REGULUS: A Generic Multilingual Open Source Platform for Grammar-Based Speech Applications

Manny Rayner[*†], **Pierrette Bouillon**[*], **Beth Ann Hockey**[†], **Nikos Chatzichrisafis**[*]

[*]University of Geneva, TIM/ISSCO
40, bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
Pierrette.Bouillon@issco.unige.ch, Nikos.Chatzichrisafis@vozZup.com

[†]ICSI/UCSC/NASA Ames Research Center
Moffett Field, CA 94035
mrayner@riacs.edu, bahockey@email.arc.nasa.gov

Abstract

We present an overview of Regulus, an Open Source platform that supports corpus-based derivation of efficient domain-specific speech recognisers from general linguistically motivated unification grammars. We list available Open Source resources, which include compilers, resource grammars for various languages, documentation and a development environment. The greater part of the paper presents a series of experiments carried out using a medium-vocabulary medical speech translation application and a corpus of 801 recorded domain utterances, designed to investigate the impact on speech understanding performance of vocabulary size, grammatical coverage, presence or absence of various linguistic features, degree of generality of the grammar and use or otherwise of probabilistic weighting in the CFG language model. In terms of task accuracy, the most significant factors were the use of probabilistic weighting, the degree of generality of the grammar and the inclusion of features which model sortal restrictions.

1. Introduction

The most common architecture for speech understanding systems is a combination of speech recognition based on n-gram language models, together with robust parsing. For many applications, however, grammar-based language models offer concrete advantages. Training a good n-gram model requires corpus data that are usually not available at the beginning of the project. Comparisons also show that grammar-based recognition gives better results for expert users who have time to learn the coverage of the system (Knight et al., 2001; Rayner et al., 2005a). Another advantage of the grammar-based approach is that a grammar developed for recognition can also be used for syntactic analysis, making it unnecessary to write a robust parser.

Recogniser platforms which support grammar-based language models, like the Nuance Toolkit, generally require the grammars to be specified in a CFG-based framework. In this type of low-level formalism, large grammars are difficult to develop. Even small ad-hoc grammars quickly acquire multiple redundant rules and become difficult to maintain, especially if they are to cover multiple related sub-domains. For all these reasons, several attempts have been made to develop systems that permit language models to be specified in higher-level formalisms, normally some type of unification grammar (UG), and then compile these grammars down into the low-level CFG formalism required by the recogniser (Moore, 1998; Dowding et al., 2001; Bos, 2002). Regulus (Regulus, 2006; Rayner et al., 2003; Rayner et al., 2006) is an Open Source system of this general kind, which has been under development since 2001.

In comparison to earlier compilers, Regulus aims to take the level of abstraction higher. Instead of building separate domain-specific UGs for each new application, Regulus provides one general UG per language, which can be reused for different domains and tasks. In Section 2., we briefly describe how Regulus allows a domain-specific recogniser to be derived from a general UG, and list Open Source resources available under Regulus and related projects which can be used to support this process. The rest of the paper focusses on evaluation, and presents experiments which investigate the impact of various factors on the performance of Regulus-derived recognisers.

2. Regulus resources and processing

The Regulus website (Regulus, 2006) makes available a number of resources, including compilers, an integrated development environment, a Regulus resource grammar for English, online documentation and a set of examples. This material is all described in detail in (Rayner et al., 2006). Open Source Regulus resource grammars for several other languages have been developed under MedSLT (Bouillon et al., 2005), a medical speech translation project which uses the Regulus platform, and are available from the project website (MedSLT, 2005). Descriptions of the French/Catalan, Finnish and Japanese grammars appear in (Bouillon et al., 2006), (Santaholma, 2005) and (Rayner et al., 2005b) respectively; French and Catalan are handled by a single parameterised grammar which covers both languages. A grammar for Spanish is under development.

The process of creating an application-specific Regulus recogniser starts with a general UG, together with a supplementary lexicon containing extra domain-specific vocabulary. An application-specific UG is then automatically derived using Explanation Based Learning (EBL) specialisation techniques (van Harmelen and Bundy, 1988). This corpus-based EBL method is parameterised by 1) a small domain-specific training corpus, from which the system learns the vocabulary and types of phrases that should be kept in the specialised grammar, and 2) a set of "operationality criteria", which control the specialised grammar's

generality (as we will see later, this roughly corresponds to the maximum permitted depth of a derivation). The application-specific UG is then compiled into a Nuance-compatible CFG. Processing up to this point is all carried out using Open Source Regulus tools. Two Nuance utilities then transform the output CFG into a recogniser. One of these uses the training corpus a second time to convert the CFG into a PCFG; the second performs the PCFG-to-recogniser compilation step.

The top-level goal, to be able to compile a specialised form of the general grammar into a CFG-based language model, informs most of the non-standard design decisions in the Regulus resource grammars. In particular, it makes it difficult to use to use modern, heavily lexicalised, formalisms like HPSG (Pollard and Sag, 1994) or LFG (Bresnan and Kaplan, 1985). The key problem is that the Regulus UG-to-CFG compilation algorithm requires all features to be finite-valued, so structure-valued features are not allowed unless they reduce to finite-valued features. Regulus grammars consequently have a somewhat less elegant and general structure than LFG or HPSG grammars; in particular, there is one VP rule for each subcategorisation pattern, instead of a single lexicalised rule schema. The grammars for European languages also use a slightly non-standard treatement of subject-verb inversion, whose motivation is to reduce the total number of features, and thus the size of the derived CFG language model. Despite these restrictions, the more mature grammars (in particular, the English one), offer good coverage of a large range of constructions.

Since sortal constraints are extremely important in grammar-based language models (cf. Section 3.3.), sortal features are included in most rules; the set of permitted sortal values is domain-dependent, and is specified in the domain lexicon. For example, a noun has a feature which encodes its sortal type, and a transitive verb has features which specify the sortal types of its subject and object.

Although grammar based recognition is common in commercial applications, and has been used in a number of research projects, there has to date been little systematic investigation of the performance characteristics of grammar-based recognisers. Since Regulus derives each recognition grammar by specialising it out of the same general base grammar, it is both possible and meaningful to compare the different recognisers produced by varying the parameters of the specialisation process. In the study described in Section 3., we varied both quantitative aspects (the vocabulary size and linguistic coverage), and also qualitative aspects (the global feature-set and the generality of the derived grammar), in order to investigate how these features impact recognition performance.

3. Experiments

The experiments were carried out using the MedSLT system mentioned above. This system was also used for the experiments described in (Rayner et al., 2005a), which demonstrated that the Regulus-derived PCFG language model performed very much better than a conventional n-gram language model on sentences within the coverage of the grammar, and about equally well on out-of-coverage sentences. The coverage of the baseline MedSLT grammar used in both studies is centered on yes/no questions, but also includes WH-questions and phrases. It was derived from four resources: 1) the general English grammar, containing 184 rules written in the Regulus feature-grammar formalism; 2) the core English lexicon, supplied with the general grammar and containing about 450 lemmas, mostly for function words; 3) a MedSLT-specific lexicon, containing 525 lemmas and 4) a training set of 650 MedSLT domain utterances. The vocabulary of the derived MedSLT-specific grammar is 429 surface words. The test set consisted of 801 American-dialect MedSLT utterances recorded during the data collection described in (Rayner et al., 2005a) and a similar earlier data collection.

Each of the variant recognisers was trained and tested in a similar manner. Since performance is very different on in-coverage and out-of-coverage utterances, we present separate figures for each subset. In each case, we measure the coverage of the different derived grammars on the test set, and the performance of the associated recognisers. We measure performance using three parameters: Word Error Rate (WER), Sentence Error Rate (SER) and Task Error Rate (TER). To make TER relevant to the MedSLT speech translation task, we proceed as in (Rayner et al., 2005a) and define a recognition result R to be correct at the task level if and only if the result of translating R into an interlingual representation and then back into English results in a paraphrase of R acceptable in the context of the task. This implies that several types of recognition errors will usually be considered acceptable at the task level. For example, "does **the headache** usually last for more than an hour" would be regarded as an acceptable recognition of "do **the headaches** usually last for more than an hour", since the interlingua does not represent the difference between singular and plural; both versions thus produce the same interlingual representation and are translated uniformly. TER is in general considerably lower than SER, since correctly recognized utterances almost always produce good translations (Bouillon et al., 2005). We evaluated the significance of differences between two different versions of the recogniser for both the SER and TER metrics, using the McNemar test. In tables reporting performance, we mark figures in **bold** if they constitute differences against the baseline significant at $P < 0.05$.

3.1. Varying vocabulary size

In the first set of experiments, we created the variant recognisers by increasing the size of the lexicon, keeping the grammar rules, features and operationality criteria constant. We derived the new lexicon entries from the Medical Subject Headings thesaurus,[1] a controlled vocabulary produced by the National Library of Medicine which is used for indexing, cataloging, and searching for biomedical and health-related information and documents. We used 3868 thesaurus entries, belonging to four different semantic classes potentially relevant to the MedSLT task: these broke down as 829 body part nouns, 1899 symptom nouns, 817 therapy nouns and 323 names of drugs. We built 10 different versions of the lexicon by successively adding larger

[1] `http://www.nlm.nih.gov/mesh/-introduction2005.html`

subsets of the lexical entries from the thesaurus, in randomly chosen increments of about 10% at a time. The results for some of the versions are shown in Table 1.

None of the extra vocabulary added to the variant versions appeared in the test material, so our expectation was that performance would gradually degrade as the vocabulary size increased; the search space becomes larger, but the only possibilities added are irrelevant ones. This is indeed what happened, though the rate of decline was fairly slow. For Version 1, vocabulary has increased from 429 to 946, a factor of 2.2; TER increases from 7.6% to 8.3% (9% relative to the baseline). At Version 4, vocabulary has increased to 2096, a factor of 2.2 relative to Version 1; this time, TER increases to 9.3% (12% relative to Version 1). By the time we reached Version 10, vocabulary is 3788, a factor of 1.8 compared to Version 4, while TER increases to 11.0% (18% relative).

In general, we can say that a doubling of vocabulary results in a degradation in task performance by about 1–2% absolute, or 10–20% relative. All versions were significantly worse than the baseline on the TER metric. The smaller number of significant differences on the strict SER metric is due to recognition errors involving articles; these are essentially a source of noise, which is filtered out by the TER metric.

	WER	SER	TER
In coverage			
(Baseline)	4.9%	15.9%	7.6%
Version 1	5.3%	16.2%	**8.3%**
Version 2	5.2%	16.2%	**8.5%**
Version 4	5.8%	17.1%	**9.3%**
Version 5	5.9%	**17.2%**	**9.5%**
Version 9	6.6%	**18.5%**	**10.9%**
Version 10	6.5%	**18.3%**	**11.0%**
Out of coverage			
(Baseline)	50.4%	96.6%	77.8%
Version 1	53.1%	95.5%	80.1%
Version 2	52.6%	95.5%	80.5%
Version 4	53.7%	95.5%	80.5%
Version 5	54.2%	95.9%	80.5%
Version 9	54.6%	95.5%	81.9%
Version 10	54.2%	95.5%	81.9%

Table 1: Word, Sentence and Task error rates for versions of the MedSLT recogniser built adding different numbers of extra lexical entries, on in-coverage and out-of-coverage data. Results significantly different from the baseline according to the McNemar test are marked in **bold**.

3.2. Varying linguistic coverage

The second set of experiments investigated the effect of changing the coverage; starting with the baseline grammar, we created different versions which successively removed coverage of various constructions. As we made the grammar smaller, we measured the change in recognition performance on in-coverage material. We expected to find that as the grammar's coverage shrank, performance on the material still in coverage improved. We used six versions of the grammar, as follows:

Version 0 Baseline grammar.

Version 1 Remove the rules for subordinate clauses, ("is the pain better **when you lie down**?") and gerunds ("does **lying down** make the pain better?").

Version 2 As (1), but also remove rules for phrasal utterances, in particular lone NPs ("chocolate?") and lone PPs ("in the morning?").

Version 3 As (2), but also remove rules for passives ("is the pain **accompanied by nausea**?").

Version 4 As (3), but also remove rules for WH-questions, ("**where** is the pain?", "**how long** do the headaches last?")

Version 5 As (4), but also remove rules for adverbs, ("does bright light **usually** give you headaches?")

We present the results in two tables; Table 2 tracks the coverage of the different variant grammars, and Table 3 recognition performance of the associated recognisers. In order to be able to make clear comparisons, all the tests in Table 3 were carried out on material within the coverage of Version 5, the most restricted grammar, and hence within the coverage of all the other grammars too.

Table 2 shows that all the groups of deleted rules are relevant to the domain; as each group is removed, coverage drops substantially. The largest drop occurs between Versions 2 and 3: removing passive constructions reduces relative coverage by 26.5%, reflecting the importance in the medical query domain of words like "caused", "relieved", "aggravated" and "preceded". The smallest drop (1.9%) is between Versions 3 and 4, where WH-questions are removed.

Table 3 is most naturally read from bottom to top. In this direction, we conceptualise it as starting with the most restricted grammar, and then successively adding coverage; the table measures how recognition performance degrades on the original coverage as the new rules are added. The largest drop occurs between Versions 2 and 1 (addition of rules for phrasal utterances), where WER increases from 5.4% to 6.1% (13% relative), and TER from 9.2% to 10.5% (14% relative). Over the whole set, WER increases from 5.0% to 6.0% (16% relative), and TER from 8.2% to 10.5% (28% relative). Scalability with respect to coverage extensions was quite good; on the TER metric, the McNemar test showed that the addition of no individual construction resulted in a significant degradation in recognition performance, though the composition of several additions was significant. None of the versions displayed significant differences against the baseline on the SER metric.

3.3. Varying the feature set

The next set of experiments investigated the effect on recognition performance of global constraints encoded in

#	Removed from previous	In-C	Loss
(Baseline)		(580)	(0)
1	Subordinates + gerunds	537	7.4%
2	Phrasal utterances	510	12.1%
3	Passives	356	38.6%
4	WH-questions	345	40.5%
5	Adverbs	306	47.2%

Table 2: Coverage of versions of the grammar, reduced by removing sets of rules. "In-C" measures the number of in-coverage sentence out of a total of 801; "Loss" measures the proportion of utterances for each grammar that are out of coverage for that grammar, but in coverage for the original baseline grammar.

	WER	SER	TER
Baseline	6.0%	18.3%	10.5%
Version 1	6.1%	18.6%	10.5%
Version 2	5.4%	17.7%	9.2%
Version 3	5.3%	17.7%	8.5%
Version 4	5.0%	17.3%	**8.2%**
Version 5	5.0%	17.3%	**8.2%**

Table 3: Word, Sentence and TER error rates for versions of the MedSLT recogniser, derived from the different grammars in Table 2, on the set of 306 utterances that are in-coverage for Version 5. Significant differences against the baseline are in **bold**.

the general grammar's feature set, and inherited by the specialised grammars. This time, we created the variant grammars by suppressing groups of related features; as we remove features and their associated constraints, the language models become looser, and we expect recognition to become less accurate. We created four variant grammars, as follows:

No sortal features Suppress the features that encode sortal restrictions on nouns, verbs and adjective. With these features removed, the grammar would for example permit "does the pain spread to the **coffee**" or "do you get headaches when you drink **neck**".

No agreement features Suppress the features enforcing agreement constraints between nouns, verbs and DETs. Without these features, the grammar permits examples like "**are** the pain frontal" or "**does** you get headaches in the morning".

No PP features Suppress the features that restrict modification of nouns and verbs by PPs. Without them, we get examples like "does the pain last **in the head**" or "does the pain occur **for more than ten minutes**".

No DET features Suppress the features that encode domain-specific restrictions on cooccurrence of nouns and DETs. Without these features, we can for example get "**a head**" or "**the minutes**".

Table 4 shows performance results. It is interesting to see that the WER, SER and TER metrics once again paint very different pictures. In terms of WER on the in-coverage data, removing the sortal features produces the largest degradation in performance (4.9% to 6.7%; 37% relative), fairly closely followed by removing the agreement features (4.9% to 6.2%; 27% relative). In terms of SER, the largest difference in performance results from removing the agreement features (15.9% to 23.3%; 47% relative), ahead of the sortal features (15.9% to 19.8%; 25% relative). With the TER metric, removing the sortal features results in a huge difference (7.6% versus 14.2%; 87% relative), but removing the agreement features makes no difference at all; the second largest difference arises from removing the PP features (7.6% versus 8.8%; 16% relative).

	WER	SER	TER
In coverage			
No sortal	6.7%	**19.8%**	**14.2%**
No agreement	6.2%	**23.3%**	7.6%
No DET	5.7%	**20.8%**	8.1%
No PP	5.3%	17.0%	8.8%
(Baseline)	4.9%	15.9%	7.6%
Out of coverage			
No sortal	48.6%	96.3%	84.7%
No agreement	50.9%	96.8%	77.8%
No DET	50.4%	96.4%	78.1%
No PP	50.4%	96.6%	83.6%
(Baseline)	49.7%	96.4%	77.8%

Table 4: Word, Sentence and Task error rates for versions of the MedSLT recogniser formed by removing features from the grammar, on in-coverage and out-of-coverage data. Significant differences are in **bold**.

In terms of the strict SER metric, the differences for "no agreement", "no DET" and "no sortal" are all significant at $P < 0.001$. This is consistent with the findings of (Rayner et al., 2001), which reported small but statistically significant differences in recognition performance when agreement constraints were removed from three different grammar-based recognisers. In contrast, only the "no sortal" version displays a significant difference in performance when we evaluate using the task-based TER metric.

The fact that we get widely differing results from the various metrics should not be surprising; WER often correlates badly with task error rate (Wang et al., 2003). In this case, the nature of the task means that singular/plural distinctions are usually irrelevant to the semantic representation. Although including agreement constraints makes a substantial difference to the surface measures, this does not translate into corresponding semantic differences.

3.4. Varying generality

In the fourth set of experiments, we investigated the effect of varying the generality of the grammar. We can do this by changing the operationality criteria to create specialised grammars which differ with respect to "flatness"; we start with a completely flat grammar, and then introduce successive levels of intermediate structure. As the grammars become more complex, they also become looser. As in Section 3.2., we expected that this would result in the derived

recognisers offering less accurate performance on the material that they covered. The compensation is that coverage increases, since the more complex grammars are also more flexible. Specifically, we used the following operationality criteria:

Flat The specialised grammar is completely flat, with the root node directly dominating all non-pre-terminals in each derivation.

Two-level The specialised grammar contains two levels, for `utterance` and `np` constituents.

Recursive 1 The grammar contains the three possible non-pre-terminals `utterance`, `np` and `post_mods`[2]. The grammar is potentially recursive, since `post_mods` can be a constituent under `np`, and `np` can be a constituent under `post_mods`.

Baseline The set of operationality criteria supplied with the MedSLT release, which produces a complex recursive grammar. This set of operationality criteria was used for the baseline recogniser in all the experiments in this chapter.

Table 5 shows the coverage of the four different specialised grammars on the 801 utterance test set, and Table 6 shows recognition performance.

The general grammar covers 604 utterances. As we would expect, the Flat grammar loses a great deal of coverage (31.7%) since it has no ability to generalise; it has however the best recognition performance on the material it does cover, with WER at 4.0% and TER at 5.8%.

Moving to the Two-level grammar reduces the relative coverage loss from 31.7% to only 8.4%, at the cost of an insignificant degradation in WER (4.0% to 4.1%) and TER (5.8% to 6.0%). This is clearly a large win.

As we move to the more complex sets of operationality criteria, the two effects come more closely into balance. Moving from Two-level to Recursive 1 reduces the relative coverage loss from 8.4% to 4.5%, while WER increases from 4.1% to 4.4%, and TER from 6.0% to 6.6%. This is still a significant improvement on both evaluation criteria, but the gain is much smaller than the one we saw when moving from Flat to Two-level.

The final transition, from Recursive 1 to Baseline, roughly marks the point where the process of enriching the specialised grammar tops out. The relative coverage loss still decreases, but only from 4.5% to 3.9%. This is counterbalanced by an increase in WER from 4.4% to 4.9%, and in TER from 6.6% to 7.6%. All the test data was produced by naive users, and for these subjects moving from Recursive 1 to Baseline is in fact a backwards step; the loss in recognition performance is slightly worse than the gain in coverage, though the McNemar test shows no significant difference on either the SER or the TER metric. For expert users, anecdotal evidence suggests that the Baseline version is somewhat better, with the increased flexibility of the recogniser appearing more important than the slight degradation in performance.

[2]In this domain, `post_mods` is essentially equivalent to `pp`.

Version	In-C	Loss
Flat	412	31.7%
Two-level	553	8.4%
Recursive 1	577	4.5%
Baseline	580	3.9%
(General)	604	(0)

Table 5: Coverage of grammars built using different operationality criteria. "In-C" measures the number of in-coverage sentence out of a total of 801; "Loss" measures the proportion of utterances for each grammar that are out of coverage for that grammar, but in coverage for the general grammar.

	WER	SER	TER
In coverage			
Flat	4.0%	13.8%	5.8%
Two-level	4.1%	15.2%	6.0%
Recursive 1	4.4%	15.3%	6.6%
Baseline	4.9%	15.9%	7.6%
Out of coverage			
Flat	49.9%	98.5%	72.2%
Two-level	52.4%	97.2%	78.6%
Recursive 1	50.6%	96.4%	77.7%
Baseline	50.4%	96.6%	77.8%

Table 6: Word, Sentence and Task error rates for versions of the MedSLT recogniser built using different operationality criteria, on in-coverage and out-of-coverage data. The proportion of the data that is in coverage depends on the version, as shown in Figure 5. There are no significant differences against the baseline.

3.5. Comparing CFG and PCFG language models

All the recognisers that we have discussed so far use PCFG language models. The training corpus is used twice; first for grammar specialisation, to produce a CFG language model, and then for PCFG training, using the Nuance `compute_grammar_probs` utility. Our final set of experiments evaluates the contribution made by PCFG training. We used the 11 grammars from Section 3.1., and built versions of the recognisers which omitted the PCFG training step, and instead compiled the recogniser directly from the CFG grammar produced by Regulus. We then evaluated the resulting recognisers in the same way as we did for the original PCFG versions in Table 1.

Although we have only a few hundred sentences of training data, it turned out that probabilistic training of the CFG language model made a huge difference to recognition quality. Comparing the results for versions with and without PCFG training, we found that PCFG training on the baseline version reduced in-coverage WER from 8.9% to 4.9% (45% relative), and in-coverage TER from 12.1% to 7.6% (37% relative). As we add more vocabulary, the difference becomes even greater. By the time we reach Version 10 (vocabulary 3788 words), PCFG training reduces in-coverage WER from 14.5% to 6.5% (55% relative), and in-coverage

TER from 22.8% to 11.0% (52% relative).

4. Summary and conclusions

We have presented an overview of the Regulus platform, and described a case study where we compared multiple versions of domain-specific recognisers derived from a single general grammar and domain lexicon. As expected, all the versions performed enormously better on in-coverage than on out-of-coverage data. Somewhat more surprisingly, given the small amount of training data, probabilistic training of the CFG language models also improved performance very substantially on all metrics.

The structural factor which most affected performance was generality. Flat grammars incurred a large coverage loss due to their lack of ability to generalise, but did not offer significantly better recognition performance on in-coverage material compared to the more complex versions. Sortal features, most of which distinguish semantically distinct types of nouns, also had a large impact on performance. Degradation in performance resulting from the addition of new grammatical constructions varied greatly depending on the nature of the construction, with elliptical phrases making the largest difference. Although the addition of no single new construction was significant on its own, the combined effect of adding several new constructions was significant.

Other factors had a smaller effect. The grammars were fairly robust to addition of new vocabulary; the vocabulary size needed to be approximately doubled to produce a significant drop in task performance. On the task criterion, all groups of features except the sortal ones could be omitted without a significant effect, though the difference was significant in terms of strict sentence error.

Up to now, it has been difficult to find methodologically sound ways to evaluate grammar-based language models, and it has been unclear what factors affect their performance; grammar-based language model design has been an art rather than a science, and the academic community has been justifiably somewhat suspicious of it. The experiments we describe here are in contrast clearly defined, and could equally well be carried out on other grammars that had been derived in the same way. By performing similar studies in other domains, it seems reasonable to hope that it will be possible to arrive at general conclusions about the performance characteristics of this type of language model.

5. References

J. Bos. 2002. Compilation of unification grammars with compositional semantics to speech recognition packages. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

P. Bouillon, M. Rayner, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, Y. Nakao, K. Kanzaki, and H. Isahara. 2005. A generic multi-lingual open source platform for limited-domain medical speech translation. In *In Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, Budapest, Hungary.

P. Bouillon, M. Rayner, B. Novellas, Y. Nakao, M. Santaholma, M. Starlander, and N. Chatzichrisafis. 2006. Une grammaire multilingue partagée pour la reconnaissance et la génération. In *Proceedings of TALN 2006*, Leuwen, Belgium.

J. Bresnan and R. Kaplan. 1985. *The mental representation of grammatical relations*. MIT Press, Cambridge, MA.

J. Dowding, B.A. Hockey, J.M. Gawron, and C. Culy. 2001. Practical issues in compiling typed unification grammars for speech recognition. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling, and I. Lewin. 2001. Comparing grammar-based and robust approaches to speech understanding: a case study. In *Proceedings of Eurospeech 2001*, pages 1779–1782, Aalborg, Denmark.

MedSLT, 2005. http://sourceforge.net/projects/medslt/. As of 30 Oct 2005.

R. Moore. 1998. Using natural language knowledge sources in speech recognition. In *Proceedings of the NATO Advanced Studies Institute*.

C. Pollard and I. Sag. 1994. *Head Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

M. Rayner, G. Gorrell, B.A. Hockey, J. Dowding, and J. Boye. 2001. Do CFG based language models need agreement constraints? In *Proceedings of the 2nd NAACL*, Pittsburgh, PA.

M. Rayner, B.A. Hockey, and J. Dowding. 2003. An open source environment for compiling typed unification grammars into speech recognisers. In *Proceedings of the 10th EACL (demo track)*, Budapest, Hungary.

M. Rayner, P. Bouillon, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, H. Isahara, K. Kankazi, and Y. Nakao. 2005a. A methodology for comparing grammar-based and robust approaches to speech understanding. In *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*, Lisboa, Portugal.

M. Rayner, N. Chatzichrisafis, P. Bouillon, Y. Nakao, H. Isahara, K. Kanzaki, and B.A. Hockey. 2005b. Japanese speech understanding using grammar specialization. In *HLT-NAACL 2005: Demo Session*, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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

Regulus, 2006. http://sourceforge.net/projects/regulus/. As of 15 March 2006.

M. Santaholma. 2005. *Linguistic representation of Finnish language in speech-to-speech translation system*. Mémoire de DEA en traitement informatique multilingue, ETI, Geneva.

T. van Harmelen and A. Bundy. 1988. Explanation-based generalization = partial evaluation (research note). *Artificial Intelligence*, 36:401–412.

Y.-Y. Wang, A. Acero, and C. Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *Proceedings of Eurospeech 2003*, pages 609–612, Geneva, Switzerland.