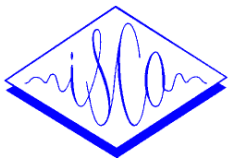

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

A MULTI-PHASE PARSING STRATEGY FOR UNRESTRICTED TEXT

A. I. C. Monaghan

Centre for Speech Technology Research, University of Edinburgh,
80, South Bridge, Edinburgh EH1 1HN, Scotland.

ABSTRACT

This paper describes work done on the Edinburgh University CSTR text-to-speech (TTS) system, a linguistically sophisticated speech output system based around a morph lexicon and a complex morphological decomposition module. The major problem with this and many other TTS systems is the lack of a reliable syntactic parse: this paper outlines a strategy designed to remedy that problem. The approach described here is, of course, still to be proven in application, but it is intended to produce a practical, efficient and flexible parsing strategy for unrestricted text by combining the best of both statistical and linguistic approaches.

1. PARSING FOR TEXT-TO-SPEECH CONVERSION

From the very crude linguistics-based island parsing of MITalk (Allen et al. 1987) to the highly-sophisticated statistical knowledge used in systems such as CLAWS and UCREL (Garside et al. 1987), almost every conceivable combination of parsing techniques has been applied to the problem of analysing unrestricted text. Until very recently, however, the criteria for deciding what parsing techniques would be implemented in a given TTS system had more to do with the researchers' interests in syntax than with the requirements of text-to-speech conversion: it is only in the last couple of years that work such as that reported in Fitzpatrick & Bachenko (1989) or Willemse & Boves (1989) has advocated sacrificing full syntactic parsing in order to achieve efficient extraction of the information most important to TTS systems, and even this work has so far concentrated in each case on applying a particular parsing technique which has been roughly tailored to the perceived needs of a TTS system.

There is an important unanswered question at the root of this approach: what does a TTS system require from syntax? The obvious things are word-class disambiguation ("Is it a noun or a verb?") and syntactic dependencies ("What does this NP dominate?" "Does the PP go with the noun or the verb?"): the former is required to determine stress and pronunciation for many orthographic forms, and the latter is assumed to be crucial for assigning prosody. However, it is clear that neither disambiguation nor dependency relations can be obtained from a purely syntactic analysis. For example, given an NP such as

(1) The damned

no amount of syntactic analysis can determine with certainty whether *damned* is a noun, an adjective or a verbal participle: it is an arbitrary choice depending on which rule the parser finds first. Similarly, the well-known example sentence

(2) I saw the man in the park with the telescope.

demonstrates the impossibility of assigning PPs a place in a syntactic tree on any principled basis, and hence the impossibility of determining what the NPs are in (2). These are serious problems for any full parse which relies on deterministic rules, but the point is that for any TTS system currently under development it doesn't much matter which of the possible analyses is chosen: the overall performance of the system will not be significantly altered. Unrestricted text includes much more problematic examples than these, of course, but the argument put forward here is that, at least until the end of this century, they are not worth worrying about.

What of the claim that prosody will suffer if such cases are not resolved? Aside from the fact that they CANNOT be resolved by syntax, there are various reasons for believing that the "correct" syntactic structure is not essential for producing good prosody. Firstly, it is widely accepted that prosodic structure is much flatter than syntactic structure (Pierrehumbert 1980, Selkirk 1984). There must therefore be levels of structure in syntactic analyses which are not relevant to prosody, i.e. a many-to-one syntax-to-prosody mapping, and if these levels are omitted in the syntactic analysis there will be no corresponding degradation in the prosodic realisation. Secondly, there is the long-standing problem of the one-to-many syntax-to-prosody mapping (Schmerling 1976, Selkirk 1984) which results in different accent patterns on what is syntactically the same sentence and provides the basis for the less-widely-accepted view that syntax has very little to do with prosody (Ladd 1980, Bolinger 1986) which is given at least lip service in many TTS systems (Quazza et al. 1989, Quené & Kager 1989, Monaghan 1990). According to this view, it is pragmatics and semantics which determine prosody and any correlation with syntax is an artefact of the syntax-semantics correlation. Thirdly, the syntax of spontaneous speech is known (Brown et al. 1984) to be much less complex and varied than that of written text: it is not clear that human readers actually realise the types of syntactic structure which can be found in technical writing, and indeed professional broadcasters make frequent errors in reading aloud from even moderately complex material with which they are unfamiliar. To what extent TTS systems should be expected to cope with text which was not designed to be spoken is a difficult question, but it is obviously unrealistic to expect them to perform better than humans and it may well be that users of any successful TTS system would simply not produce such text.
In any case, speech output systems must be able to respond reasonably to any input if they are to claim unrestricted applicability: even if this does not mean that they should read T. S. Eliot's poetry as well as Eliot himself would have done, it does mean that the parser should be failsafe and that the information available at every stage should be exploited to the full. The strategy which is proposed in the remainder of this paper is an attempt to do just that, and is justified if at all on purely pragmatist grounds.

2. A MULTI-PHASE PARSER

Given that the information necessary to produce a "perfect" acoustic realisation of a text sentence is not available to automatic systems, and given that the system must produce as acceptable a realisation as possible for any input text without ever failing altogether, it seems obvious that a simple phrase-structure parser will not suffice: such parsers are quite capable of failing for trivial reasons, and are prone to serious errors if given wordclass-ambiguous input of the type generally produced by TTS systems. There is also the question of punctuation, abbreviations, and other non-words. These problems need to be handled before any phrase-structure analysis is attempted, i.e. by some sort of pre-processor, so that the parse is guaranteed not to fail. As was stated above, a purely syntactic analysis cannot determine the attachment of constituents such as PPs or adverbs, and so this level of structure must also be supplied by heuristics. The final analysis must then be passed to phonetic or phonological modules, and there are bound to be elements of structure which the syntax has built up but which are irrelevant to the flatter, more linear prosodic structure: some sort of post-parse interface is therefore needed to ensure that the syntactic information is passed on with minimal redundancy. These observations are the basis for the CSTR multi-phase parsing strategy, which includes the following components:

PRE-PARSER: This phase makes use of reliable collocational and other statistical or heuristic information to remove needless word-class ambiguity (e.g. noun/verb ambiguity after determiners, main/aux verb ambiguities), and recognises and pre-process elements which the parser cannot handle (sentential adverbs, impermissible sequences (e.g. determiner + verb), clitics, punctuation, etc.). It is essential that the input to this phase from dictionaries, morphology, text pre-processors, etc. is optimised: for example, the word class of *damned* in (1) above could conceivably come out of some morphological analysis as four-ways (or more)

ambiguous (MAINVERB, PARTICIPLE, ADJECTIVE, NOUN) but in view of the practical limitations of the parser it is advisable to apply a morphological rule along the lines of

(3) MAINVERB + ed --> EDFORM

which would produce the unambiguous wordclass EDFORM as output and leave the parser to determine what type of phrase will be built from the syntactic context. An initial version of this pre-parser is implemented in our current system. The output of this phase should be guaranteed not to produce fatal errors in subsequent phases, and the pre-processed elements should be passed without further processing to the post-parse phase.

PHRASE-LEVEL PARSE: This needs to be failsafe, so it must be kept simple. An intelligent control structure (i.e. not just ordered rewrite rules) would also be advantageous: the problem of disambiguating noun/verb ambiguous items so as to identify the correct VPs requires a solution in terms of search strategies and control structure (stated informally, "Use breadth-first search, and look for verbs before nouns."). The depth of embedding, order of search, and so forth will ideally be variable, at least for development purposes until the rules have been satisfactorily tuned. The main purpose of this phase is to parse its input unambiguously into minimal constituents (NP, VP, PP). Each constituent may contain only one possible head, so that a sequence of three nouns produces three separate NPs: the general principle to be observed is that no spurious structure should be generated at this stage which has to be dismantled by subsequent processes. This phase is currently being implemented as a set of phrase-structure rules taking wordclass-ambiguous input and producing a string of constituents spanning the input in which all word-classes have been disambiguated. The disambiguation of word-class is determined largely on the basis of frequency information, in that if the most frequent word-class for that item results in a possible parse then that word-class will be taken. The phrase-structure rules are intentionally limited in coverage, so that only the most common phrase-types of English are covered (all other constructions must be handled by later phases) and distinctions such as that between adjectives and participles are not preserved: we have found (Monaghan 1990) that the effect of such distinctions on prosody is negligible, whereas their effect on parsing time is considerable.

CLAUSE-LEVEL PARSE: This will probably be based more on statistical than on syntactic knowledge, but given a phrase-level parse of the sort detailed above we can certainly make an intelligent guess at the location of clause boundaries. Together with a heuristic approach to the construction of major phrases and their attachments, a flattened clause structure will be produced. This phase will use information such as punctuation and verb subcategorisations, and can be as simple or as complex as is practical, although it must be failsafe and sensitive to the capabilities of the prosody modules. The first step is to collapse the minimal phrases identified in the phrase-level parse into larger phrases, and then the head of the clause must be identified: finally, pre- and post-modifiers will be attached according to subcategorisation information and general default rules. An initial version of this is under development, based on the assumption that there is one VP per clause. This is the phase where PP attachment and compounding are performed. Our current heuristic approach to PP attachment is simply to attach PPs in linear order and as low as possible in the tree, and this seems to produce reasonable prosody most of the time. *N*-noun compounds are a more serious problem, as identifying the head is virtually impossible except on a per-case basis (Sproat & Liberman 1987) and yet incorrect accent placement results in very low acceptability of output. Our current approach involves a version of the Compound Stress Rule (Chomsky & Halle 1968) with various exception clauses.

POST-PARSE: This phase is necessary to ensure compatability between syntax and prosody. Its main purpose is to remove any prosodically-irrelevant syntactic structure (e.g. internal structure of adjective phrases, complex prepositions and the like) which has been built up during parsing, and to ensure that the structure which is passed on to the prosodic rules is concise and coherent. This phase also slots adverbs, abbreviations, etc. back into place on the basis of the original linear order. The post-parse phase will subsume our current syntax-intonation mapping rules (Monaghan 1989), and will contain additional rules to integrate discourse-level information whenever this is available. The eventual shape of this phase depends to a very large extent on the details of the prosodic processing which it feeds, so

that notions of prosodic well-formedness and statistical heuristics would be equally appropriate in, say, determining at what level adverbs were attached.

3. CONCLUSIONS

The multi-phase parsing strategy discussed above is presented as an alternative to the single-pass or double-pass, purely linguistic or purely statistical parsers common in TTS systems. This strategy is designed both to maximise the use of linguistic and statistical knowledge at each phase and to be a development tool which can be extended as and when required: many phases are under construction already, and the other elements can be functioning/deliverable in a very short time but will allow for development over a longer term.
The pre-and post-parse stages above are clearly highly application-specific, in that they serve as interfaces between the syntactic processing and a specific system: the other phases, however, are seen as application- and domain-general, being limited to a core syntax and being relatively unambitious in the structures they produce. It is therefore anticipated that this strategy could be applied to any TTS system with the minimal problems of designing specific interfaces.
We make no apologies for the lack of theoretical syntactic motivation in this presentation: in our view, the role of syntax in TTS systems is largely as a woefully inadequate substitute for semantic and pragmatic analyses. We therefore consider the mixing of different approaches to parsing as perfectly justifiable insofar as they complement each other, and in the absence of discourse information we believe it is essential for a high-quality speech-output system to make use of all the available knowledge sources in analysing written text.

REFERENCES

Allen, J., S. Hunnicutt & D. Klatt (1987), *From Text to Speech: The MITalk System.* Cambridge: CUP.

Bolinger, D. (1986), *Intonation and its Parts.* Stanford: University Press.

Brown, G., A. Anderson, R. Shillcock & G. Yule (1984), *Teaching Talk.* Cambridge: CUP.

Chomsky, N. & M. Halle (1968), *The Sound Pattern of English.* New York: Harper & Row.

Fitzpatrick, F. & J. Bachenko (1989), "Parsing for Prosody: What a Text-to-Speech System Needs from Syntax", Proceedings of the Annual AI Systems in Government Conference, pp. 188-194. Washington DC: IEEE Computer Society Press.

Garside, R., G. Leech & G. Sampson (eds) (1987), *The Computational Analysis of English: A Corpus-Based Approach.* London: Longman.

Ladd, D. R. (1980), *The Structure of Intonational Meaning: Evidence from English.* Bloomington: Indiana University Press.

Monaghan, A. I. C. (1989), "Phonological Domains for Intonation in Speech Synthesis", Proceedings of Eurospeech 1989, vol. 1 pp. 502-505.

Monaghan, A. I. C. (1990), "Rhythm & Stress Shift in Speech Synthesis", *Computer Speech and Language* 4 (1), pp. 71-78.

Pierrehumbert, J. B. (1980), *The Phonology and Phonetics of English Intonation.* Ph.D. dissertation, MIT.

Quazza, S., G. Varese & E. Vivalda (1989), "Syntactic Pre-Processing for High Quality Text-to-Speech", Proceedings of Eurospeech 1989, vol. 1 pp. 506-509.

Quené, H. & R. Kager (1989), "Automatic Accentuation and Prosodic Phrasing for Dutch Text-to-Speech Conversion", Proceedings of Eurospeech 1989, vol. 1 pp. 214-217.

Schmerling, S. (1976), *Aspects of English Sentence Stress.* Austin: Texas University Press.

Selkirk, E. O. (1984), *Phonology and Syntax: The Relation between Sound and Structure.* Cambridge, Mass.: MIT Press.

Sproat, R. W. & M. Y. Liberman (1987), "Toward Treating English Nominals Correctly", Proceedings of the 25th Annual Meeting of the ACL, pp. 140-146.

Willemse, R. & L. Boves (1989), "Context Free Wild Parsing in a Speech-to-Text System", *Univ. Nijmegen Dept. of Language & Speech (Phonetics) Proceedings* 13, pp. 65-81.