The MOLTO Phrasebook

K. Angelov, O. Caprotti, R. Enache, T. Hallgren, A. Ranta

Chalmers, University of Gothenburg
{krasimir,caprotti,ramona.enache,hallgren,aarne}@chalmers.se

Abstract

This Phrasebook is a program for translating touristic phrases between 14 European languages: Bulgarian, Catalan, Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Polish, Romanian, Spanish, Swedish. The Phrasebook is implemented in the Grammatical Framework programming language as the first demonstration for the MOLTO EU project (molto-project.eu) and will be extended during the project.

1. Introduction

The MOLTO Phrasebook is a multilingual grammar application developed within the EU MOLTO project to showcase the features of the Grammatical Framework, GF, system. It demonstrates how reliable multilingual translations can be derived from an abstract grammar unifying these translations and allowing to translate from any language to the others. The interlingua used by GF, rather than translating words, focuses on meanings or concepts. The GF programming language combines features from grammar languages to functional programming with categorical grammar formalisms and logical frameworks (Ranta, 2004).

From the programmer's perspective, any GF application builds upon a large library of resource grammars and functors: the GF Resource Grammar Library, that currently makes available programmatic primitives to handle syntax, lexicon and inflection for 22 languages with variable coverage. GF deals with the structural differences between languages at compile time, yielding maximal run-time efficiency. Ideally, leaving the linguistic aspects to the GF libraries, the author of an application grammar needs only basic skills in order to add a new language to an application. In the specific case of the Phrasebook application, many of the grammars were created semi-automatically by generalization from examples and grammar induction from statistical models (Google translate). The various configurations of skills tested during the development of the Phrasebook are presented in Section 3.

GF is distributed for all platforms and GF applications can be compiled to JavaScript making them suitable to the web browsers, irrespective of the device. This possibility alone makes GF a convenient tool for fast prototyping of mobile multilingual applications, such as the MOLTO Phrasebook. From the users' perspective, a GF application can be accessed via a web browser on any device, including mobile phones. Off the shelf JavaScript functions are available to construct a friendly user interface in which allowed word choices guide the selection and/or textual input. Not only does the system use incremental parsing to prompt the possibilities, but it also produces quasi-incremental translations of intermediate results from words or complete sentences. The user interface is presented in Section 4.

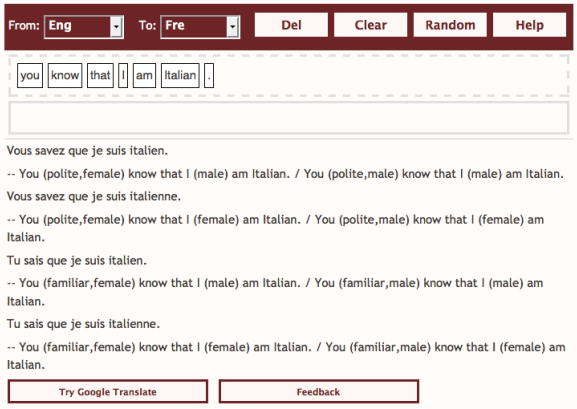

Figure 1: Screen-shot of the online demo

2. Abstract and Concrete Grammars

The GF abstract grammar that captures the object entities and domain of the Phrasebook handles several categories, from units of discourse such as phrases, sentences and questions, to objects like numerable or mass items (*three pizzas* but *some water*), and places, currencies, languages, nationalities, means of transportation, date, and time. It has a collection of constructors that allow to represent for instance a question such as *How far is the zoo?* abstractly as `HowFar(Zoo)` using `HowFar : Place -> Question ; Zoo : PlaceKind`. Each language is produced by linearizing the abstract tree with specific rules that use the GF resource grammar to capture the specific linguistic characteristics. In the example, the French concrete grammar rules are `Zoo = mkPlace (mkN "zoo" masculine) dative` and `HowFar place = mkQS (mkQCl what_distance_IAdv place.name)`. The GF resource grammar for French knows how to build a noun with morphology, `mkN`, a question `mkQS`, and a question clause `mkQCl`. The concrete grammar rule for Swedish is slightly different `HowFar place = mkQS (mkQCl far_IAdv (mkCl (mkVP place.to)))`, yet it is the same as that for Norwegian because of how the resource grammars are designed. Combining it all, the French translation will be *À quelle distance est le zoo?* and the Swedish *Hur långt är det till djurparken?*.

GF application grammars strive for quality. In the

Language	Fluency	GF skills	Informed dev.	Informed testing	Ext. tools	RGL edits	Effort
Bulgarian	***	***	-	-	?	*	**
Catalan	***	***	-	-	?	*	*
Danish	-	***	+	+	**	*	**
Dutch	-	***	+	+	**	*	**
English	**	***	-	+	-	-	*
Finnish	***	***	-	-	?	*	**
French	**	***	-	+	?	*	*
German	*	***	+	+	**	**	***
Italian	***	*	-	-	?	**	**
Norwegian	*	***	+	-	**	*	**
Polish	***	***	+	+	*	*	**
Romanian	***	***	-	-	*	***	***
Spanish	**	*	-	-	?	-	**
Swedish	**	***	-	+	?	-	**

Table 1: Effort estimate

Phrasebook, the kind of quality that can be achieved is exemplified e.g. by sentences that have many translations, each one capturing a flavor of politeness (e.g. "you" in English will have to be disambiguated to polite you, colloquial you and male/female when translating to, say, Italian or French). The abstract grammar makes distinctions between various cases of personal pronouns that identify gender and familiarity, e.g. in greetings or in questions, so that it knows about `IMale` versus `IFemale`, or `YouPolMale` versus `YouFamFemale`. If an ambiguous sentence such as *How old is your daughter?* is entered for translation, it leads to several choices in most languages, for instance in Swedish to *Hur gammal är er dotter?* for the cases of `your(polite,female)` and `your(polite,male)` whereas *Hur gammal är din dotter?* for `your(familiar,female)` and `your(familiar,male)`.

Currently the grammar does not yet cover directions, time and problematic situations, for instance when compared to http://wikitravel.org/en/Phrasebook. With a lexicon of 100 words, the grammar yields 2582 abstract syntax trees of depth 2, which become 656399 of depth 4.

3. The Phrasebook as a Case Study

Developing a multilingual application covering some domain in 14 languages is demanding in terms of language knowledge and quality testing. In Figure 1, we have tracked the type of expertise and effort that was devoted to crafting each single language. Native speakers, fluent in GF and with linguistic background, worked on Bulgarian, Catalan, Polish, and Romanian. However, developers had no knowledge of Danish and Dutch, and had to request the help of native speakers, who were presented with examples generated by a bootstrapped version of the concrete grammars, based on similar languages or on idioms and literal translation taken from the Internet. The full legend for the table is described in (Angelov et al., 2010).

The overall aim is to devise a MOLTO methodology that lowers the cost of adding a new language to a GF application by using automated example-driven grammar generation. The correct design of the batch of examples is language dependent and assumes analysis of the resource grammar, for instance to be able to build inflected words. More precisely, for some languages it is enough to generate examples that show one form of a noun in order to obtain its GF representation (the full inflection table), whereas for other languages, such as German, one has to know up to 6 forms.

4. The Phrasebook at Your Hands

The Phrasebook is distributed as open-source software, licensed under GNU LGPL, from http://code.haskell.org/gf/examples/phrasebook/. It is also available online from the MOLTO project web pages, as a demo and as a mobile application for the Android platform. Users are welcome to send comments, bug reports, and better translation suggestions using the feedback button, as shown in Figure 1. Fall-back to statistical translation is currently implemented just as a link to Google translate, however in future versions, GF will be integrated with tailor-made statistical models.

5. Acknowledgments

The Phrasebook has been built in the MOLTO project funded by the European Commission (FP7/2007-2013) under grant agreement FP7-ICT-247914. The authors are grateful to Inari Listenmaa, Jordi Saludes, and Adam Slaski and to the native speaker informants helping to bootstrap and evaluate the grammars: Richard Bubel, Grégoire Détrez, Rise Eilert, Karin Keijzer, Michał Pałka, Willard Rafnsson, Nick Smallbone.

6. References

K. Angelov, O. Caprotti, R. Enache, T. Hallgren, I. Listenmaa, A. Ranta, J. Saludes, and A. Slaski. 2010. D10.2 molto web service, first version. Project Deliverable D10.2, Gothenburg, Sweden, 06/2010.

A. Ranta. 2004. Grammatical Framework: A Type-Theoretical Grammar Formalism. *The Journal of Functional Programming*, 14(2):145–189.