is no harm in having too many production rules, but if the grammar engineer would like to remove the unnecessary productions, he or she could easily attain that goal due to the semantic and syntactic information coded into the production rule name (e.g., by searching the grammar for the keyword "`FORMAL`").

3. `album_noun`: one example of meta information that the grammar interprets is which noun of a group of synonyms the user utilizes to express a semantic entity, e.g. an album. The system echoes the synonym that the user speaks, and uses the word "album" if the user asks for "albums", but "record" if the user demands for "records". Since it is improbable that those synonym groups contain the exact same number of entries in all languages, and even more improbable that there is a direct one-to-one-mapping between the elements of those synonym groups, this meta information may not be transferred completely automatically, but as all the words in one group mean the same thing, being synonyms, this is not a problem.

2.3. Word Alignment and Terminal Substitution

The bilingual corpus resulting from the manual translation is aligned on sentence level. For the terminal substitution, we need an alignment on word level to transfer the syntactic and semantic information from the source to the target language. This word alignment is obtained by the GIZA++ tool (Och, 2003) plus manual correction where necessary. The necessity is determined by human inspection first, but can also be done automatically by finding language-pair and phrase mood typical characteristics. Also, wrong alignment links can be found by inspecting those sentence pairs from which reordering problems emerge, see Section 2.4. According to the alignments found, the source language terminal symbols in the syntax trees are substituted by the target terminals. Source terminals without matching target terminals are deleted, target terminals without matching source terminals are inserted at the appropriate position in the syntax tree automatically, but only after the reordering, see Section 2.4.1.

2.3.1. Alignment Definitions

Terminal substitution occurs in blocks, as we adhere to a general definition of alignments that allows for many-to-many alignments (Och, 2003): an alignment is defined on a source string $s_1^J = s_1, ..., s_j, ..., s_J$ that is aligned to a target string $t_1^I = t_1, ..., t_i, ..., t_I$. We define an alignment between the two word strings as a subset of the Cartesian product of the word positions; that is, an alignment $\mathcal{A}$ is defined as $\mathcal{A} \subseteq \{(i, j) : j = 1..., J; i = 1..., I\}$. Such an arbitrary relation between source and target language positions allows for a source word to be aligned to none, one, or many target words and vice versa, where the many source or target words need not form a sequence. An *alignment block* is a subset of the previously defined alignment relation with the restriction that the one or more source words $s_x^y = s_x, ..., s_y$ and the one or more target words $t_q^r = t_q, ..., t_r$ that form a connected component are also sequences, i.e. there is no source word s_k with $x < k < y$ and no target word t_l with $q < l < r$ that is not part of the alignment block.

The terminal substitution algorithm is designed to substitute words blockwise, but often, the alignment blocks are trivial one-to-one mappings. Connected components that are not alignment blocks can also be substituted by the algorithm, but pose a problem to the swivel algorithm in Phase III, which is why we implemented a fallback solution (cf. Section 2.4.) for uncontinuous many-to-many alignment mappings.

2.3.2. Alignment Quality

Current cross-lingual approaches work with large corpora, for instance the EUROPARL (Koehn, 2002) corpus which contains 34K sentences and dwarfs our small bilingual corpus by a factor of 10. However, the sentences in our domain are less complex in structure and meaning than the political debates in the EUROPARL corpus, and consist of a smaller vocabulary.

In a pilot study, 30 sentences of the generated 3856 were translated manually and word-aligned by GIZA++, using a manually crafted dictionary and resulting in an F-Score[6] of 0,76. In order to calculate precision and recall, a gold standard was established.

The comparatively high median frequency of words in our corpus, due to the limited vocabulary, suggested that further improvements in alignment quality could be expected when the complete corpus of 3856 bi-sentences is considered. As Table 1 shows, this is true, the F-Score for the full corpus with dictionary (0,94) proved to be significantly higher than for the mini corpus considered before.

Number of Sentences considered	30	3856
Distinct words	60	482
Total Words	342	31173
Median Word Frequency	5,7	64,7
F-Score	0,76	0,94

Table 1: Comparison of preliminary mini corpus and full training corpus.

[6] We used the harmonic mean of precision and recall.

To determine the F-Score, we established a gold standard for the full corpus. This took one annotator about 5 hours, as the alignments were already of quite high quality due to the use of a weighted dictionary. The alignments that had to be corrected were mainly function words.

The induction process was conducted with the manually corrected alignment where incorrect word alignments were indicated by a problem to the "`swivel`" operation described in the next section.

2.4. Parse Tree Adaptation

To establish order in the modified syntax trees, we compare them to the original target language sentences and change the linear order of the syntax tree's terminals to match the well-formed, translated target sentence. The reordering is done by the "`swivel`" operation, which changes the order of children in a node, but cannot alter node dominance. Yamada and Knight also used that view on syntax tree structure (Yamada and Knight, 2001), which Gildea compared to an "Alexander-Calder-Mobile" (Gildea, 2003). We found the "`swivel`" operation to be sufficient to produce valid adaptions of a syntax tree in 97% of the cases, since the trees are of fairly flat structure. However, 3% of the Spanish corpus contained disconnected alignments, i.e. one-to-many alignments where the many terminals did not form an *alignment block* as defined in Section 2.3.1. Three courses of action were possible: change the alignment paradigm to align only adjacent terminals, eliminate the problematic sentences from our corpus and not consider them for the merging of the grammar, or implement a fallback solution for this special case.

We chose to implement an alternate solution for the 135 sentences that could not be reordered completely by the swivel operation, which is to cut the branch off at the appropriate level and reassign it to its rightful parent node. This is done only for the problematic alignment link in question, the remaining reordering of the tree is done via the "`swivel`" operation.

2.4.1. Insertion of Terminals

So far, the syntax tree contains only the terminals that could be projected across the languages. Therefore, the target terminals that are still missing in the tree have to be inserted. There are three options for the insertion strategy: post-order (attach to the parent of the terminal that precedes the inserted terminal in the correct sentence), pre-order (attach to the parent of the terminal that succeeds the inserted terminal), or in-order (both post- and pre-order, i.e. attach the terminal in question under the first common parent of the preceding and the succeeding terminal).

Figure 2 shows an example where an English sentence (Figure 2(a)) is projected to Spanish sentence (Figure 2(b)) by literal translation of the single words. The correct Spanish translation of the sentence, however, reads "Pedro quiere **a** María" [7].

Figure 2(c) shows the terminal "a" that needs to be inserted - to the resulting string, the insertion strategy is of no relevance, but it affects the structure of the syntax tree, as can be seen in Figure 2(d-f), where the syntax tree structures resulting from the different insertion strategies are displayed. Note that the syntactic category that "a" belongs to is not known to us, therefore, the terminal "a" is inserted directly under on of the discussed nodes. Analyzing the correct target sentence with a Part-Of-Speech tagger could probably give us the syntactic category for "a", which could be used to structure the induced grammar more nicely, however, we chose to insert the words directly to avoid overgeneration, as the syntactic category found by a POS tagger would be too general for our approach.

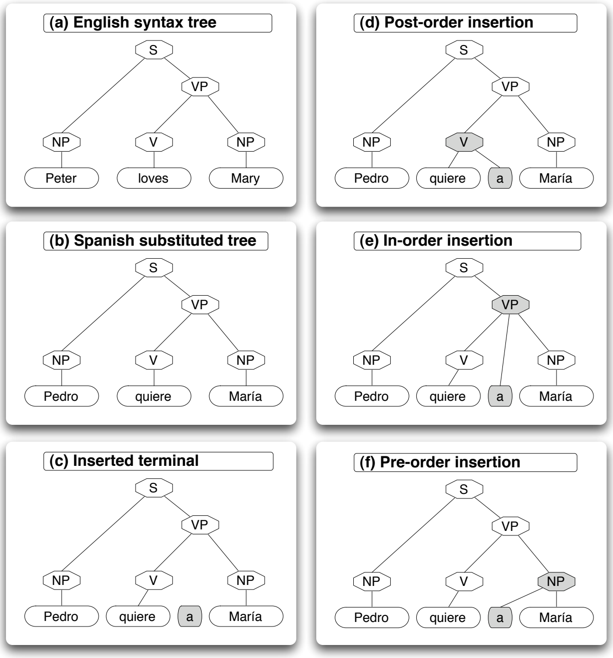

Figure 2: Insertion Strategies. The English syntax tree from (a) is projected to Spanish (b), the terminal "a" must be inserted (c). The terminal can be inserted in post-order (d), in-order (e), or pre-order (f).

We chose the in-order insertion strategy as we expect it to produce the most reliable results because it uses two information sources instead of one (the successor and the predecessor).

After the insertion of missing terminals at the appropriate places, the corrected syntax trees reflect a valid derivation for the corresponding target sentences, and implicitly contain the production rules that were needed for their derivation.

2.5. Production Rule Merging

Since a grammar is defined by a set production rules (which implicitly contain the information needed for a formal definition, namely set of terminals, set of nonterminals, and start production), we obtain a grammar by combining and merging all the production rules that the grammar should contain. Every syntax tree is `split` in its participating production rules, all rules are collected, "uniqued" (only one of a set of identical production rules is used) and `merged`. The sum of merged production rules form the

[7] In Spanish, prepositions are used for case identification whereas in English, case identification is shown via word order.

induced target grammar. Merging must take account of the three grammar rule operators introduced in Section 2.1. above. They are treated by the following rules:

1. Alternatives with matching identifiers are merged to alternative lists.

2. Concatenation lists are merged “modulo” optional operators if the non-optional elements are identical.

3. If two concatenation lists cannot be merged, they are “ORed” by adding a new alternative containing the two concatenation lists.

To guarantee the correctness of our implementation, the `split` and `merge` operations were tested monolingually by splitting the source sentence syntax trees without substituting terminals, and then merging the resulting productions back together. This resulted in a source grammar equivalent to the initial source grammar. Equivalence was shown by generating 200,000 random sentences with one grammar and successfully parsing them with the other grammar, and vice versa, several times.

3. Evaluation

We evaluated the complete induced target grammar according to three criteria:

1. Does the grammar correctly parse and interpret the training corpus of target sentences?

2. Does the grammar perform in a sufficient way on a test corpus of user utterances in the target language?

3. How does the grammar compare to a grammar which is obtained by monolingual grammar induction based on the training corpus only (van Zaanen, 2000), with respect to coverage on the training corpus and the test corpus?

3.1. Results for Coverage and Interpretation

Coverage and correct interpretation was tested on two corpora for each language: the “training corpus” of the 3856 generated and translated Spanish/Japanese sentences that were used to induce the two CLIoS grammars, and the “test corpus” of Spanish/Japanese user utterances collected via a Wizard-of-Oz experiment.

3.1.1. Training Corpus

We found that of the 3856 sentences from our Spanish training corpus, i.e. the generated and manually translated Spanish sentences, 100% are parsed and interpreted correctly by the induced grammar. As expected, the same can be said about the Japanese training corpus, the induced Japanese grammar covers 100% of the training corpus.

3.1.2. User Utterance Corpus

Designing a grammar is an iterative process: the more potential speakers are consulted on how they would express a given concept, the more possible user utterances emerge. These utterances tend to converge, of course, so that the probability that a given utterance was already made and added to the grammar before converges asymptotically to 100%, but unknown utterances can always emerge and a (non-recursive) grammar can only cover a fix amount of possible phrases. This situation can be compared to Nielsen’s view (Nielsen, 1993) on heuristic evaluation, where *“individual experimenters can perform a heuristic evaluation of a user interface on their own, but the experience from several projects indicates that any single evaluator will miss most of the usability problems in an interface. However, since different evaluators tend to find different problems, it is possible to achieve substantially better performance by aggregating the evaluations from several evaluators.”*

We can see the subjects in our evaluation as heuristic evaluators of the grammar, finding problems (out-of-grammar utterances). Projecting Nielsen’s concept of heuristic evaluation to our case, complete grammar coverage over all possible utterances is improbable to be achieved for small numbers of evaluators.

The user utterance corpora were collected via a Wizard-of-Oz-style experiment with native speakers, 10 subjects for Spanish and 5 for Japanese. The subjects were shown the German SAMMIE system and received a short introduction relying to them the possible actions within the MP3 domain which they could carry out by speaking to the system (e.g. listening to an album or modifying playlists). Then, the investigator explained the tasks to them in *German* to prevent delivering certain words to the subjects that should be chosen freely, and the subject formulated what he/she would say in this situation in the car, in his/her native language (*Spanish*/*Japanese*).

If the subject could think of several possible utterances, all the utterances that were made entered the evaluation corpora. The whole session was recorded and the user utterances were transcribed manually. For Spanish, this resulted in a corpus of 281 utterances for 27 different tasks, for Japanese we collected 135 utterances for 27 tasks.

Figure 3 shows a picture of the experimental setup.

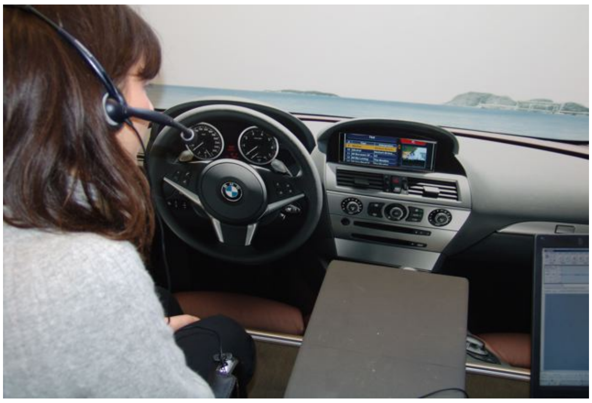

Figure 3: Experimental Setup. The subject in the mock-up system’s driver seat, investigator recording utterances and explaining tasks from passenger seat.

Of the 281 Spanish utterances, 17 (6,04%) could not be interpreted by the induced grammar. Analysis showed that of the 17 problematic utterances, 15 (88%) addressed command tasks such as stopping playback or scrolling the screen. The problem source was the use of words that do not exist in the grammar so far, to some extent because the words that the subjects used were colloquial or uncommon[8].

To summarize: 93,96% of the Spanish user utterance corpus was parsed and interpreted correctly. The remaining 6.04% could be inserted easily into only 5 different production rules of the induced grammar. Table 2 displays the achieved coverage for both corpora and both grammars (CLIoS-ES is the Spanish grammar, CLIoS-JP the grammar for Japanese).

Induced Grammar	CLIoS-ES	CLIoS-JP
Coverage Training Corpus (%)	100	100
Coverage User Corpus (%)	93,96	89,63

Table 2: Coverage of the two CLIoS grammars for Spanish (CLIoS-ES) and Japanese (CLIoS-JP) on the training corpus (3856 sentences) and the test corpus of user utterances (281 utterances for Spanish, 135 for Japanese).

Of the 135 Japanese utterances, 121 (89,63%) were accepted and interpreted correctly by the induced grammar. Similarly to the Spanish utterances, one problem source was the use of words that do not exist in the grammar so far, but could be added easily. Another problem was that some users chose to say “I-want-to” sentences that would formally begin with “watashi-wa” (“I”), but left out the “watashi-wa”, which was not an option in our induced grammar. By making this part of the production rule optional after the induction, which was possible through simple search and replace commands, the utterances lacking the “watashi-wa” could be parsed and interpreted correctly.

3.2. Results for Comparison of Monolingual and Cross-Lingual Grammars

We used ABL (van Zaanen, 2000) to monolingually induce a grammar (called ABL-ES) from the 3856 Spanish sentences, and likewise a Japanese ABL grammar (ABL-JP) from the 3856 Japanese sentences. We compare the two ABL grammars to the two created by cross-lingual induction (called CLIoS-ES and CLIoS-JP) with respect to coverage on the training corpus and the test corpus.

However, the ABL framework does not permit the interpretation tags that are projected to the CLIoS grammar from source to target language via the word alignment. Thus the ABL grammar can be used for syntactic analysis rather than semantic interpretation.

Qualitatively, it can be said that the ABL grammars are more difficult to comprehend by humans than the CLIoS grammars, first due to the lack of production rule names which convey an idea of what to expect from the right side of the production rule to the grammar engineer, and second due to the amount of recursive rules which complicate human comprehension.

Left recursion in the ABL grammars had to be removed as common speech recognition grammar compilers, in our case Nuance, cannot handle left recursion (Moore, 2000; Nuance Communications, 2003). The left recursion was removed by using the algorithm from (Moore, 2000) as the standard algorithm could not be applied due to memory demands resulting from the complexity of the induced ABL grammars.

Table 3 shows a comparison of size in KB and the number of nonterminals for the six respective grammars (ABL induced with left recursion for Spanish and Japanese, ABL induced without left recursion for Spanish and Japanese, and CLIoS induced for Spanish and Japanese).

The CLIoS grammars were transformed from EBNF to BNF form to allow for a meaningful comparison of grammar size and nonterminal numbers. Due to the cross-lingual projection approach, both CLIoS grammars in EBNF form (CLIoS-ES and CLIoS-JP) contain exactly as many nonterminals as the initial source grammar, namely 379. Their respective sizes are 147 KB for CLIoS-ES and 207 KB for CLIoS-JP. The Japanese ABL grammar expanded heavily when the left recursion was removed whereas the Spanish ABL grammar stayed reasonably small.

The ABL-induced, left-recursion-free grammars were compared to the CLIoS grammars based on the training corpus, where both the two ABL grammars and the two CLIoS grammars correctly accept 100% of the training sentences. Of the Spanish user utterance corpus, the CLIoS-ES grammar accepts 93,96% (cf. Table 3) and the ABL-ES grammar parses only 16,35%. Of course, the ABL-ES grammar can not parse more than the 93,96% that the CLIoS-ES grammar parses, as we already established that the problem with these utterances was the use of words unknown to the grammar, i.e. words that did not occur in the target sentence corpus.

For Japanese, the CLIoS-JP grammar correctly parses 89,63% of the user utterance corpus and the ABL-JP grammar parses 11,57%.

4. Conclusion and Outlook

We have presented a strategy for the localization of speech recognition grammars that separates the task of *translation* from the task of *grammar generation* (CLIoS). The approach taken is a pragmatic combination of automated NLP methods and manual translation. A speech recognition grammar is induced cross-lingually for the target language by using the production rules of the source language, manual translation and a statistical word alignment tool.

Two grammars were induced and evaluated from a generated corpus of 3856 German sentences and their Spanish/Japanese translations. Evaluation showed that the two induced grammars correctly parse and interpret 100% of the training corpora, both Spanish and Japanese respectively. Of the test corpus of collected user utterance corpora for both languages, the Spanish grammar successfully interpreted 93,96% and the Japanese grammar 89,63%.

We were able to show that the CLIoS approach results in grammars that are easier to read by humans and therefore easier to improve afterwards than a current state of the art monolingual unsupervised induction approach (ABL). In

[8] For instance, one subject used the word “stop” to stop playback, which is not usually used by Spanish speakers, this might be the influence of living in Germany.

Induced Grammar	ABL-ES-LR	ABL-ES	CLIoS-ES	ABL-JP-LR	ABL-JP	CLIoS-JP
Size (KB)	196	526	1.497	196	5.181	6.345
Nonterminals (in BNF)	5.232	12.603	14.622	6.670	115.968	76.514
Coverage Training Corpus (%)	-	100	100	-	100	100
Coverage User Corpus (%)	-	16,35	93,96	-	11,57	89,63

Table 3: Comparison of the grammars induced by ABL and CLIoS. ES stands for Spanish, JP for Japanese, LR stands for the left recursion contained in the original ABL grammars that could not be processed by the Nuance compiler (hence no coverage data). Note that the size and the nonterminal numbers for the CLIoS grammars are taken from their BNF forms instead of the EBNF forms, to allow for a meaningful comparison.

addition to that, the CLIoS grammars performed much better on a test corpus of user utterances than the ABL induced grammars.

We evaluated the method proposed in this paper on the interpretation grammar of one specific dialogue system. However, the approach will clearly apply with similar results to other grammars, as long as they have no recursive rules and a comparably shallow syntax. Due to the shallow syntax, realizing semantic interpretation via slot-filling works well, therefore the semantic interpretation slots are integrated into our approach and mapped across languages. It will be a matter of future research to investigate how the induction approach generalises to grammars with a deeply structured syntax and more complex semantic interpretation rules.

5. Acknowledgements

We would like to thank Menno van Zaanen for letting us use his ABL grammar induction and for helpful discussions, and Robert C. Moore for removing the left recursion from the ABL-induced grammars. We are grateful for Stephan Thoma's invaluable programming help and for the fruitful discussions with Sebastian Padó and Hans Leiß. The SAMMIE grammar was written for the TALK research project[9] by Jochen Steigner, Peter Poller, and Rosemary Stegmann.

6. References

T. Becker, C. Gerstenberger, I. Kruijff-Korbayova, A. Korthauer, M. Pinkal, M. Pitz, P. Poller and J. Schehl. 2006. Natural and intuitive multimodal dialogue for In-Car Applications: The SAMMIE System. *4th Prestigious Applications of Intelligent Systems (PAIS 2006)*, pp. 612-616, Riva del Garda, Italy.

F. Bodmer. 1997. *Die Sprachen der Welt*. Parkland Verlag.

P. Bouillon, G. Flores, M. Starlander, N. Chatzichrisafis, M. Santaholma, N. Tsourakis, M. Rayner, B. A. Hockey. 2007. A Bidirectional Grammar-Based Medical Speech Translator. *In Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing (ACL-07)*, pp. 41-48, Prague, Czech Republic.

D. Gildea. 2003. Loosely Tree-Based Alignment for Machine Translation. *In Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL-03)*, Supporo, Japan.

M. Johannson. 2006. Globalization and localization of a dialogue system using a resource grammar. *Master's thesis*, University of Gothenburg.

P. Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. *Master's thesis*, University of Southern California.

J. Kuhn. 2004. Experiments in parallel-text based grammar induction. *In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-04)*, Main Volume, pp. 470-477, Barcelona, Spain.

J. Kuhn. 2005. Parsing word-aligned parallel corpora in a grammar induction context. *In Proceedings of the ACL Workshop on Building and Using Parallel Texts (ACL-05)*, pp. 17-24, Ann Arbor, Michigan.

R. Moore. 2000. Removing left recursion from context-free grammars. *In Proceedings of 1st Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 249–255, Seattle, Washington.

J. Nielsen. 1993. *Usability Engineering, Heuristic Evaluation*, 155-160. Morgan Kaufmann.

Nuance Communications, Inc. 2003. Nuance Speech Recognition System 8.5. Grammar Developer's Guide. 1005 Hamilton Avenue, Menlo Park, California 94025, USA.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), pp. 19-51.

N. Perera and A. Ranta. 2004. Dialogue system localization with the GF resource grammar library. *In Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing (ACL-07)*, pp. 17-24, Prague, Czech Republic.

A. Ranta. 2004. Grammatical framework: A type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2), pp. 145–189.

M. Rayner, B.A. Hockey, and P. Bouillon. 2006. *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler.* CSLI Press, Chicago.

K. Yamada and K. Knight. 2001. A Syntax-based Statistical Translation Model. *In Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL-01)*, pp. 523–530, Toulouse, France.

M. van Zaanen. 2000. ABL: Alignment-Based Learning. *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pp. 961–967, Saarbrücken, Germany.

[9] funded by EU FP6 [ref. 507802]