AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models

Eric Atwell, John Hughes, and Clive Souter
Centre for Computer Analysis of Language And Speech
School of Computer Studies, Leeds University, Leeds LS2 9JT, UK
eric@scs.leeds.ac.uk john@scs.leeds.ac.uk cs@scs.leeds.ac.uk

Abstract

Several Corpus Linguistics research groups have gone beyond collation of 'raw' text, to syntactic annotation of the text. However, linguists developing these linguistic resources have used quite different wordtagging and parse-tree labelling schemes in each of these annotated corpora. This restricts the accessibility of each corpus, making it impossible for speech and handwriting researchers to collate them into a single very large training set. This is particularly problematic as there is evidence that one of these parsed corpora on its own is too small for a general statistical model of grammatical structure, but the combined size of all the above annotated corpora should deliver a much more reliable model.

We are developing a set of mapping algorithms to map between the main tagsets and phrase structure grammar schemes used in the above corpora. We plan to develop a Multi-tagged Corpus and a MultiTreebank, a single text-set annotated with all the above tagging and parsing schemes. The text-set is the Spoken English Corpus: this is a half-way house between formal written text and colloquial conversational speech. However, the main deliverable to the computational linguistics research community is not the SEC-based MultiTreebank, but the mapping suite used to produce it - this can be used to combine currently-incompatible syntactic training sets into a large unified multicorpus. Our architecture combines standard statistical language modelling and a rule-base derived from linguists' analyses of tagset-mappings, in a novel yet intuitive way. Our development of the mapping algorithms aims to distinguish notational from substantive differences in the annotation schemes, and we will be able to evaluate tagging schemes in terms of how well they fit standard statistical language models such as n-pos (Markov) models. [1]

Introduction

Several research projects around the world are building grammatically analysed corpora; that is, collections of text annotated with part-of-speech wordtags and syntax trees. Tagged and parsed English corpora (Bank of English [54]; BNC [44], [22]; Brown [24]; ICE [12], [28], [64]; Lancaster-IBM [26], [22]; LOB [1], [3], [38], [42]; London-Lund [62]; Nijmegen [12]; PoW [23], [55], [57]; SEC [63]; TOSCA [46], [31], [12]; UPenn [53], [45]; etc) are used, among other things, as authoritative examples by researchers in English Language Teaching and Lexicography (e.g. [44]), and as training data for statistical syntactic constraint models to improve recognition accuracy in speech and handwriting recognisers (e.g. [37], [10]).

However, projects have used quite different wordtagging and parsing schemes. In contrast to the Speech research community, which has reached broad agreement on an uncontentious set of labelling conventions for phonetic/phonemic analysis, there is no general consensus in the international Natural Language research community on analogous conventions for grammatical analysis. Developers of corpora adhere to a variety of competing models or theories of grammar and parsing, with the effect of restricting the accessibility of their respective corpora, and the potential for collation into a single fully parsed corpus.

In view of this heterogeneity, we have begun to investigate and develop methods of automatically mapping between the annotation schemes of the most widely known corpora, thus assessing their differences and improving the reusability of the corpora. Annotating a single corpus with the different schemes allows for comparisons, and will provide a rich test-bed for automatic parsers.

The most widely known tagged corpora for English are: the Lancaster-Oslo/Bergen (LOB) Corpus; the Brown Corpus; and the London-Lund Corpus. In addi-

[1] This research began with grants from the UK Science and Engineering Research Council (SERC) and the Universities Funding Council's Knowledge Based Systems Initiative (UFC KBSI), and is now funded by the UK Engineering and Physical Sciences Research Council (EPSRC) and the Higher Education Funding Councils' New Technologies Initiative (HEFCs' NTI); we gratefully acknowledge their financial support. We are also grateful to the Corpus Linguistics research teams who have generously provided background information on their tagging and parsing schemes.

tion, the International Corpus of English (ICE) should be included as its tagset has now been published [28]. Parsed corpora for English include: the Lancaster-IBM Treebank; the Lancaster-IBM Spoken English Corpus (SEC) Treebank; the Lancaster-Leeds Treebank; the Polytechnic of Wales (POW) Corpus; the Nijmegen Corpus; the TOSCA Corpus;; and the University of Pennsylvania (UPenn) Treebank. We plan to include the parsed ICE-GB (Great Britain component of ICE) and the BNC (British National Corpus) in the project when they become available.

As a development and testing resource, we are using the text of the Lancaster-IBM Spoken English Corpus (SEC). The SEC is a collection of recordings of radio broadcasts with accompanying annotated transcriptions, collected by Lancaster University and IBM UK as a general research resource. The SEC is available from the International Computer Archive of Modern English (ICAME) based at the Norwegian Computing Centre for the Humanities (in Bergen, Norway). The corpus exists in several forms and annotations: the digitised acoustic waveform; the graphemic transcription annotated with prosodic markings; and a part-of-speech analysis (using the LOB Corpus tagset). Skeletal parsing has been added to create the SEC Treebank, and this forms a subset of the Lancaster-IBM Treebank. Gerry Knowles (Lancaster) and Peter Roach (Leeds) are collaborating in an ESRC-funded project to set up a time-aligned database of recorded speech, accompanied by phonetic and graphemic transcriptions. Our proposal will produce, as a side-effect, several alternative tagged and parsed versions of the SEC which will be made available to the SEC database project collaborators. It will also be able to act as a test-bed for the comparison and evaluation of parsing schemes.

Objectives of the project

The main objectives are as follows :

To design and implement algorithms for mapping between corpus annotation schemes; for both wordtag sets and phrase structure grammar schemes.

To empirically evaluate the accuracy and shortcomings of the developed mapping algorithms, by applying them to the tagged SEC and the SEC Treebank. The outcome of this evaluation will be to highlight the notational and substantive differences between the alternative tagging and parsing schemes.

To build a Multi-Tagged Corpus, by enhancing the Spoken English Corpus with different wordtagging schemes.

To build a Multi-Treebank, by enhancing the Spoken English Corpus with grammatical analyses according to several alternative grammatical theories.

To investigate the use of the Multi-Treebank as a benchmark for grammars and parsers.

Initially, we considered adopting the 'Interlingua approach' to mapping, as used in Machine Translation projects such as EUROTRA. This would require us to develop tagset mappings between the LOB Corpus (our primary tagset Interlingua) and each of the 'major' tagged corpora: BROWN, ICE, Lancaster-IBM, and UPenn. Next full grammar mappings would be developed between the Lancaster-IBM Treebank (our primary parsing scheme Interlingua) and each of: the UPenn Treebank and the Lancaster-Leeds Treebank. The ICE and BNC tagsets and parsing schemes could be included when they become available. Mapping between tagsets will involve relabelling of words, whereas mapping between grammar schemes also involves structural manipulation. These treebanks have been chosen for their skeletal parsing schemes, which are of relatively similar structure apart from a small number of systematic differences.

We have chosen the SEC as a 'core' text for this project, because

1. the tagged SEC uses the same tagset as the LOB Corpus (widely considered to be the UK standard and our proposed primary tagset);
2. the parsed SEC uses the same grammatical scheme as the Lancaster-IBM Treebank (our proposed primary parsing scheme);
3. these are the annotation schemes which we have most prior experience of;
4. the text material, BBC radio broadcasts, are a neutral compromise between written and conversational spoken English genres.

Our aim is to develop bidirectional mappings for the above tagsets and grammar schemes, although we appreciate that for mapping from simple to delicate schemes this will not be possible, and that mappings will be imperfect. As mapping algorithms are developed and tested, and whilst building the Multi-Tagged Corpus and Multi-Treebank, we will compile "handbooks" of common errors (i.e. mismatches) and their corrections. These will help future users of the developed mapping algorithms to straightforwardly post-edit their mapped corpora and treebanks, thus maximising resource reusability. To map between two tagsets other than LOB, two mappings will be necessary (via the primary tagset, our "interlingua" representation); similarly for non-terminal grammar schemes. We appreciate the danger of propogating incorrect mappings.

If there is sufficient time, we hope to go on to investigate mapping algorithms for other (more detailed) grammar schemes; for example the parsed POW Corpus (Systemic Functional Grammar), and the parsed Nijmegen Corpus (Extended Affix Grammar). The non-corpus-based Generalised Phrase Structure Grammar (GPSG) (as used in the Alvey Natural Language Toolkit ANLT) should also be included. Mapping from these to the Lancaster-IBM Treebank grammar scheme would only be uni-directional i.e. from a detailed to a skeletal analysis.

The Multi-Treebank will be produced by applying the final version of each grammar scheme mapping al-

gorithm to the SEC Treebank. Similarly, for the Multi-Tagged Corpus, the final version of each tagset mapping algorithm will be applied to the tagged SEC. The resulting annotations will then be intensively proofread and post-edited. This will require consultations with authorities in each of the tagsets and grammar schemes involved.

Progress to date

We envisage three main stages to the project: implementation of algorithms for mapping between tagsets; implementation of algorithms for mapping between phrase structure grammatical analysis schemes; and investigating applications of the mapping programs, multi-tagged corpus, and multi-treebank.

We are currently in the first of these. Mapping algorithms are being designed and implemented between the LOB Corpus tagset and each of: the tagged BROWN Corpus, the tagged ICE, the Lancaster-IBM Treebank, the UPenn Tagset (and the BNC tagset will be added when published). Each tagset is being considered in turn:

1. Analysis of the notational and substantive differences between the LOB tagset and the 'current' tagset.
2. Design and implementation of a mapping algorithm (two-way, where possible).
3. Evaluate success of algorithm by applying it to the tagged SEC; incrementally improve in light of common errors and linguistic intuition.

A side-effect of this phase is the production of a Multi-Tagged Corpus: the SEC text annotated with each tagset.

A standard format for tagged and parsed corpora

As well as using different tagsets and parsing schemes, different annotated corpora come in a range of different formats - see [57], [59], [60]. A non-trivial first step in merging tagged and parsed corpora is to decide on a unitary standardised format. Although the Text Encoding Initiaitive (TEI) [61], [18] offers general guidelines for text formatting standards, and some corpora (including BNC, ICE) aim to be "TEI-conformant", in practice it seems almost as hard for Corpus linguists to agree to accept a single annotation format as it is to agree on a single annotation scheme. Our mapping software will use a standardised internal format for taggings and parse-trees, but will have to be able to accept input and produce output in a range of existing formats.

Hand-crafting a detailed mapping

One approach to obtaining a mapping between two tagsets is to use expert linguistic knowledge in identifying the relationship between particular tags, and is exemplified in, for example, [58]. In this work, Souter drew up a mapping between the parts of speech used in the CELEX database [17], (which were themselves derived largely from those in LDOCE [51]), and the systemic functional grammar (SFG) used to hand parse the Polytechnic of Wales corpus [23], [55].

The aim was to provide a large lexicon to support SFG-based parsing programs. The original CELEX lexicon, which contained some 80,000 English wordforms, was transformed into a lexicon with SFG tags, using a semi-automatic mapping program, written in the AWK programming language. The resulting lexicon was then compatible with a large corpus-based systemic grammar consisting of over 4,000 phrase-structure rules [56]. Together they can then support relatively robust probabilistic parsing programs.

The problems encountered in trying to specify such a mapping result from disparity in the level of delicacy in the two tagging schemes. Mapping from a coarse- to fine-grained grammar must be achieved manually, use subcategorisation information contained in the lexicon, contextual information, or exception lists. Souter's program contained simple one-to-one mappings, many-to-one mappings, and one-to-many mappings supported by exception lists and subcategorisation information. In his work, contextual information could not be used to support the mapping because the source material was a lexicon and not a tagged corpus. A small part of the mapping code (used to map between pronoun labels) is shown in Figure 1.

```
else if (category == "PRON") # pronouns
{
    if (\$8 == "Y") printf("HWH)(") # WH-pronouns
    else if ((\$1 == "no-one")||\
             (\$1 == "nobody")||\
             (\$1 == "nothing")||\
             (\$1 == "noone")||\
             (\$1 == "none"))
        printf("HPN)(") # negative pronouns
    else printf("HP)(") # other pronouns
}
```

Figure 1: Fragment of an AWK mapping from CELEX to SFG tagsets

Here, the coarse-grained CELEX tag PRON (pronoun) is mapped to three SFG tags, HWH, HPN and HP. The default mapping is to HP (pronominal head of the nominal group), but if the CELEX lexicon contains subcategorisation information in the form of a Y in column 8, then we can assign the label for wh-pronoun head (HWH). An exception list is used to map to the third SFG pronoun label (HPN), for negative pronoun heads.

Incremental refinement through feedback

Earlier work at Leeds [32] explored ways that a probabilistic grammar may be improved with positive feedback from a human user; this has direct implications

for how to improve the mappings incrementally. As the mapped annotations are to be hand-corrected by experts this provides positive feedback. Rather than tag the core text completely using the best derived mapping a better idea would be to do it in sections and then have the expert correct the errors in each section in turn. After each section is complete the mapping rules will be updated to incorporate the new information. Hopefully this will enable future sections to be mapped more accurately. This method is similar to that used by the Nijmegen corpus parsing group, [47]

Problems with the interlingua-based approach

The interlingua idea seems sound for several reasons. Amongst these is the saving made in required mappings. For instance, five tagging schemes would require twenty mappings if each pair is mapped directly in both directions. However only eight mappings are required if one of the tagging schemes acts as an interlingua. This saving becomes greater as more tagging schemes are considered. The interlingua also helps the mappings attain a level of consistency as the interlingua is the basis of all possible mappings from one tagging scheme to another. However, the interlingua may cause problems in the instances where it is coarse-grained relative to other tagging schemes. For instance, the LOB tagset has no notion of verb transitivity whereas the ICE tagset does. If a mapping is being made between two tagging schemes both of which incorporate the concept of transitivity then a tag may be wrongly allocated as the sense of transitivity is lost via the LOB tagset interlingua.

One problem with the work plan is the strong emphasis on this problem of imperfect mappings due to course-grained parts of the interlingua. It may turn out after experimentation that the contextual information of the surrounding tags and words make up for this. The sentence, S, the interlingua tags, a, and the other tagging scheme's tags, b, can be represented as follows:

Sentence	Tagging 1	Tagging 2
S_1	a_1	b_1
S_2	a_2	b_2
S_3	a_3	b_3
$\vdots$	$\vdots$	$\vdots$
S_{n-2}	a_{n-2}	b_{n-2}
S_{n-1}	a_{n-1}	b_{n-1}
S_n	a_n	b_n

Using a window of the closest four neighbours, say, the tag a_i when presented with a difficult tag, b_i, has the additional context information of S_{i-2}, S_{i-1}, S_{i+1}, S_{i+2}, a_{i-2}, a_{i-1}, a_{i+1} and a_{i+2} to work on. Previous research (including [4], [5], [33], [34], [35], [36]) has indicated that there is a high degree of useful contextual information implied by the surrounding items. This contextual information can be used with clustering techniques to classify words. These techniques could be applied to the tagging scheme with little modification. The classifications could then be examined to see which types of tag are difficult to group. Tags which are difficult to cluster may be useful in identifying problem areas early. As an example, Figure 2 shows a clustering dendogram for the LOB Corpus tagset.

One possible problem might be: given the set of words in $(S_1 \ldots S_n)$ and the interlingua tags $(a_1 \ldots a_n)$ how are the other tagging scheme's tags $(b_1 \ldots b_n)$ derived? Obviously, the tagging scheme itself can be used to map directly the words $(S_1 \ldots S_n)$ onto the tags $(b_1 \ldots b_n)$ without knowledge of the interlingua tags $(a_1 \ldots a_n)$. Also, mappings could be made solely between the interlingua tags $(a_1 \ldots a_n)$ and the other tags $(b_1 \ldots b_n)$. This could be done by having an expert tag a section of the 'core text' with both the interlingua and the other tagging scheme. Probabilisitic rules could be derived indicating how the tags match up. These rules would be strengthened if the context of the surrounding tags was incorporated.

When the two pieces of information are combined it is hoped that a more accurate mapping can be achieved. This can be done by mapping directly from the word plus tag to the new tag. For instance;

$$S_i + a_i \mapsto b_i.$$

However, rules could grow very large even when regularities are used to reduce their number.

Alternatively, a mapping could be made for $S_i \mapsto b_{i1}$ according to the standard annotation rules for the non-interlingua tagging scheme; and a mapping could be done for $a_i \mapsto b_{i2}$ according to the procedure outlined above. These mappings produce two potential tags *independently*. One algorithm might always accept b_{i1} when $b_{i1} = b_{i2}$. When $b_{i1} \neq b_{i2}$ a decision needs to be made as to which, if any, of the tags should be chosen. Brill [14] developed a clever and highly accurate tagging scheme which could have implications for this problem. He tagged every occurrence of a word with its most probable tag if there was more than one choice. A second pass of the corpus would update the tags according to a set of automatically acquired rules. A similar idea could be utilised to choose between the tags. Perhaps the most probable tag would always be selected on the first pass but we would allow that decision to be altered on a second pass according to rules derived from earlier sections of the corpus that had been tagged and checked by the expert linguist. This, then, is another example of incremental learning.

Combining Symbolic and Statistical Approaches to Language

Our research is particularly relevant to this workshop, as it is clear we will have to combine rule-based symbol-mapping knowledge with statistical disambiguation models. We envisage that the bulk of a

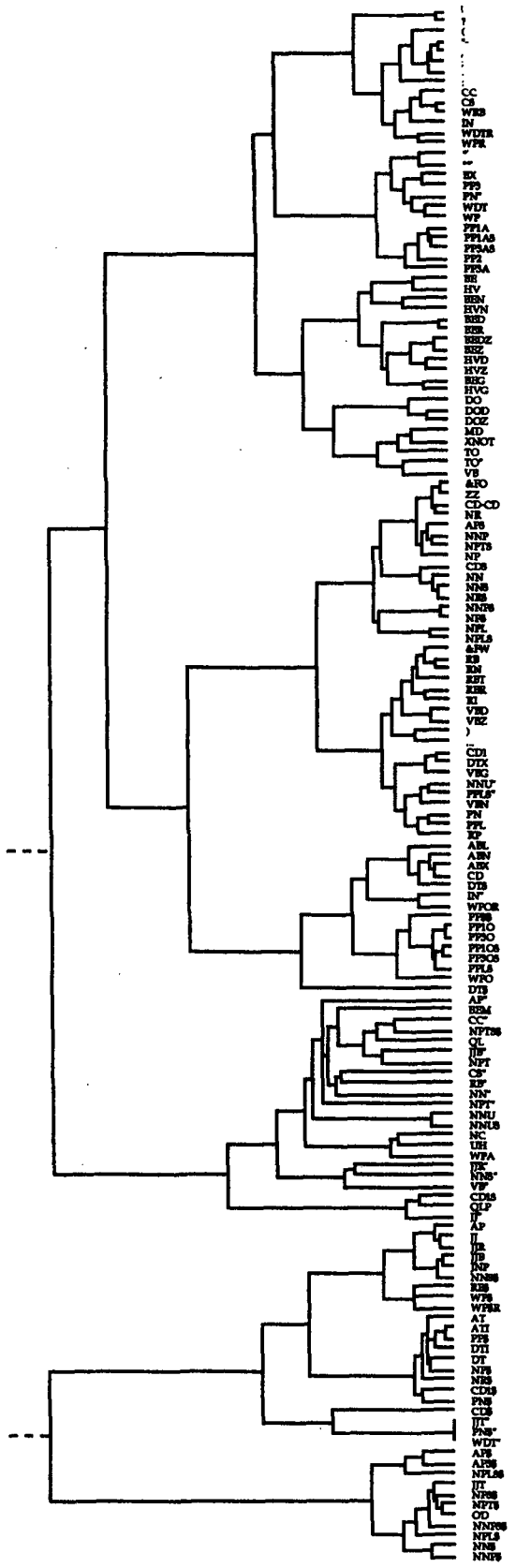

Figure 2: A Clustering of LOB Tags

mapping can be done using simple one-to-one symbol replacement rules; but some rules will map one source symbol onto a set of more than one target symbol. A statistical part-of-speech tagger along the lines of the CLAWS tagger used on LOB [42], [3], [25] can then be used to select between the reduced candidate set using a statistical context model; probably a 1st-order Markov or Bi-Pos model is most appropriate as this is the simplest 'standard' statistical language model and is widely used and understood, see for example [2], [48], [7], [37], [9], [49].

We can arrive at such a model of source-to-target mapping by 'working backwards': first run a CLAWS-style Markovian target-tagset tagger over the text, ignoring the source tags; proofread the output to note where this makes mistakes (assigns incorrect target tags); and then devise source-to-target tag mapping rules only for these cases. We are aware from our own initial attempts at deivising tag-mappings that this requires a high level of specialist linguistic knowledge, of both source and target tagset; this "symbolic patching of the statistical model" approach minimises the 'linguistic expertise' we need to capture (and first develop!) to devise symbolic mapping rules. We have learnt of corpus-tagset mapping work by a number of other researchers (including [13], [30], [20] [39] [41], [52]), but generally such research in the past has been merely a means to an end (to create a re-tagged Corpus), so the full mapping algorithms have not been formalised or published; but if all we need is a limited nuber of mapping-rules to "patch" the Markov model then we may be able to glean sufficient details from informal notes etc. It may appear that we are promoting bad Software Engineering principles in advocating symbolic "patches" to fix the flaws in the statistical model as and when we spot them - patching up a program as the bugs seep out. However, we prefer to view this as a principled, well-founded approach to combining symbolic and statistical models, minimising the 'overlap' by ensuring that each has a separate useful contribution to make to the overall mapping task.

We envisage combining a CLAWS-style tagger module for each of the target tagsets into a single Multi-tagger program. This accepts as input a stream of words annotated with tag(s) from one (or more) of the source tagsets; to output is a stream of words plus tags from ALL target tagsets. This model allows for inclusion of mapping rules both direct from source to target tagset, and via an interlingua (a backup default to try if there are no direct mapping rules).

Future Work

The remainder of the project will be devoted to the two other phases of the plan listed earlier, mapping between phrase-structure parsing schemes, and investigating applications of the multi-tagged corpus and multi-treebank.

Implementation of Algorithms For Mapping Between Grammar Schemes

Initially mapping algorithms will be designed and implemented between the Lancaster-IBM Treebank grammar scheme, and each of the UPenn Treebank and the Lancaster-Leeds Treebank. Each grammar scheme will be considered in turn:

1. Analysis of the notational and substantive differences between the Lancaster-IBM grammar scheme and the 'current' grammar scheme.
2. Manually parse a subset of the SEC according to the 'current' grammar scheme. This subset should be sufficient to allow a prototype mapping algorithm to be induced.
3. Apply mapping algorithm to the parsed SEC; incrementally improve in light of common errors and linguistic intuition.

Depending on how much time is available, mapping algorithms for more detailed grammar schemes will be investigated: parsed POW Corpus, parsed Nijmegen Corpus, GPSG, and the BNC grammar scheme (when published). A side-effect of this phase will be the production of a Multi-Treebank; the SEC automatically annotated with each grammar scheme.

The all-in-one Multi-tagger architecture outlined above can be carried over to a Multi-parser. Instead of a CLAWS-style Markovian tagger, for each target parsing scheme a grammar and parser can be extracted directly from the corresponding training Treebank. A Context-Free Grammar can be elicited directly by extracting each non-terminal and its immediate-daughter-sequence, to become the left-hand-side and right-hand-side respectively of a context-free grammar rule [6]; frequencies of constituents in the training treebank can be used to make this a Probabilistic Context Free Grammar [49], useable in a treebank-trained probabilistic parser such as those in [2], [29], [9], [50]. Rather than producing a single, fully correct parse-tree for each input sentence, these probabilistic Treebank-trained parser generally output an ordered list of possible parse-trees, with a probability or weight attached to each. As with the procedure for developing a partial tag-mapping, we need only devise source-to-target parse-tree-constituent mappings in cases where the target-parser's 'best' parsetree is not fully correct.

Assessment of the Multi-Treebank as a Benchmark for Grammars

This requires analysis of the substantive differences between different parses of the SEC sentences; detailed analysis of how many and which constructs differ in the different language models. It may be possible to divide the sentences in the SEC into two subsets: a common core of "uncontentious" sentences which all or most theories analyse in much the same way; and a "troublesome" subset of sentences which linguists can concentrate their debate on.

One possible criticism of a lot of work in Corpus Linguistics, including the AMALGAM proposed work-plan, is that we restrict ourselves to variants of existing tagging and parsing schemes which are specifically crafted for Corpus annotation, but which are quite different from grammar models being advocated and developed by non-Corpus-based theoretical linguists, such as GPSG or HPSG (see e.g. [27]). Unfortunately, we know of no English corpus parsed according to such a feature-based unification-oriented formalism, so one cannot readily be included in the AMALGAM project; however, we would like to hear from theoretical linguists who we could collaborate with in extending the multi-parser to a unificational grammar formalism. It is not clear that our multi-corpus will be a 'fair' benchmark for testing grammars and parsers from such widely-differing theories; it will be interesting to see whether the partition between "uncontentious" and "troublesome" sentences is also applicable in assessment of unificational grammars.

Another constraint of the AMALGAM project is that we are not considering Corpus-based semantic tagging schemes (e.g. [40], [22], [21]), only syntactic tagging schemes. Again, it will be interesting to see whether the syntactically "troublesome" sentences are also semantically complex or anomalous; but this is a question for another, follow-up, project.

Comparison of the Multi-Treebank with other parsed corpora

We will compare the SEC data with other parsed texts (LOB, UPenn, POW, Nijmegen, etc), to assess differences in the range and frequency distributions of grammatical constructs. The SEC consists of transcripts of scripted (and probably rehearsed) radio broadcasts. Some natural language researchers may feel that the Spoken English dataset is thus inappropriate for their work, since the grammars and parsers they are developing are designed for a different type of language, for example, unrehearsed informal spoken dialogue as found in the London-Lund Corpus and British National Corpus spoken section, or more formal published (written) text as found in the Brown and Lancaster-Oslo/Bergen Corpora. It may be appropriate to augment the SEC dataset with additional material from alternative sources. On the other hand, it may be that the main differences are in vocabulary rather than syntax, and that the coverage of the SEC, though not complete or perfect, is adequate for most applications. We will try to find empirical evidence for or against the acceptability of 'scripted' Spoken English to the NL community.

Assessment of Multi-Treebank as a Benchmark for Parsers

This will involve attempting to parse the SEC text with other parsers, available from a variety of sources. To avoid the need for intensive manual proofreading or checking of results, a (semi-)automatic assessment procedure will be developed.

Anticipated Results

The tangible 'deliverables' of use to the Speech and Language research community include:

- Final implementations of algorithms for mapping between pairs of tagsets
- Final implementations of algorithms for mapping between pairs of Treebanks
- Handbooks of common errors and corrections for post-editing
- The Multi-Tagged Corpus
- The MultiTreebank
- Reports on the above

The mapping software, Multi-tagged Corpus and MultiTreebank (along with postediting handbooks and documentation) will be delivered to ICAME and Oxford Text Archive for public distribution; they will also be available for incorporation into the SEC Speech Database. Reports on the findings of the three stages of investigations will be made widely available to all interested parties through SALT and ELSNET (UK and European Networks of Excellence) and other channels including conference presentations and journal papers.

Applications

The implemented mapping algorithms will be made widely available to the UK and international speech and language research community. They will allow research groups who are using corpus-based training data to make use of other corpora straightforwardly, without substantial modifications. Any current and future users of corpora will have a much expanded resource.

The Multi-Tagged Corpus and the Multi-Treebank will be distributed, along with the main Spoken English Corpus, through ICAME. They will also be available for incorporation into the SEC Speech Database currently being created by Gerry Knowles and Peter Roach, further enhancing the SEC as a general research resource.

Both the Multi-Treebank and the Multi-Tagged corpus will potentially be used by speech and language technology groups for many research and teaching purposes, including: training data for speech-recognisers, optical text recognisers, word processor text-critiquing systems, machine translation systems, natural language interfaces, and NLP applications generally; and for providing examples for English Language Teaching (ELT) grammar textbooks and training material. In addition, the Multi-Treebank may be used as a testbed and benchmark for parsers (explored in the workplan). It would also be a rich resource for grammar-learning experiments - a research topic of growing interest (see e.g. [8], [11], [16], [33]).

We envisage supplying the computational linguistics research community with a valuable research resource, and the ACL Workshop will be an invaluable opportunity for us to survey potential customer requirements and preferences!

References

[1] Eric Steven Atwell. 1982. *LOB Corpus Tagging Project: Manual Post-edit Handbook.* Departments of Computer Studies and Linguistics, Lancaster University.

[2] Eric Atwell. 1983. Constituent Likelihood Grammar. In *Journal of the International Computer Archive of Modern English (ICAME Journal), No. 7*, pages 34–66. Norwegian Computing Centre for the Humanities, Bergen University

[3] Eric Steven Atwell, Geoffrey Leech and Roger Garside 1984. Analysis of the LOB Corpus: progress and prospects in Jan Aarts and Willem Meijs (ed), *Corpus Linguistics: Proceedings of the ICAME 4th International Conference on the Use of Computer Corpora in English Language Research* pp40-52, Amsterdam: Rodopi.

[4] Eric Steven Atwell. 1987. A parsing expert system which learns from corpus analysis. In Willem Meijs, editor, *Corpus Linguistics and Beyond: Proceedings of the ICAME 7th International Conference*, pages 227–235. Amsterdam, Rodopi.

[5] Eric Steven Atwell and Nikos Drakos. 1987. Pattern Recognition Applied to the Acquisition of a Grammatical Classification System from Unrestricted English Text. In Bente Maegaard, editor, *Proceedings of the Third Conference of European Chapter of the Association for Computational Linguistics*, New Jersey, Association for Computational Linguistics.

[6] Eric Steven Atwell. 1988. Transforming a parsed corpus into a corpus parser. In Merja Kyto, Ossi Ihalainen, and Matti Risanen, editors, *Corpus Linguistics, Hard and Soft: Proceedings of the ICAME 8th International Conference*, pages 61–70. Amsterdam, Rodopi.

[7] Eric Steven Atwell. 1988. Grammatical analysis of english by statistical pattern recognition. In Josef Kittler, editor, *Pattern Recognition: Proceedings of the 4th International Conference Cambridge*, pages 626–635. Berlin, Springer-Verlag.

[8] Eric Steven Atwell. 1992. Overview of grammar acquisition research. In Henry Thompson, editor, *Workshop on sublanguage grammar and lexicon acquisition for speech and language: proceedings*, pages 65–70. Human Communication Research Centre, Edinburgh University.

[9] Eric Steven Atwell. 1993. Corpus-based statistical modelling of English grammar. In Clive Souter and Eric Atwell, editors, *Corpus-Based Computational Linguistics*, pages 195–214. Amsterdam, Rodopi.

[10] Eric Steven Atwell. 1993. Linguistic Constraints for Large-Vocabulary Speech Recognition In Eric Steven Atwell (ed), *Knowledge at Work in Universities: Proceedings of the second annual conference of the Higher Education Funding Councils' Knowledge Based Systems Initiative*, pp26-32. Leeds, Leeds University Press.

[11] Eric Steven Atwell, Simon Arnfield, George Demetriou, Stephen Hanlon, John Hughes, Uwe Jost, Rob Pocock, Clive Souter, and Joerg Ueberla. 1993. Multi-level disambiguation grammar inferred from English corpus, treebank and dictionary. In *Proceedings of the IEE Two One-Day Colloquia on Grammatical Inference : Theory, Applications and Alternatives*, (Ref 1993/092). London, Institution of Electrical Engineers (IEE).

[12] Henk Barkema. 1994. *The TOSCA Analysis Environment for ICE.* Technical Report, Department of Language and Speech, Katholieke Universiteit Nijmegen, The Netherlands.

[13] Nancy Belmore. 1991. Tagging Brown with the LOB tagging suite. In *Journal of the International Computer Archive of Modern English (ICAME Journal). No. 15*, pages 63–86. Norwegian Computing Centre for the Humanities, Bergen University.

[14] Eric Brill. 1991. *A Simple Rule-Based Part of Speech Tagger.* Technical Report: Department of Computer Science, University of Pennsylvania.

[15] Eric Brill and Mitchel Marcus. 1992. Tagging an Unfamiliar Text with Minimal Human Supervision. In Robert Goldman, editor, *Working notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, AAAI Press.

[16] Eric Brill, David Magerman, Mitchell Marcus, and Beatrice Santorini. 1992. Deducing Linguistic Structure from the Statistics of Large Corpora. In Carl Weir and Ralph Grishman, editors, *Proceedings of AAAI-92 Workshop Program: Statistically-Based NLP Techniques* San Jose, California.

[17] Gavin Burnage. 1990. CELEX - A Guide for Users. Nijmegen: Centre for Lexical Information (CELEX).

[18] Lou Burnard. 1991. What is the TEI? In D. Greenstein, editor, *Modelling Historical Data.* Goettingen: St. Katharinen.

[19] K. Church. 1992. *Parts of Speech Tagging.* Fifth Annual CUNY Conference on Human Science Processing.

[20] Aviv Cohen. 1994. personal communication.

[21] George C. Demetriou and Eric Steven Atwell. 1994. *Machine-Learnable, Non-Compositional Semantics for Domain Independent Speech or Text Recognition* to appear in Proceedings of 2nd Hellenic-European Conference on Mathematics and Informatics (HERMIS), Athens University of Economics and Business.

[22] Elizabeth Eyes and Geoffrey Leech. 1993. Progress in UCREL research: Improving corpus annotation practices. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: design, analysis and exploitation; Proceedings of the 13th ICAME conference*, pages 123–144. Amsterdam: Rodopi.

[23] Robin Fawcett and Michael Perkins. 1980. *Child Language Transcripts 6-12. (With a preface, in 4 volumes).* Department of Behavioural and Communication Studies, Polytechnic of Wales.

[24] W.N. Francis and H. Kučera. 1979. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Corrected and Revised edition).* Department of Linguistics, Brown University, Providence, Rhode Island.

[25] Roger Garside, Geoffrey Leech, and Geoffrey Sampson (editors). 1987. *The Computational Analysis of English : A Corpus-Based Approach.* Longman, London and New York.

[26] Roger Garside, Geoffrey Leech and Tamás Váradi. 1990. *Manual of Information for the Lancaster Parsed Corpus.* Technical Report, Department of Linguistics and Modern English, University of Lancaster, UK.

[27] Gerald Gazdar and Chris Mellish. 1989. *Natural Language Processing in POP-11 : An Introduction to Computational Linguistics.* Addison Wesley.

[28] Sidney Greenbaum. 1993. The Tagset for the International Corpus of English. In Clive Souter and Eric Atwell (eds.) *Corpus-based Computational Linguistics* Amsterdam: Rodopi.

[29] Robin Haigh, Geoffrey Sampson and Eric Atwell. 1988. Project APRIL – a progress report on the Leeds annealing parser project. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 104–112. New Jersey, Association for Computational Linguistics (ACL).

[30] Robin Haigh. 1993. personal communication.

[31] Hans van Halteren and Nelleke Oostdijk. 1993. Towards a syntactic database: the TOSCA analysis system. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: design, analysis and exploitation; Proceedings of the 13th ICAME conference*, pages 145–162. Amsterdam: Rodopi.

[32] John Hughes. 1989. *A Learning Interface to the Realistic Annealing Parser.* Technical Report: School of Computer Studies, The University of Leeds.

[33] John Hughes and Eric Steven Atwell. 1993. utomatically acquiring and evaluating a classification of words In *Proceedings of the IEE Two One-Day Colloquia on Grammatical Inference : Theory, Applications and Alternatives*, (Ref 1993/092). London, Institution of Electrical Engineers (IEE).

[34] John Hughes. 1994. *Automatically Acquiring a Classification of Words.* PhD Thesis: School of Computer Studies, The University of Leeds.

[35] John Hughes and Eric Steven Atwell. 1994. A Methodical Approach to Word Class Formation Using Automatic Evaluation. In Lindsay Evett and Tony Rose, editors, *Proceedings of AISB workshop on Computational Linguistics for Speech and Handwriting Recognition.* Leeds University.

[36] John Hughes and Eric Steven Atwell. 1994. The Automated Evaluation of Inferred Word Classifications. In Tony Cohn (ed), *Proceedings of the 11th European Conference on Artificial Intelligence*, Amsterdam.

[37] F. Jelinek. 1990. Self-organised language modelling for speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann.

[38] Stig Johansson, Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The Tagged LOB Corpus - Users' Manual.* The Norwegian Centre for the Humanities, Bergen.

[39] Stig Johansson. 1994. personal communication.

[40] Uwe Jost and Eric Steven Atwell. 1993. Deriving a probabilistic grammar of semantic markers from unrestricted English text In *Proceedings of the IEE Two One-Day Colloquia on Grammatical Inference : Theory, Applications and Alternatives*, (Ref 1993/092). London, Institution of Electrical Engineers (IEE).

[41] Judith Klavans. 1994. personal communication.

[42] Geoffrey Leech, Roger Garside and Eric Atwell. 1983. The automatic grammatical tagging of the LOB Corpus. In *Journal of the International Computer Archive of Modern English (ICAME Journal), No. 7*, pages 13–33. Norwegian Computing Centre for the Humanities, Bergen University.

[43] Geoffrey Leech and Roger Garside. 1991. Running a grammar factory: The production of syntactically analysed corpora or "treebanks". In Stig Johansson and Anna-Brita Stenström, editors, *English Computer Corpora: Selected Papers and Research Guide.* Berlin: Mouten de Gruyter.

[44] Geoffrey Leech. 1993. *100 Million Words of English: The British National Corpus (BNC) Project.* English Today.

[45] Mitch P. Marcus and Beatrice Santorini. 1992. Building Very Large Natural Language Corpora: The Penn Treebank. In N. Ostler, editor, *Proceedings of the 1992 Pisa Symposium on European Textual Corpora.*

[46] Nelleke Oostdijk. 1989. *TOSCA Corpus Manual.* University of Nijmegen.

[47] Nelleke Oostdijk. 1991. Corpus linguistics and the automatic analysis of English. Amsterdam: Rodopi.

[48] Marian Owen. 1987. Evaluating automatic grammatical tagging of text. In *Newsletter of the International Computer Archive of Modern English (ICAME NEWS), No. 11*, pages 18–26. Norwegian Computing Centre for the Humanities, Bergen University.

[49] Rob Pocock and Eric Atwell. 1993. Extracting statistical grammars from the Lancaster–IBM Spoken English Corpus Treebank. Technical Report 93.29, School of Computer Studies, Leeds University.

[50] Rob Pocock and Eric Atwell. 1993. Probabilistic grammatical models for treebank-trained lattice disambiguation. Technical Report 93.30, School of Computer Studies, Leeds University.

[51] Paul Procter. 1978. *Longman Dictionary of Contemporary English.* London: Longman.

[52] Geoffrey Sampson. 1994. "personal communication".

[53] Beatrice Santorini. 1990. *Part-of-speech tagging guidelines for the Penn treebank project.* Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

[54] John Sinclair. 1987. *'Looking Up: An Account of the COBUILD Project in Lexical Computing.* Collins, Glasgow.

[55] Clive Souter. 1989. *A short handbook to the Polytechnic of Wales Corpus.* Bergen: Norwegian Computing Centre for the Humanities, Bergen University.

[56] Clive Souter. 1990. Systemic functional grammars and corpora. In J. Aarts and W. Meijs, editors, *Theory and Practice in Corpus Linguistics*, pages 179-211. Amsterdam: Rodopi.

[57] Clive Souter and Eric Steven Atwell. 1992. A richly annotated corpus for probabilistic parsing. In Carl Weir and Ralph Grishman, editors, *Proceedings of AAAI workshop on Statistically-Based NLP Techniques, San Jose, CA*, pages 28–38.

[58] Clive Souter. 1993. Harmonising a lexical database with a corpus-based grammar. In Souter and Atwell, editors, *Corpus-based Computational Linguistics*, pages 181–193. Amsterdam: Rodopi.

[59] Clive Souter. 1993. Towards a standard format for parsed corpora. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: design, analysis and exploitation; Proceedings of the 13th ICAME conference*, pages 197–214. Amsterdam: Rodopi.

[60] Clive Souter and Eric Steven Atwell. 1994. Using Parsed Corpora: A review of current practice In Nelleke Oostdijk and Pieter de Haan (eds), *Corpus-based Research Into Language*, pp143-158. Amsterdam, Rodopi.

[61] C. Sperberg-McQueen and L. Burnard. 1990. *Guidelines for the encoding and interchange of machine-readable texts, TEI P1*, Technical report, Universities of Chicago and Oxford.

[62] Jan Svartvik (ed). 1990. *The London-Lund Corpus of Spoken English: Description and Research.* Lund University Press, Lund, Sweden.

[63] L.J. Taylor and G. Knowles. 1988. *Manual of Information to Accompany the SEC Corpus.* Technical Report, Unit for Computer Research on the English Language,University of Lancaster, UK.

[64] Ni Yibin 1993. *The ICE Tagset - A Complete List of Tags used by the Tag-Selector for the Reference of Tag-Selectors and Researchers.* Technical Report, Department of English, University College London, UK.