Published on Multilingual Online Translation (http://www.molto-project.eu)

D11.3 Evaluations of ACE-in-GF and of AceWiki-GF

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO — Multilingual Online Translation
Deliverable:	D11.3 Evaluations of ACE-in-GF and of AceWiki-GF
Security (distribution level):	Public
Contractual date of delivery:	M38
Actual date of delivery:	2013-05-20 (v1.0)
Type:	Prototype
Status version:	v1.0
Author(s):	Laura Canedo, Norbert E. Fuchs, Kaarel Kaljurand, Maarit Koponen, Tobias Kuhn, Jussi Rautio, Victor Ungureanu
Task responsible:	UZH
Other contributors:	UHEL

Abstract

This report describes the user evaluation of two related software products — ACE-in-GF and AceWiki-GF. The multilingual grammar ACE-in-GF is implemented in the Grammatical Framework (GF) with the goal to provide a multilingual interface to a large subset of Attempto Controlled English (ACE). We measure the accuracy with which the ACE-in-GF grammar translates ACE sentences to the other languages that it implements, and show that its translations are preferred to the translations obtained with a state-of-the-art statistical translation system. The semantic wiki engine AceWiki-GF enables collaborative knowledge engineering environments that are based on controlled natural language and implemented as GF grammars. We set AceWiki-GF up with the ACE-in-GF grammar and a geography domain lexicon, and then ask speakers of different languages to supply the wiki with geographical knowledge. We show that the automatic translation does not affect the basic functioning of the wiki: users who view the wiki content in a language different from that in which it was originally written are as likely to agree or disagree on the verity of the content than users who view the content in the same language.

Contents

1 Introduction

This report describes the user evaluation of the multilingual grammar ACE-in-GF [CFK12] and the semantic wiki system AceWiki-GF [FKK13, KK13].

ACE-in-GF is a partial implementation of Attempto Controlled English (ACE) [FKK08] in the Grammatical Framework (GF) [Ran11] which allows the automatic translation of texts from any of the many languages supported by GF into any other of those languages. Extending the preliminary evaluation performed in [CFK12] the current evaluation of ACE-in-GF covers more languages and a deeper analysis of syntactic and semantic structures. Specifically, we measure the correctness of translations of ACE sentences into the languages supported by the grammar.

AceWiki-GF, which is derived from AceWiki [Kuh08], is a wiki engine that can be used with any GF grammar. For our evaluation we based AceWiki-GF on the multilingual grammar of ACE-in-GF extended by a vocabulary for a geographical domain. The evaluation uses AceWiki-GF in a collaborative setting where users can edit the wiki content in the three languages English, German and Spanish guided by a look-ahead editor that enforces the syntactical correctness of the input. For the evaluation we look at the syntactic and semantic features of the resulting multilingual content, measure how far users agree on the meaning of the content in the three languages, and establish the overall usability of AceWiki-GF.

2 ACE-in-GF

ACE-in-GF (for a detailed description see [CFK12] and the project website[1]) is an implementation of a large subset of Attempto Controlled English (ACE) [FKK08] in the Grammatical Framework (GF) [Ran11]. This subset corresponds roughly to the subset supported by the semantic wiki engine AceWiki [Kuh08] and to the input language of the ACE→OWL translator [Kal07]. Thus this subset of ACE establishes a natural language interface to the semantic web language OWL [Gro12]. The implementation of ACE-in-GF relies heavily on the GF Resource Grammar Library (RGL) [Ran09] by importing most of the language structures of ACE from the GF resources for standard English via the language-independent API of RGL. This approach allows us to easily plug in other RGL-implemented natural languages thus creating a multilingual grammar which bidirectionally maps ACE to fragments of other natural languages. The result is an automatic definition of multiple controlled natural languages (CNLs) based on the RGL languages. The RGL-based design also guarantees that the quality and language-coverage of the grammar increases almost automatically with the increasing quality and coverage of the RGL. Still, some fine-tuning is necessary to override certain grammatical structures that cannot be specified in a completely language-independent way. The current grammar supports almost all the RGL languages including all its European languages (Bulgarian, Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Latvian, Norwegian, Polish, Romanian, Russian, Spanish and Swedish). However, for some languages, some ACE constructs (e.g. verb phrase coordination and some complex types of questions) have not been implemented since they were not directly available via the RGL. As ACE supports a certain degree of syntactic sugar, the incomplete syntactic coverage does not necessarily reduce the semantic expressivity of the grammar. Also, one has to recall that ACE is derived from English which influenced many design decisions, specifically the way how to handle ambiguity. Thus an ACE text translated into a language dissimilar from English may have a different behaviour concerning ambiguity.

In order to use the grammar in applications, one needs to supply a multilingual lexicon where words are classified according to ACE content word classes, namely common nouns (`CN`), proper names (`PN`), transitive verbs (`V2`) and transitive adjectives (`A2`). The implementation of this lexicon can rely on the RGL smart paradigms which automatically derive the complete morphological features of a word from a few input arguments, e.g. the dictionary form and the gender information.

Like any GF-implemented grammar, which describes a language as a set of functions operating on a set of categories, ACE-in-GF is very modular and it is thus easy to create subsets of the grammar by overriding the parsing start category or by disabling certain functions. This is useful in applications like query interfaces that only need a subset of the language.

[1] `https://github.com/Attempto/ACE-in-GF`

3 AceWiki-GF

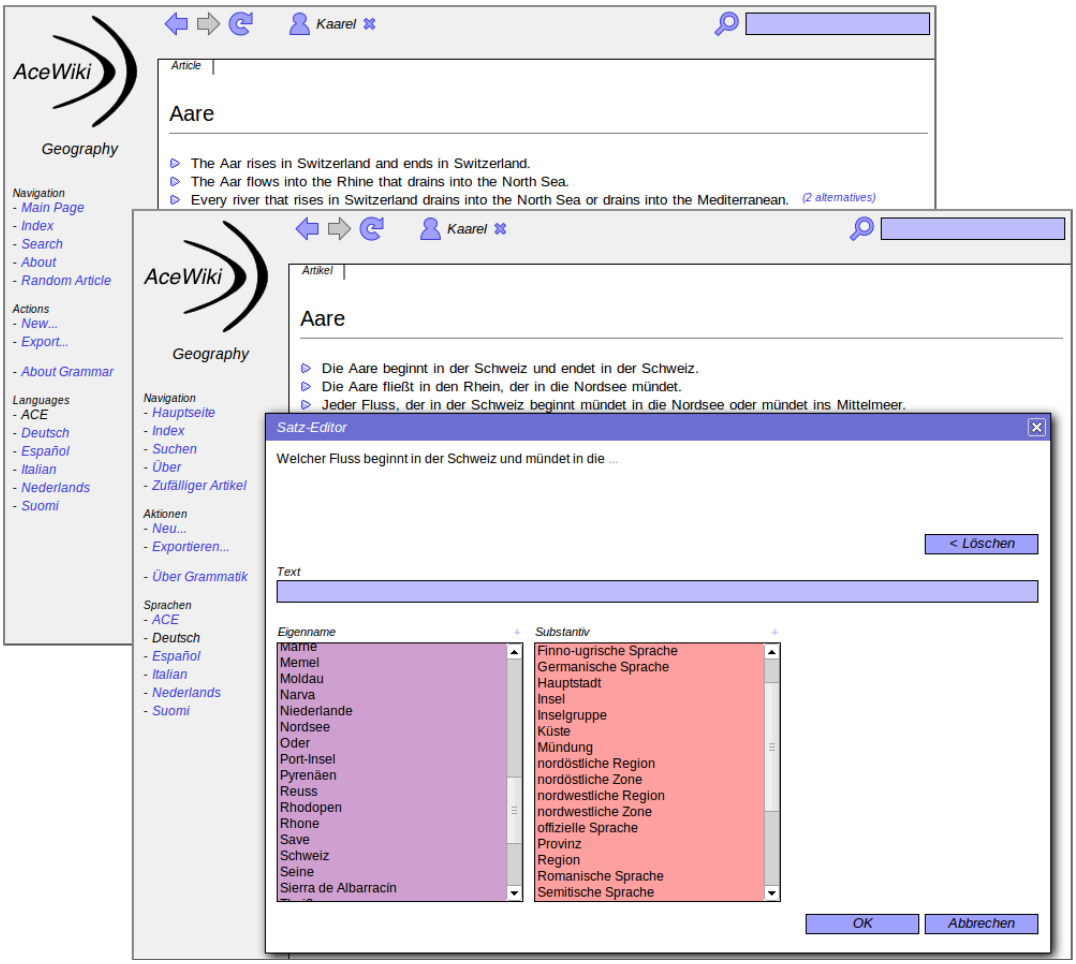

Figure 1:
Multilingual geography article that is based on the ACE-in-GF grammar and displayed in ACE and German. The wiki language of both the content and the user interface can be changed in the left sidebar. The look-ahead editor helps users to input syntactically correct sentences, in this case offering proper names and common nouns as possible continuations of the already entered partial sentence.

AceWiki-GF (for a detailed description see [FKK13, KK13] and the project website[2]) is a software tool that supports the creation of multilingual semantic wikis, i.e. environments where users with different language background collaboratively build a knowledge base. Such wikis integrate a GF grammar that defines the content language including the domain vocabulary of the wiki that could refer to "European Geography", "A Tourist Phrasebook" or "A Museum Catalogue". The wiki offers the content language to the users via a look-ahead editor that guides them when adding or editing wiki entries, which are sentences in the user's language. Any language supported by the grammar can be used via the look-ahead editor, therefore every wiki entry can be viewed, created and edited in multiple languages. This means that a sentence introduced in language A by user X can be viewed and updated in language B by user Y. Figure 1 shows a screenshot of the AceWiki-GF environment.

With our approach of controlled languages, the entire content of the wiki becomes automatically processable. The grammar offers a precise translation of wiki articles into all supported languages. Additionally, our approach allows for formal reasoning over the wiki content if a subset of ACE is among the

[2] https://github.com/AceWiki/AceWiki

supported languages. With the grammar ACE-in-GF, for example, the wiki content is available as an OWL ontology, offering the standard OWL reasoning services such as consistency checking and classification which can be used to implement various user-level features such as pinpointing semantic errors and question answering.

4 Evaluation of ACE-in-GF

A preliminary evaluation of the ACE-in-GF grammar was reported in [CFK12]. The evaluation described in this report is more comprehensive and detailed, covering more languages, using more participants, and featuring an improved evaluation methodology.

4.1 Evaluation Methodology

The evaluation of the multilingual translation accuracy of the ACE-in-GF grammar had three goals:

1. assess the grammaticality and acceptability of the sentences in multiple languages;
2. collect information about textual changes considered necessary by native speakers;
3. compare the accuracy to that of an off-the-shelf translation system, concretely Google Translate.

The evaluation material comprises a total of 111 sentences (848 words in English) generated by two different methods and then converted into the evaluation languages. A detailed description of the methods used for generating the sentences is given in Section 4.2.

The multilingual evaluation was carried out in 10 European languages: Catalan, Danish, Dutch, Finnish, French, German, Italian, Norwegian, Spanish and Swedish. Other languages included in the grammar, namely Bulgarian, Greek, Latvian, Polish, Russian, and Romanian, were not evaluated, either because suitable evaluators were not available or because the grammar still lacked the implementation of some important constructs, e.g. 'if-then', at the time of the evaluation. English was used as the source language to be evaluated against. For comparison purposes, the English sentences were also machine translated using Google Translate.

Two native speakers of each target language were recruited for the evaluation. The evaluators were not familiar with CNL or ACE and were not told that the translations were automatically created or introduced to the involved translation technologies beforehand. The evaluation was carried out on Appraise [Fed12], a web-based open-source system for the evaluation of machine translation (MT). The evaluators were presented with a source sentence in English and two translation options, ACE-in-GF or Google Translate, into the target language. The order of the options was randomized by Appraise and the evaluators were not aware of the option used to generate each translation. The task was to choose the translation result they considered best and either accept it as-is or post-edit it as necessary, making only the minimal corrections needed. If neither translation result was considered good enough, the test subjects had the option of creating a translation from scratch. The evaluation system automatically records the choices made by the evaluators as well as the time taken for selection and post-editing.

After receiving the results from the evaluators, system-level automatic metrics were calculated using Asiya software [GM10] using the two reference translations created by the reviewers. The metrics used were BLEU [PRWZ02], NIST [Dod02], TER (Translation Edit Rate) [SDS$^+$06], WER (Word Error Rate), and PER (Position-independent word error rate). The agreement between the evaluators was also observed.

4.2 Selection of Evaluation Input

The multilingual translations of ACE sentences were generated by first compiling the ACE-in-GF grammar against the RGL built-in lexicon, and then using two GF tree generation methods to obtain the trees to be linearized multilingually. In the first method, a set of hand-picked ACE sentences were parsed into GF tree structures using the ACE concrete syntax of the ACE-in-GF grammar. In the second method, a set of GF tree structures were automatically generated based on the abstract syntax of the ACE-in-GF grammar.

Finally a manual post selection was applied to the combined output of both methods in order to prune out unwanted entries like repetitions. More details of the complete process are given below.

4.2.1 Lexicon

In order to cover more languages and avoid some lexical mistakes made during the evaluation reported in [CFK12] we chose to use the RGL built-in lexicon from which we extracted entries that correspond to the ACE content word classes, namely common nouns, proper names and transitive verbs, resulting in $\sim$300 words in total. Transitive adjectives were not included for this evaluation, as there were only two transitive adjective entries in the RGL lexicon. These words do not fall into any specific domain but are available in all RGL languages and their lexical entries have been tested extensively as part of the RGL development effort. This allowed us to concentrate on the evaluation of the ACE-specific syntactic structures, and not on the correct use of GF smart paradigms. Note however that these content words were automatically "plugged into" the sentences as explained below, without trying to model any semantic or pragmatic constraints. This resulted in odd sentences like 'Which airplanes win themselves?' or 'Paris is a bird and is a horse.'. The participants were instructed not to pay any attention to such violation of common sense and only focus on the syntactic aspects of the sentences and their translations.

4.2.2 Trees Based on the ACE Sentences of the ACE-in-GF Regression Test Set

Over the course of the ACE-in-GF project we have built up a set of ACE sentences that are used for regression testing to verify that after each grammar update there is no unwanted change in the ACE parsing correctness and multilingual linearization correctness. The main principles behind this set are to

1. cover all $\sim$90 functions of which the grammar consists, i.e. include sentences whose trees collectively contain all the functions of the grammar;
2. favour short and readable sentences that are likely to occur in real application scenarios;
3. cover ACE sentences that correspond to the main axiom structures (e.g. domain, range, transitivity, class assertion) found in formal ontologies.

This set is a superset of the sentences used in the evaluation of [CFK12] since it adds interrogative sentences, sentence negation and some more complex structures to achieve a complete coverage of the grammar. At the time of the evaluation the size of this set was 96 sentences.

We automatically replaced the content words in these sentences by content words of the RGL built-in lexicon as this gives a better guarantee for the morphological correctness of the linearizations. For example 'Bill' was changed to 'John' as the former is not available in the RGL lexicon. We then parsed and linearized all the resulting ACE sentences, thus obtaining their multilingual translations.

4.2.3 Automatically Generated Trees

The GF command line tool offers two ways to generate trees automatically: `generate_tree` for exhaustive generation and `generate_random` for probability-biased random generation. Both methods can start generating from a given start category, or from a given partial tree where the expansion of some nodes is unspecified and to be filled in by the generator. Both methods operate top-down and can be bounded by a depth-parameter. We found both methods unsuitable for generating a representative and easy to evaluate set of trees for the ACE-in-GF grammar. This might be due to the many recursive structures of this grammar. At low generation depths the coverage was very small, at higher depths the output got very noisy both in the sense of a lot of repetition and excessively large trees, whose linearizations cannot be used in a human-based evaluation setting.

To overcome these limitations we designed a new tree generation technique that combines a bottom-up partial tree generation with the top-down probability-biased random generation. This ensures that the resulting trees have high coverage, because the bottom-up generator is applied to every function of the

grammar, thus obtaining a partial tree for every function. This strategy also results in smaller trees and consequently shorter linearizations because the random generator starts from a sufficiently deep "backbone" of the tree and can be therefore run at lower depths.

We generated trees for every non-lexical function in the grammar up to the top-level categories of `Sentence` and `Question` in the following way:

1. A single partial tree was generated for every non-lexical function by tracing a path from the function to the given top-level category and then converting this path to its corresponding partial tree. The path is a sequence of pairs of the form (f, k), where f is a function symbol and $k \in \{1 \dots n\}$ is a position of its argument category in the simple type $C_1 \to \dots \to C_n \to C$, where C_i are argument categories and C is the value category. For each function on the path, its value category must be equivalent to the argument category of the following function at its specified argument position, or — if it is the last function on the path — equivalent to the given top-level category. In general, there can be many different paths but we chose only one by greedily preferring functions with a low number of arguments and low argument positions, breaking ties by random selection. Also, the path could not contain the same function more than once. This strategy excludes certain structures — e.g. double sentence negation that ideally should also be evaluated — but results in a more maintainable output in the form of a single reasonably small partial tree for each function.

2. Generic trees, that is to say partial trees that subsume another tree in the generated set, were removed. For example, the partial tree for the function that constructs sentences could be removed because this function is already represented in the partial tree of the function that constructs noun phrases as every noun phrase is part of a sentence. This pruning avoids some repetition in the output and makes post selection less time consuming.

3. All partial trees were completed by binding the unspecified nodes using probability-biased random generation. We used a depth level of 3 and a short hand-coded probability file that avoided some content words by setting their probability to 0, and gave a lower chance to indefinite pronouns, negation, and other structures that the grammar generated with a higher probability than what might be considered the default in natural language. Note that this step can decrease the coverage if a too low generation depth is chosen. We had to make a small compromise between the coverage and the size of the trees. Note also, that this step can reintroduce some structures that the partial tree generation step avoided by design.

This fully automatic process using only minor manual parameter tuning resulted in just 80 trees that were then linearized. The result is a representative and readable set of sentences that is suitable for evaluations or simply for introducing available syntactic structures to a new user of the grammar.

4.2.4 Post Selection

The two sets of trees generated with the above methods were combined and underwent a post selection step were some entries were removed because they still contained a lot of repetition — for example, 'There is a train that Paris does not hit and that Paris wipes and that Paris does not squeeze and that John does not learn.' — and then using random selection to reduce the amount to be suitable for the evaluation. We verified that the post selection did not reduce the full coverage of non-lexical functions that was present in the originally generated sentences. The final evaluation set had 111 treescorresponding to both declarative and interrogative sentences and having 3–18 words in their ACE linearizations.

4.3 Results

4.3.1 Automatic metrics

The system-level automatic metrics calculated from the sample are presented in Table 1. The best score of each metric is shown in bold. For BLEU and NIST, higher scores are better, whereas for the edit distance based scores (WER, TER, PER), lower scores are better. All metrics used measure the lexical level similarity of the translation suggestions and the reference translations.

Table 1: Automatic metrics: ACE-in-GF vs. Google Translate

	ACE-in-GF					Google Translate				
	BLEU	**NIST**	**TER**	**WER**	**PER**	**BLEU**	**NIST**	**TER**	**WER**	**PER**
Catalan	0.809	8.803	0.101	0.231	0.223	0.716	7.993	0.151	0.265	0.232
Danish	0.716	8.233	0.142	0.263	0.208	0.623	7.452	0.186	0.324	0.244
Dutch	0.899	9.335	0.056	0.223	0.158	0.735	8.371	0.133	0.275	**0.170**
Finnish	**0.948**	**9.336**	**0.026**	**0.147**	**0.132**	0.446	6.053	0.321	0.401	0.365
French	0.873	8.998	0.073	0.221	0.179	0.784	8.284	0.128	0.258	0.217
German	0.850	9.027	0.060	0.162	0.152	0.660	7.943	0.166	0.289	0.187
Italian	0.822	8.626	0.090	0.191	0.173	0.793	8.186	0.116	**0.204**	0.181
Norwegian	0.718	8.142	0.116	0.248	0.187	0.687	7.795	0.152	0.240	0.199
Spanish	0.788	8.835	0.095	0.224	0.198	0.708	7.994	0.167	0.281	0.212
Swedish	0.889	9.303	0.056	0.300	0.226	**0.794**	**8.723**	**0.093**	0.260	0.194
Average	**0.831**	**8.864**	**0.081**	**0.221**	**0.184**	**0.695**	**7.879**	**0.161**	**0.280**	**0.220**

All average scores for ACE-in-GF translations are better than the respective results for Google Translate. ACE-in-GF gets the best scores with Finnish, while Google Translate fares the worst. As far as ACE-in-GF is concerned, this is not surprising as the Finnish concrete syntax in ACE-in-GF — together with German and Spanish — has received more developer attention than the other languages.

4.3.2 Human evaluation

The translation suggestion that the evaluators preferred is presented in Table 2 and Figure 2. The numbers show whether both, either or neither of the two evaluators chose the ACE-in-GF translation suggestion or the exact same Google one as an acceptable translation or a post-editable one. For example, 87% of the ACE-in-GF translation suggestions in Finnish were accepted without editing by both or one of the evaluators, and 10% was chosen for post-editing. Only 3% of the Finnish Google suggestions were preferred as such or for post-editing and the ACE-in-GF suggestion rejected.

Table 2: Evaluator preference (total 111 sentences)

	Accepted ACE-in-GF				Post-edited ACE-in-GF				Pref. Google	
	Both	**One**	**Total**	**Acc. %**	**Both**	**One**	**Total**	**PE %**	**Either**	**Total**
Catalan	40	22	62	**56 %**	9	20	29	**26 %**	20	**18 %**
Danish	29	19	48	**43 %**	23	12	35	**32 %**	28	**25 %**
Dutch	44	38	82	**74 %**	9	3	12	**11 %**	17	**15 %**
Finnish	71	26	97	**87 %**	11	0	11	**10 %**	3	**3 %**
French	47	30	77	**69 %**	7	2	9	**8 %**	25	**23 %**
German	57	18	75	**68 %**	13	5	18	**16 %**	18	**16 %**
Italian	45	16	61	**55 %**	0	8	8	**7 %**	42	**38 %**
Norwegian	32	20	52	**47 %**	20	11	31	**28 %**	28	**25 %**
Spanish	27	25	52	**47 %**	27	26	53	**48 %**	6	**5 %**
Swedish	32	48	80	**72 %**	4	13	17	**15 %**	14	**13 %**
Average	**42.4**	**26.2**	**68.6**	**61.8**	**12.3**	**10.0**	**22.3**	**20.1**	**20.1**	**18.1**

4.3.3 Issues in human evaluation

The evaluators had some difficulties with certain issues in the ACE-in-GF translations. Even though the evaluators were instructed to ignore the content words of the sentences and focus on the syntax and morphology, some evaluators found the material hard to evaluate. For example,the lack of an ellipsisin the source sentences generated by ACE-in-GF was seen as a flaw in the translations. For example the sentence 'Everything that something finds is a horse or is a bird.' was usually translated with a more natural-sounding

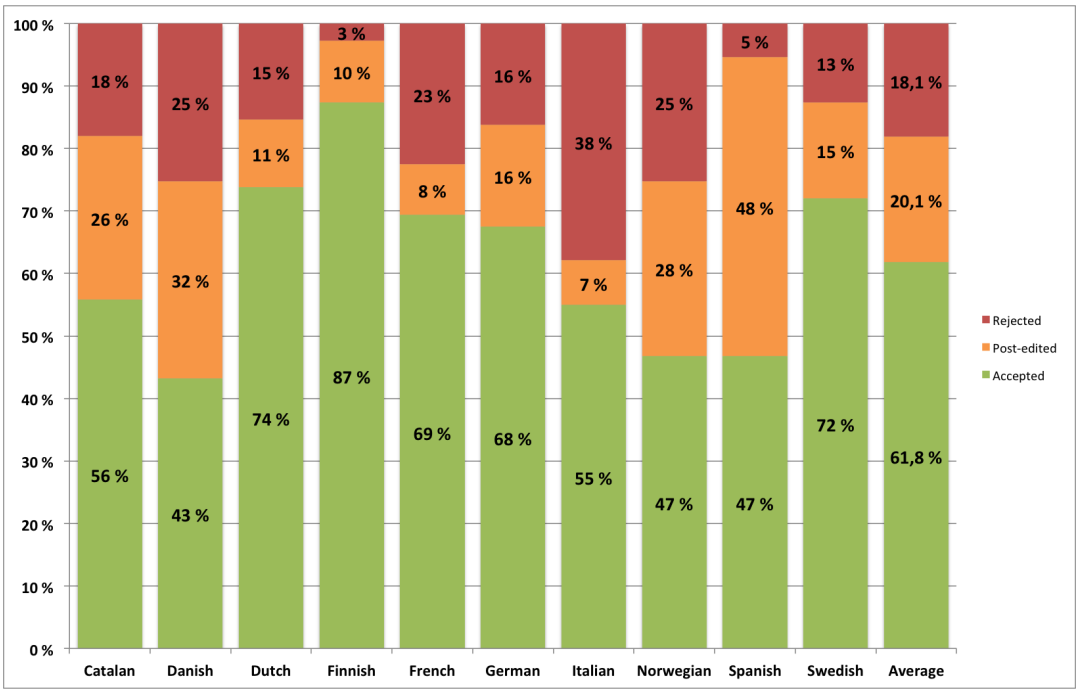

Figure 2:
Evaluator preference

ellipsis 'Everything that something finds is a horse or (is) a bird.' This had a negative effect on the scores of the GF translations, as Google Translate suggestions usually used an ellipsis in its translations.

As with all manual evaluation of translations, some choices made by the evaluators were purely subjective, for example the use of punctuation and using the active form instead of the passive. For example, the question 'What is seen by every dog?' was translated into 'What does every dog see?' by the evaluators in many cases. Also, for example one Italian evaluator chose to translate many sentences from scratch even when identical or nearly identical translation suggestions were available. This was probably due to unfamiliarity with the Appraise evaluation tool.

Even with these issues, it is evident that ACE-in-GF translation suggestions were preferred over the Google Translate ones. As the very low TER score of 0.081 for the ACE-in-GF sample shows, very little post-editing was needed to create acceptable translations.

5 Evaluation of AceWiki-GF

5.1 Introduction

AceWiki, on which AceWiki-GF is based, was shown to be useful and usable in previous work [Kuh13, Kuh09]. In this report we focus on the novel aspect of AceWiki-GF, namely that the wiki content is multilingual. The evaluation of AceWiki-GF tries to assess the quality of interlingual communication via multilingual wiki articles written in CNLs supported by the ACE-in-GF grammar. This is the hypothesis that we want to test for AceWiki-GF:

Hypothesis 5.1 *A group of users reaches almost the same level of agreement on the content of an article presented to them in different languages as when the article is presented to all of them in the same language.*

Concretely, we measured the degree to which users agree on the truth or falsehood of sentences written in AceWiki-GF. Basing the evaluation on the ACE-in-GF grammar lets us exploit the fact that ACE sentences have a clearly defined unambiguous interpretation regarding e.g. relative clause attachment, quan-

tification, and scoping of coordination and negation. Note that the translations into other languages should preserve this unambiguous interpretation. This behaviour of ACE-in-GF also makes the evaluation results for AceWiki-GF comparable to previous results obtained with the monolingual AceWiki. We also evaluated the usability of AceWiki-GF as a tool. Again, it is reasonable to use the ACE-in-GF grammar as the basis, because the original AceWiki system was based on a similar subset of ACE and the extension to AceWiki-GF adds very little to the original design as far as the user interface is concerned.

The following sections describe in detail the steps of the evaluation.

5.2 Design

The experiment is designed to evaluate the effectiveness, efficiency and usability of AceWiki-GF when users of different languages are collaboratively working on the same content. This includes the tasks of writing possibly false or incorrect sentences about a certain topic in the user's own language, reading and understanding sentences written by other users, and correcting mistakes made by other users in the form of identifying and deleting false sentences. When users work in different languages, the quality of the automatic translation is of course an important factor for the effectiveness of their communication. To determine whether two users understand each other, we ask them to write true and false sentences and to specify which sentences they consider to be true and which ones they consider to be false. This gives us a simple yet robust measure of the degree to which users can reach consensus, within and across languages. For example, if a sentence such as 'Every river flows into a lake or flows into a sea or flows into a river.' is perceived to be true by an English speaker, then after being automatically translated into another language, it should also be perceived as true by the speakers of the respective language. The cross-language understanding can then be evaluated by measuring the agreement on sentences being true or false. Thus, the experiment consisted of two tasks. In the main editing task a number of users create articles in their native language or in a language they are fluent in. In the post-editing task, users read in their respective language automatically translated articles written by other users and evaluate the truth or falsehood of the sentences.

In order to have a baseline for comparison, every participant accesses two articles during the post-editing task: one that was translated from another language, and one that was originally written in the same language, i.e. has not undergone any translation process. Even in the latter case, we cannot expect perfect agreement, as mistakes can never be completely avoided and, more importantly, people have different views on the world and tend to interpret certain sentences differently. The above exemplary sentence on rivers probably appears true on first sight to most readers, but it is false if one considers "endorheic basins", which are fed by rivers but are neither a lake nor a sea nor a river. By having the same article evaluated by a speaker of the language in which it was originally written and by another one that views an automatic translation, we are taking into consideration these natural disagreements, which may influence the results but have nothing to do with the AceWiki-GF system itself.

We developed a medium-size domain vocabulary covering the main topics of European geography. This domain was chosen because it was likely to be equally known to all possible participants. Also, objects, concepts and relations in this domain can be effectively illustrated in the form of a map, a medium also likely to be familiar to everybody. The vocabulary was developed for English, German and Spanish. The small number of languages compared to what ACE-in-GF can potentially offer is mainly caused by the fact that building large multilingual vocabularies is very time consuming. Furthermore, we had to make sure that we are able to recruit a sufficient number of participants for each of the chosen languages. Yet, having a linguistic scenario based on controlled structures will make it easy and reliable to extrapolate the results to other languages.

The participants were asked to enter both simple, existentially quantified sentences and complex, universally quantified sentences. Examples of the first type are 'The Limmat flows through Zurich.' and 'Zurich is not a capital of Switzerland.'. These are sentences that always mention specific domain objects like 'Zurich' or 'Limmat', and their information can in general easily be extracted from a map. Examples of the second type are 'Every country that does not border a sea is a landlocked country.' and 'If a river X flows into a river Y then Y does not flow into X.'. Such universal statements express generally accepted knowledge, or reference collections of domain objects like 'country' or 'river'. We were interested

in how our multilingual grammar for ACE-in-GF and AceWiki-GF in general perform with both types of sentences.

Candidates for the evaluation were recruited via university mailing lists from different European countries and by word-of-mouth spreading without any restrictions of sex, age, or background. Candidates were only required to have a good command of one of the three languages involved in the evaluation — English, German or Spanish — and to be somewhat familiar with computers so that they could easily get the gist of the wiki and of controlled languages. The communication with the evaluation participants occurred via email and the experiment was performed via a dedicated AceWiki-GF instance that was available online. There were no strict time limits given to the participants.

The wiki was presented to each participant as a monolingual environment, i.e. participants could see the wiki only in their own language. This is to ensure that the GF-based automatic translation is the only way to communicate across languages. The participants were working in a localized version of the wiki, i.e. not just the wiki content but the entire graphical user interface was shown in their language.

Finally, we designed a questionnaire to gather feedback from the participants, mostly as free-form comments. Such feedback often highlights issues which cannot be detected by a fully automatic log file analysis.

5.2.1 Lexicon

We developed a 500-word domain vocabulary covering the main topics of European geography. The vocabulary was developed for English, German and Spanish, and consisted of:

1. common nouns such as ‘country’ and ‘language’, internally mapped to OWL concepts (named classes) (13% of the total vocabulary);

2. proper nouns such as ‘Germany’ and ‘Danube’, mapped to OWL objects (named individuals) (83%);

3. transitive verbs and transitive adjectives such as ‘borders’, ‘capital of’, ‘to the east of’, ‘flows through’, mapped to OWL relations (named object properties) (4%).

The vocabulary was limited to the geography of Europe, and we tried to cover all countries, their capitals, languages and major natural features.

Geographical names tend to be quite ambiguous. For instance, the word ‘Etna’ can denote a volcano in Sicily, a river in Norway, or several settlements in the United States. AceWiki-GF offers a disambiguation dialog that allows the users to choose the intended reading of an ambiguous entry ([FKK13]). This works most effectively if the grammar contains a set of “disambiguation languages” [RED12] with explicit lexical entries like ‘Etna (river)’ for each visible concrete natural language. To reduce the overall effort of the lexicon editing and in order not to complicate the evaluation by including the disambiguation dialog, we decided to avoid lexical ambiguity in the grammar in the first place. We also removed cross-category ambiguity, e.g. the word ‘Bergen’, which is the name of a Norwegian city as well as the dative plural form of the noun ‘Berg’ (‘mountain’) in German. In most cases though such ambiguity does not manifest itself in actual sentences as words from different categories do not typically occur in the same context.

Regarding the support for synonymous names for a given geographical entity, we relied on the GF variants-feature. In most cases this was used when we were unsure which lexical form to include between several possible candidates, e.g. ‘Balaton’ vs ‘Plattensee’. In this case all forms were added as variants of each other. However, in the completed lexicon this feature was used only in 21 entries.

We note that one could develop a more principled guideline with respect to ambiguity and the use of variants, e.g. to always include a variant to make sure that one of the linearizations of the entry is unambiguous. In the geography domain this can be achieved to a large extent by including the type in the entry of the geographical object, for instance ‘the river Etna’ vs ‘the volcano Etna’ which in some languages already happens naturally, e.g. ‘el río Etna’ in Spanish. The unambiguous form can then be used to help to disambiguate the ambiguous form ‘Etna’. However, this would mean that the type assignment happens already at the level of the lexicon and not at the level of sentences — such as ‘Etna is a volcano.’ — as the current functioning of AceWiki and AceWiki-GF pre-supposes.

Table 3: Lexicon table with 6 entries. The first column lists unique keys of the lexical entries and maps directly to the GF function and the OWL entity of the entry. The second column lists the ACE-in-GF categories of the entries and also determines the OWL entity type (class, individual, object property). The rest of the columns provide the encoding of the lexical entries as calls to GF smart paradigms for all the supported languages. Such a table was collaboratively edited in Google Spreadsheets and automatically converted to the GF grammar. For convenience reasons it was possible to deviate from the strict notation of the smart paradigms or even leave the table cell empty. The converter filled in the missing parts on the basis of the first two columns and mapped our ad hoc convenience conventions to the correct paradigm calls. The encoding of the Spanish entry for the relation 'border' demonstrates a GF variant (|), i.e. this entry can be linearized as a form of 'limitar con' or 'hacer frontera con'.

Entity	Cat	ACE (English)	German	Spanish
located in	`A2`	aceA2 (mkA "located") (mkPrep "in")	mkA2 (mkA "platziert") (mkPrep "in" dative)	mkA2b (mkA "situado") (mkPrep "en")
longer than	`A2`	aceA2 (mkA "longer") "than"	mkA2 (mkA "länger") (mkPrep "als" nominative)	mkA2 (mkA "más largo") (mkPrep "que")
capital	`CN`		aceN die "Hauptstadt" "Hauptstädte"	"capital" feminine
Semitic language	`CN`		aceN die "Semitische Sprache"	"lengua semítica" "lenguas semíticas" feminine
Danube	`NP`	defsg "Danube"	die "Donau"	el "Danubio"
border	`V2`	"border" "borders" "bordered"	prepV2 (regV "grenzen") (mkPrep "an" accusative)	mkV2 (mkV "limitar") (mkPrep "con") | mkV2 (I.hacer_V) (mkPrep "frontera con")

The lexical entries were written as calls to GF smart paradigms, i.e. operators in the form of a constructor (e.g. `mkA2`, `mkV2`) with arguments that are strings such as the dictionary forms of the words, and categories like `feminine`. We extended the smart paradigm notation by ad hoc language-specific shorthand operators, e.g. the gender of geographical names could be tagged by definite articles in German and Spanish as in 'die "Donau"' and 'el "Danubio"'. This improved the productivity of the lexicon editors as it was more natural to work with language-specific labels than with universal categories like "feminine" or "dative". However, the latter option remained available. Table 3 shows an example of six terms belonging to the available categories. The entries in such a table were automatically converted to a GF grammar. To make the lexicon editing even simpler, the operator names could sometimes be omitted. In this case the conversion script derived them automatically from the category tag.

The vocabulary was collaboratively built by three people including native speakers of German and Spanish and a GF expert. AceWiki-GF supports lexicon editing but currently only in a very basic form as GF source editing. For that reason, the lexicon was developed using a shared Google Spreadsheets table which offered a structured view to the lexicon where rows represented lexical entries, columns represented languages and table cells contained GF smart paradigm calls for the lexical entries. This environment offered some collaboration features such as simultaneous editing and version control which allowed us to quickly build up a rather large vocabulary. However, since this environment had no special support for the GF language — e.g. in the form of auto completion and syntax highlighting — the feedback on GF syntax errors or wrongly used smart paradigms took long: the spreadsheet had to be downloaded, converted to the GF grammar, compiled, regression tested, and uploaded to a test wiki. We wrote a number

of tools to automate this process — see the project website[3]— but an integrated grammar checker in such collaborative lexicon editors would still be necessary, especially for linguists who are less familiar with GF and the structure of the RGL API.

The vocabulary was integrated with the ACE-in-GF grammar and loaded into a test wiki on a regular basis. This allowed for an easy checking of the entries in actual sentences, e.g. to determine if the automatically generated forms like the plural form of nouns and singular forms of verbs were correct.

Constructing a domain-specific lexicon for use in a CNL environment has advantages, such as the reduction of ambiguity and synonymy. It was not difficult to provide the terms in the three languages, since mostly only one translation was possible. For example, 'rise in' can only be translated to Spanish as 'nacer en' with regards to rivers. However, we also encountered some difficulties, mostly with the syntactic realization of relations across languages. For example, 'located in' was translated into German as 'platziert in' in order to allow for the construction with the verb 'to be', which is acceptable, but not pragmatically used in German. Also, the transitive verb 'drain' ('A river drains a lake.') sounds natural in English but does not have a counterpart in the form of a simple transitive verb in German and Spanish, so it was decided to remove it from the evaluation. We decided to split the transitive verb category to two, separating out transitive verbs which do not allow for a passive form. These more restricted verbs were mostly used to encode the domain relations because of the many prepositional verbs in German and Spanish, which do not allow for a passive form. We also created an additional operator for Spanish adjectives in order to distinguish between the 'ser' and 'estar' forms of 'to be'.

We expect that such issues with the encoding of domain relations also come up with other languages, slowing down the increase in the language coverage of the lexicon as the described encoding issues require the attention of a GF expert. Then again, the amount of relations in such ontological applications is naturally small as most of the vocabulary is built up from objects and concepts.

5.2.2 Setup

In order to obtain clearly interpretable evaluation results the functionality of the AceWiki-GF environment was customized for the evaluation, which mostly resulted in disabling certain features:

1. only the grammar-backed content was allowed in the wiki, i.e. it was not possible to enter free-form comments;
2. the grammar could not be changed, i.e. users had to work with the available vocabulary and available syntactic structures;
3. the wiki language could not be changed, i.e. users could read/edit the wiki in only one language, i.e. they could not observe how their contribution was interpreted in other natural and formal languages such as OWL or GF tree structures;
4. the OWL reasoning capability of the wiki was disabled;
5. users had to register and log in so that everybody's contribution could be identified and tracked;
6. direct collaboration with other users was forbidden (participants were instructed to edit only their own pages);
7. using the "Assert/Retract" flag in AceWiki-GF, each sentence could be tagged as "true" or "false", which renders the sentences in black or red, respectively;
8. the wiki was configured to automatically disambiguate ambiguous entries by always picking the first reading. The ACE-in-GF grammar allows by design for very little ambiguity which made it unlikely that many users would be exposed to the wiki disambiguation dialogue, so we decided to exclude this feature from our evaluation.

In addition, the coverage of the ACE-in-GF grammar was somewhat reduced for the evaluation purposes:

[3] https://github.com/Kaljurand/GF-Utils

1. multi-sentence statements were disabled (AceWiki-GF originally allows more than one sentence per entry);
2. question sentences were disabled because we wanted to base the evaluation on the agreement over the truth value of the sentences (which questions do not have);
3. animate indefinite pronouns ('everybody', 'nobody', 'somebody') were removed as these do not fit the inanimate domain of geography;
4. the quantifier 'for every' and the reflexive construct 'itself' were removed as their implementation contained bugs in German and Spanish.

The grammar was fixed prior to the evaluation and was not changed during the evaluation.

5.2.3 Procedure

Participants were divided into three language groups of equal size: English, German, and Spanish. Each member of such a group had a good command of the respective language. The procedure for the participants consisted of the following four steps:

1. Introduction
2. Main editing task
3. Questionnaire
4. Post-editing task

In the introduction step, the participants of the evaluation were presented with a brief description of the evaluation experiment, including an overview of the sequence of tasks they would have to carry out. Other than a description of the tasks, the real aim of each task was not known to the participants, nor were they aware of what exactly was under evaluation. The description of the main editing task and a link to the questionnaire were provided to the participants in the introduction, but the detailed description of the post-editing task was provided only after the participants had filled out the questionnaire, i.e. when we were sure they had finished the main editing task. The introduction was also meant to familiarize the participants — most of whom we expected to have no prior knowledge of ACE nor AceWiki — with the concept of controlled natural languages and the AceWiki-GF environment, to the extent that they would be able to successfully participate in the evaluation. The participants were asked to watch a 6-minutes-long screencast of the AceWiki-GF environment which contained instructions for the registration and login procedure, basic article creation and sentence editing. Also the domain of the wiki — European geography — was introduced, pointing out some existing sources that describe this domain in the form of maps and encyclopedic articles.

The main editing task asked the participants to create a new wiki page and write at least four true sentences for which examples were provided, such as 'Every river that rises in Switzerland drains into the North Sea or drains into the Mediterranean.', and at least four false sentences, e.g. 'Every country that borders Germany is a member of the European Union.', and to tag them explicitly as "true" or "false" using the Assert/Retract flag. They were additionally encouraged to use some more complex structures and expressions, and they were allowed to edit and delete their sentences as often as they wished during the period of the experiment. In general, they were given freedom with regard to how many sentences to write or which words to use in their sentences, with the understanding that they could not create new words.

Once the main editing task was completed, the participants were presented with a questionnaire asking for their background, impressions of the system and of the constraints enforced by the controlled environment. Specific questions were asked about the user-friendliness of the look-ahead editor and the complexity of the sentence formation task.

For the final post-editing task, the participants were directed to two articles written by other participants and were asked to remove all false sentences in these articles. They were instructed to ignore minor syntactic errors possibly present in the sentences and only remove a sentence if it was clearly false in terms

of its meaning, as explained above. Both articles were copies of articles that other participants had produced, with a randomized order of the sentences and with their true/false color-coding removed. For each participant, one of the articles was originally written its own language and the other was translated from another language. The participants did not know which was which (in fact, they did not even know that there was such a difference). Half of the participants did the post-editing on the translated article first; the other half started with the article that did not involve translation. In this way, each article of the main task was post-edited exactly twice: once with and once without translation. The distribution of the six possible translation directions between the given three languages was perfectly balanced.

5.3 Results

The results of the experiment were collected in the form of the final wiki content, system logs that also registered edits/deletions of sentences that otherwise do not end up in the wiki content, and feedback via the questionnaire. We analyzed these data in a number of different ways.

5.3.1 General Numbers

We had ten participants for each of the three languages, i.e. 30 participants in total. The ratio of female to male participants was 14 to 16. They spent on average 37 minutes using AceWiki-GF, creating in total 316 sentences — not counting two sentences that led to an internal error. Of these, 171 sentences were marked as true and 145 as false. Therefore, each participant wrote on average 5.7 sentences marked as true and 4.8 sentences marked as false.

5.3.2 User Agreement

The degree of agreement between participants from the same and different languages is the main outcome of this experiment. To get a feeling for why participants might disagree on the truth or falsehood of a statement even in the absence of translation, let us have a look at an example. One of the participants wrote (in German) 'Every mountain contains a valley.' and marked it as true. Both post-editing participants (accessing it in German and English, respectively), however, deleted this sentence, thereby stating that they thought it to be false. Apparently, the participants had different views of the world and interpreted at least some of the terms 'mountain', 'contains', and 'valley' differently. The first participant might have thought that every mountain must have a certain size (otherwise it would just be a hill), is a geographical entity that is subject to rain and erosion, and therefore must have some trenches that, in this participant's view, are to be called valleys. The other participants, in contrast, might have followed the line of argument that there is no logical necessity for a mountain to have valleys, or that small mountains only have trenches that are not big enough to be called valleys, or that the verb 'contain' does not apply to mountains and their valleys. In any case, this example shows that this kind of disagreement can easily arise without any translation process involved.

Coming back to our hypothesis, we wanted to show that translation does not significantly lower the level of agreement. The agreement level can be defined as $(T_k + F_d)/S$, where S is the total number of sentences, T_k is the number of sentences originally marked as true and not deleted in the post-editing, and F_d is the number of sentences originally marked as false and deleted in the post-editing. In the case where the same language was used during main task and post-editing — meaning that there was no translation — the average agreement level was 82.2%. This means that the respective participants disagreed on 17.8% of the sentences with respect to whether they were true or false. In the case where post-editing was performed in a different language — meaning that there was a translation process — the average agreement level was 84.0%. That means that the agreement level was even slightly higher with translation, but we have every reason to assume that this is just a statistical artifact, because the difference is not significant at all: We get a p-value of 0.87 when applying a Wilcoxon signed rank test to compare the two samples. The null hypothesis that the two samples come from identical distributions cannot be refuted. Therefore, the level of agreement is about the same in our two samples, which is consistent with our hypothesis.

The above calculation, however, does not prove that there is no difference between the two scenarios: It only shows that we have so far no reason to assume that they are different. In fact, no statistical test can

prove that two samples come from the same distribution, as no sample size can eliminate the possibility of an arbitrarily small difference of the distributions. For that reason, we need to come up with a more specific hypothesis that is testable. First, let us assume that translation introduces a constant translation error rate r with $0 < r < 1$ that has the effect that the level of agreement between two given users is $(1 - r) \times a$ if translation is involved, where a would have been the level of agreement under the same circumstances if no translation was necessary. With this definition we can state a more specific hypothesis:

Hypothesis 5.2 *The translation error rate for AceWiki-GF is less than 5%.*

Under these assumptions, we can use our sample that did not involve translation and calculate the agreement level it would have produced with a translation error rate of 5%. This we can now compare to the original sample that involved translation, and we can evaluate with a one-tailed test whether we can refute the null hypothesis that the translation error rate is 5% or more. Doing this with a one-tailed Wilcoxon signed rank test gives us a p-value of 0.046. This means that we can reject the null hypothesis, which verifies our hypothesis that the translation error rate of AceWiki-GF is less than 5%.

Whether a sentence was initially marked as true or false did not have a large effect on the level of agreement. Sentences originally marked as true were kept in the post-editing phase in 81.8% (without translation) and 82.7% (with translation) of the cases; sentences originally marked as false were deleted in 84.1% (without translation) and 84.7% (with translation) of the cases.

5.3.3 Syntactic Features

We were interested in the syntactic features of the 316 original entries, specifically the frequency distribution of their words and grammatical constructs and whether the distributions differ depending on the language and the true/false tag. Our main concern was that certain grammatical constructs are not easily usable or easily discoverable in German and Spanish, compared to ACE. This is a reasonable hypothesis because ACE has been carefully designed over many years, while its German and Spanish counterparts were obtained almost automatically and little effort has so far gone into fine-tuning them.

This analysis was done in a completely language-neutral way by looking at the underlying tree of every entry. This is possible because the functions in the ACE-in-GF abstract syntax are relatively syntactic in nature, referring to syntactic objects like “relative clause” and “noun phrase”, rather than semantic objects like “relation” or “concept”. In other words, this means that if an ACE user chose to express a sentence with a relative clause then the readers in German and Spanish will also see a relative clause, just in their respective languages. In a more semantic grammar such syntax-based analysis of the accessibility of the language features would not work, and one would have to look directly at the logical form as we do in the next section.

There were 315 unique trees, i.e. only one tree was repeated, which happened in the same language, namely German. It turned out that almost all non-lexical functions of the grammar were used by the participants. The exceptions were “variable in apposition”, e.g. ‘a country X’, and negated object relative clause, e.g. ‘that a country does not border’. This is not surprising as these constructs are conceptually relatively complex, and furthermore, can be rephrased by simpler constructs that preserve their meaning.

Looking at the most frequently used constructs no unusual distribution of grammatical constructs per language was observed, apart from two cases:

1. The German users avoided relative clauses much more that the ACE and Spanish users. This can be explained by looking at the orthographic conventions of German relative clauses, how they were implemented in the German grammar, and how the look-ahead editor presented them. Relative clauses are always surrounded by commas in German, the grammar however required the omission of the final comma. Also, as the look-ahead editor always presents single follow-up tokens, the users could only see a comma as an indicator of an available relative clause, in this case however it would be more user-friendly to show more of the possible follow-up context, i.e. a comma followed by the possible relative clause pronouns. These discrepancies might explain the low usage of relative clauses in German.

2. The sentences of the ACE users contained very few definite noun phrases ('the country', 'the capital of'). This is probably due to a bug which caused the look-ahead editor to classify the token 'the' as a proper name (because it also occurred as part of proper names such as 'the EU'). This misclassification made this token less discoverable in the look-ahead editor and the users who relied on the clicking on the category boxes for the construction of sentences were less likely to use definite noun phrases.

When looking only at the sentences marked as true, we did not observe a different distribution. This means that the sentences marked as "false" were also syntactically normal, i.e. users did not generate false sentences by just randomly clicking on the look-ahead editor word selection lists.

We conclude that the use of the different languages did not result in widely different sentence patterns, and that the possible deviations can be effectively discovered by the wiki developers using the described comparison of construct distributions in different languages. The issues that we noticed can be easily fixed by modeling the German orthographic conventions more closely and by generalizing the look-ahead editor to propose more context.

From the 500 entries in the lexicon the participants used 230 entries. The frequent content word vocabulary also differed across languages, but in an unsurprising way — German users wrote more about Switzerland and Germany, while the Spanish users more about Spain.

5.3.4 Semantic Features and Mapping to OWL

We analyzed the semantic properties of the resulting wiki entries by looking at their OWL mapping (in a few cases it resulted in a SWRL rule, which in the following we consider as a subtype of an OWL axiom). We were interested in the types of axioms that resulted and the cases where the mapping to OWL failed. Note that we cannot report how effectively does the OWL mapping reduce ambiguity (i.e. do away with the possibly spurious ambiguity in the entries) because the evaluation setup did not permit any tree ambiguity to begin with.

From the total of 316 wiki entries 4 were not ACE-compatible, i.e. due to some bugs, the ACE-in-GF grammar parsed or generated sentences which are not correct ACE (i.e. they were not accepted by the reference implementation of the ACE parser (APE, the project website[4]) which is embedded in the wiki. Examples include: verb phrase coordination with multiple elements was encoded with a comma for all but the last element (but ACE requires an explicit 'and' or 'or' token as a coordination separator) and multi-digit numbers were not correctly tokenized.

There were 26 entries (including the 4 non-ACE entries) which could not be mapped to OWL, i.e. the grammar does not always correctly model the OWL-compatible subset of ACE. The main reasons include: sentence disjunction which cannot be directly represented in OWL as it does not support axiom disjunction, cardinality constraints together with the 'of' relation some forms of which the ACE→OWL translator does not support, odd sentences with respect to anaphoric references, e.g. the 'then' part of the 'if-then' sentence does not anaphorically reference the 'if' part (which in general cannot be represented in OWL). The SWRL statements were typically nonsensical, e.g. 'Everything is everything.' and 'If Etna does not lie in Belgium then Italy contains the mountain.'. AceWiki-GF as well as the original AceWiki is designed to handle such OWL mapping deficiencies by highlighting the failing entries and pinpointing the source of the failure by an error message, so that the participants can rephrase the sentence, often in a semantically equivalent way, or otherwise delete them as nonsensical. From the entries that were marked as "true" only one entry failed to map to ACE and 11 entries failed to map to OWL. Table 4 shows the frequency distribution of OWL axiom types. 32% of the axioms correspond to the universal statements that the evaluation participants were encouraged to also enter.

When the true sentences were loaded into the wiki with reasoning turned on, then the knowledge base was found to be consistent, meaning that the sentences agreed with each other logically. The sentences that failed to map to OWL did not participate in this reasoning, this included the 4 SWRL statements. However, the wiki found also no inconsistencies if the complete set of sentences were loaded. This is also understandable as the set of sentences was small and the users were not coordinated to edit the same articles and

[4] `https://github.com/Attempto/APE`

Table 4: The OWL axiom distribution by type was (listing only types that occurred more than once)

Axiom type	All	True	Comment
ClassAssertion	104	54	Usually a simple statement asserting an individual into a class (‘Limmat is a river.’)
ObjectPropertyAssertion	92	60	Usually a simple statement asserting a relation between two individuals (‘The Aar flows into the Rhein.’)
DisjointClasses	44	22	General statement asserting a disjointness of two classes (‘No language is a country.’)
SubClassOf	41	19	General statement relating two classes (‘Every capital is a city.’, ‘Every river that flows into a lake is ...’)
SWRL rule	8	4	General statements that failed to map to OWL but could be mapped to SWRL. Typically similar to SubClassOf-statements but with a more flexible anaphoric reference structure.
Total	290	160	

write about the same objects thus there was little chance that they would write contradictory statements, especially considering that with the OWL-based open-world reasoning (even with Unique Name Assumption which the wiki enforces) some types of inconsistencies are not immediately captured.

5.3.5 User Feedback in the Questionnaire

Apart from providing information on their background, the participants had to answer the following three questions in the questionnaire.

1. Was AceWiki Geography easy or difficult to use in general?
2. Was the sentence editor easy or difficult to use?
3. Was creating true and false statements easy or difficult to perform?

This gave us a quantitative measure of the participants’ subjective experience with the wiki. For each of these questions, participants could choose from “very difficult” (value 0), “difficult” (1), “medium” (2), “easy” (3), and “very easy” (4). For all three questions, the average answer was close to but slightly below “easy”: 2.93 for the first question, 2.77 for the second, and 2.70 for the third. Given that it was indeed a rather complicated task involving a powerful tool, we consider these results very satisfactory.

Participants could also give free-form feedback in the questionnaire. Unsurprisingly, around 80% of the participants reported that the controlled environment did not let them express everything that they had in mind. From their feedback, it seems that once participants had decided which sentence to write, they occasionally hit against the wall of the controlled environment where coming up with a syntactically acceptable formulation was not always straightforward. Even users who stated having used AceWiki before, reported this obstacle. The main issue was the lack of content words that the participants wanted to use, for example ‘European’, ‘ocean’, ‘hill’, ‘Great Britain’. The sentence structures which the participants claimed they could not write included “repetitive structures”, such as the coordination of phrases (e.g. ‘The Danube flows through Germany, Austria and Hungary.’, ‘Die Hauptstadt von Portugal ist nicht Lissabon oder Madrid.’); adversative constructions (e.g. ‘Norway borders Sweden but not France.’); peripheral arguments of verbs (e.g. ‘There is a volcano in Switzerland.’, ‘French is a language spoken in France.’); triadic relations (e.g. ‘X lies between Y and Z.’); and comparative constructions (‘more than’). The lack of repetitive structures was reported most often and by speakers of all languages involved.

In most cases the grammar simply lacked the reported structures. Obviously, the 500-word lexicon did not cover all words needed to describe the European geography. Some of the syntactic structures have been excluded in ACE by design, mostly because they feature structural ambiguity that might not be always

visible to the users, but that nevertheless would make a deterministic mapping to a formal logical form impossible. In a few cases, however, the constructs were available but simply not easily discoverable (see more in Section 5.4).

The participants also reported that the categorization of some words in the look-ahead editor was confusing, such as the preposition 'in', which appeared in the category "adjectives", and the definite article 'the' that appeared in the category "proper names".

20% of the participants reported technical issues using the evaluation wiki, such as website crashing, needing to reload the webpage, and the look-ahead editor not registering mouse clicks for some seconds. However, these were reported to be only temporary hiccups from which the wiki engine recovered on its own. We therefore believe that these technical issues did not affect the evaluation.

5.4 Discussion

In our evaluation setting, all participants had the task to fill an initially blank wiki article with information. Of course, this does not accurately model the normal scenario where most of the users just read the wiki content and those few who additionally contribute were originally also readers, i.e. they have already gained knowledge of the domain, editing tools and conventions of the wiki, and — in the case of semantic wikis — learned the syntactic and semantic structures available in the content language. Therefore, some of the issues highlighted in the previous sections are likely not to occur or will be less relevant in a normal setting.

To summarize our main findings we use the handling of conjunction as a representative case. Both evaluations highlighted sentences that contain lists of domain entities, such as the sentence 'The English Channel separates France and England.' that we will use as example. Some participants wanted to write this or similar sentences containing an ellipsis, but could not. The reason is that ACE does not use ellipsis. In full English our example sentence can either be interpreted as containing a verb phrase conjunction where the second occurrence of the verb 'separates' is elided, or as containing the noun phrase conjunction 'France and England'. The intended meaning of this sentence, namely that the English Channel lies between France and England, is in English inferred from the lexical semantics of the words and from common knowledge. ACE, however, does not use semantics or common knowledge for disambiguation, but only structural means. That is to say that ACE distinguishes between 'The English Channel separates France and England.' where 'France and England' are interpreted as conjunctive noun phrase, and 'The English Channel separates France and separates England.' containing a conjunction of verb phrases. This principle is carried over to all other languages in the ACE-in-GF grammar. However, as ACE-in-GF does not support noun phrase conjunction, our example sentence is not accepted. A further complication of the wiki environment is that the look-ahead editor does not clearly indicate the absence of the construct "noun phrase conjunction" since the supported sentence conjunction looks identical to noun phrase conjunction if only the beginning of the sentences is visible. In other words, participants wanted to finish 'The English Channel separates France and England' with a period, but this was not offered because 'England' is the first word of a new conjoined sentence. In such situations, the participants felt "stuck" and had to "backtrack" losing much or all of their work done so far.

For the evaluation, we perhaps better should have disabled sentence conjunctions. However, as a general solution, the look-ahead editor technology could be generalized to offer more context, e.g. list the most frequent completions starting from the given position. Such a frequency analysis could be performed on the basis of the existing entries in the wiki. Another way to involve the existing wiki content into the editor is to automatically learn a fragment of the grammar from the existing entries to offer a simpler subset of the grammar to novice users. Yet another approach is to augment a CNL grammar with an explicit notation which avoids the discussed type and scope ambiguities. This approach is valid if the additional notation can be effectively explained to the users and made available multilingually, i.e. it should not use English keywords.

The main user-interface component with which the participants interacted is the look-ahead editor. Some of its issues derive from the fact that the original look-ahead editor of AceWiki was ported from the monolingual ACE setting to the more complex multilingual ACE-in-GF setting. For example, the GF look-ahead technology currently does not provide all features that the wiki requires, since it only returns a list of possible completions for the partial sentence but does not categorize them by word class. Since the AceWiki user interface requires such a categorization, we implemented a heuristic solution on top of the

GF look-ahead output which however misclassified certain words making these words effectively hidden, which negatively affected the editing. This problem can be fixed by implementing the categorization at the level of the GF parser.

6 Conclusions

The main goal of the work described in this report was to verify that an AceWiki-style semantic wiki could in principle function also multilingually. We have developed a multilingual grammar ACE-in-GF that maps ACE to fragments of many natural languages, and extended the AceWiki engine to support multilinguality, resulting in AceWiki-GF, which runs on a multilingual GF grammar, such as ACE-in-GF. We designed two separate evaluations. The first evaluates the correctness with which the ACE-in-GF grammar automatically translates ACE sentences. The second uses ACE-in-GF as the underlying grammar in the AceWiki-GF environment to evaluate its usability and effectiveness. The results of both evaluations are encouraging — AceWiki-GF allows users to cooperatively and concurrently build-up a shared and agreed-upon formal knowledge base in different controlled natural languages. Users accessing the wiki in different languages can reach a shared understanding of the content that has approximately the same quality as for users of the same language. Our evaluation shows that, with AceWiki-GF, only very little is lost in translation.

Most of the issues that the evaluation highlighted are easily put right in either the wiki engine or the underlying grammar, and the wiki content itself can be used to pinpoint these issues as one can — mostly automatically — analyze and compare how different languages are used to produce the content. Some issues however remain as open research questions, such as finding the most effective way of making new users familiar with the controlled environment. We are interested in ways in which the existing wiki content and the existing user experiences in the wiki can directly and automatically feed a concise user-level documentation of the grammar and the domain of the wiki, optimally integrated into the wiki itself.

As far as the evaluation itself is concerned, it could be repeated with more users, running over a longer period of time and by including more AceWiki-GF features — such as community based lexicon extension, ambiguity management and automatic reasoner feedback — which were disabled in this evaluation in order to make it more manageable. The analysis of the results could also include a more elaborate analysis of the wiki log files which show among other details how often the participants had to backtrack or how often they used auto completion. In general, however, we believe that the evaluation design could remain largely unmodified.

References

[CFK12] John J. Camilleri, Norbert E. Fuchs, and Kaarel Kaljurand. Deliverable D11.1. ACE Grammar Library. Technical report, MOLTO project, June 2012. `http://www.molto-project.eu/biblio/deliverable/ace-grammar-library`.

[Dod02] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[Fed12] Christian Federmann. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September 2012.

[FKK08] Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Attempto Controlled English for Knowledge Representation. In Cristina Baroglio, Piero A. Bonatti, Jan Małuszyński, Massimo Marchiori, Axel Polleres, and Sebastian Schaffert, editors, *Reasoning Web, 4th International Summer School 2008, Venice, Italy, September 7–11, 2008, Tutorial Lectures*, number 5224 in Lecture Notes in Computer Science, pages 104–124. Springer, 2008.

[FKK13] Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Deliverable D11.2. Multilingual semantic wiki. Technical report, MOLTO project, January 2013. http://www.molto-project.eu/biblio/deliverable/multilingual-semantic-wiki.

[GM10] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86, 2010.

[Gro12] W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012. Technical report, W3C, 2012. http://www.w3.org/TR/owl2-overview/.

[Kal07] Kaarel Kaljurand. *Attempto Controlled English as a Semantic Web Language*. PhD thesis, Faculty of Mathematics and Computer Science, University of Tartu, 2007.

[KK13] Kaarel Kaljurand and Tobias Kuhn. A multilingual semantic wiki based on Attempto Controlled English and Grammatical Framework. In *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013)*. Springer, 2013.

[Kuh08] Tobias Kuhn. AceWiki: Collaborative Ontology Management in Controlled Natural Language. In *Proceedings of the 3rd Semantic Wiki Workshop*, volume 360. CEUR Proceedings, 2008.

[Kuh09] Tobias Kuhn. How Controlled English can Improve Semantic Wikis. In Christoph Lange, Sebastian Schaffert, Hala Skaf-Molli, and Max Völkel, editors, *Proceedings of the Fourth Workshop on Semantic Wikis, European Semantic Web Conference 2009*, volume 464 of *CEUR Workshop Proceedings*. CEUR-WS, June 2009.

[Kuh13] Tobias Kuhn. The understandability of OWL statements in controlled English. *Semantic Web*, 4(1):101–115, 2013.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Ran09] Aarne Ranta. The GF Resource Grammar Library. *Linguistic Issues in Language Technology*, 2(2), 2009.

[Ran11] Aarne Ranta. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford, 2011. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

[RED12] Aarne Ranta, Ramona Enache, and Grégoire Détrez. Controlled Language for Everyday Use: the MOLTO Phrasebook. In *Proceedings of the Second Workshop on Controlled Natural Language (CNL 2010)*, Lecture Notes in Computer Science. Springer, 2012.

[SDS+06] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.