Problems in Machine Conversion of Print to "Speech."†

Jane H. Gaitenby
Haskins Laboratories, New York City

Applied research that is linguistic in character is required by a variety of enterprises and institutions these days. The research to be reported here has been made under contract to the Prosthetic and Sensory Aids Service of the Veterans Administration. The V.A. has supported a number of research projects for the purpose of developing reading machines for the blind; and among them have been studies in electronic instrumentation, in psychology, in linguistics, and in cross-disciplinary fields. Each of these investigations has been concerned with the problem of converting the printed word to a tactile or auditory output. Part of the general research problem is optical in nature (because print must be recognized electronically), part deals with printed-symbol-to-sensory-symbol correspondence, and another part deals with human perception of units of sound. A specific problem area is the analysis of the structure of spoken English, in view of the fact that the printed word is speech at one symbolic remove. Since blind people in general are obliged to approach all handwritten or printed or illustrated material either through Braille or through the intercession of a human reader, a reading machine of some kind is an obvious need.

Leaving the consideration of cost aside, we can assume that the best reading machine will be one that converts printed text

---

† Given before the XIth Annual National Conference on Linguistics sponsored by the Linguistic Circle of New York, 12 March 1966.

to an output that is as much like real speech as possible. In short, the machine should talk. This ideal machine should produce completely natural sentences, and to do this it should have the ability to vary intonations and pauses appropriately for specific texts. Ideally, it should be replete with variable voice quality, such as one shade and timing of voice for business letters and another one for romantic novels -- or letters! But such a perfect machine would require a gigantic storage capacity, starting with tens of thousands of words. It would also require a remarkable program to manipulate its memory, to account for all manner of related nuances: grammatical, semantic and intonational -- in order to duplicate the associational memory and variable voice of the human being. Bear this ideal machine in mind -- and the fact that it is an ideal machine.

Speech has been synthesized at Haskins Laboratories and elsewhere by rules applying to very small speech units, of phonemic or syllabic size, generally. A logical developmental step was to experiment with sentence production from word units that had been pre-recorded by a human speaker. A device called the interim word reading machine was built for the V.A. at Haskins to test the feasibility of generating sentences from single spoken words. Although a successor to that machine is already under way, many of the problems encountered in the course of outfitting the first interim device with a vocabulary, and other equipment, remain; so I will confine my remarks to the machine that has already served much of its purpose, but is soon to become ancestral.

The well-named interim word reading machine at Haskins deals only with the word storage, retrieval and output side of the reading machine problem, and omits the optical scanning operation (which is not our contractual obligation). My concern has been to get words recorded for the machine's storage that

can be played out one after another, to(presumably) sound like sentences. The print-scanning function has been by-passed in the machine by simulation. The contents of a printed text are typed on a Flexowriter. The punched-tape resulting is simply a letter-to-digit conversion, and this is the form of information put into the machine. This is the input.

Now about the vocabulary storage in the machine. Each separate spoken word that has been previously recorded is stored on magnetic tape, and along with each word on the tape is its digital spelling, or code. When the code for a given word is sensed on the punched tape by the machine, that word is searched out on the stored spoken vocabulary magnetic tape. The word is matched and played back if it is stored, and is recorded on another tape at the same time, where it is added to the rest of the words in the order commanded by the original text. (If the word is not stored, it is spelled out, letter by letter, unfortunately.) When the contents of the entire text have been accumulated, this new tape, of spoken words, is played back to the listener. This is the output of the word reading machine, and the entire process is in fact, a conversion of print to sound. I will let you judge to what degree that sound is speech-like, before back-tracking over certain linguistic problems that the tape will illustrate. The sample you will hear now is about a minute long. It will be played for you at the rate in which the individual words were originally spoken and recorded. The long words which are spelled out at the end will probably surprise you. (By this I mean that you may have trouble understanding them.)

[PLAY PART I TAPE, "This is speech produced word by word..."]

That was an output of 81 words per minute, which is slow compared to normal speech -- although it is not a great deal

slower than my rate. This sample and the ones to follow later were compiled by a manual method, while the machine itself was still under construction; but these are the actual words from the 7200-unit vocabulary (of words, letters, numerals, punctuation, and a few suffixes) available to the machine's own storage. The original words were spoken by John T. Wadsworth, in a long series of short recording sessions made in the course of one year, for the most part.

After the words were recorded, the original tape was copied and edited. Then each word was separately mounted on a card like this one. (There are samples to be passed around.) These test sentences you have heard were generated by playing the word cards in sequence through a machine that plays tapes from cards instead of reels, and then each word was recorded in direct sequence on another machine. The sentences as such are therefore synthetic, since they were never spoken as sentences by the original speaker. The output you have heard, and those to come, demonstrate by negative evidence, that real speech is produced not word by word, but in continuous groups of words that are compatible in respect to tempo, loudness, and melody.

Every problem in applied research has constraints. For this machine, one -- and only one -- intonational version of each vocabulary word could be stored. This is a severe restriction, because in normal speech a given word may occur in many very, VERY different prosodic forms. Also, just one pronunciation was allowed per printed form, but homographs are a minor problem, compared to the fact that each word had to be stored in a single, frozen, stress and intonation form. It is clear that the one version recorded should be a highly probable spoken form for a highly probable type of occurrence in printed form.

We chose the vocabulary to be stored from the Dewey and the Thorndike and Lorge lists of the most frequently-printed English

words. There were no published data on probable spoken forms of the vocabulary words, although there were some helpful reports in the acoustic literature dealing with measurable correlates of stress and intonation, by such investigators as Bolinger, Fry and Denes.

Dr. F.S. Cooper of Haskins gave us a take-off push in the right general direction, and some fine instruments to work with. We made exploratory acoustic and perceptual tests of real and synthetic speech, and approached the machine's speech problem with the hypothesis that: very frequent phrases and polysyllabic words were structurally similar in a way that might be useful for us. That is, the syllables of a polysyllabic word have a persistent stress relationship, even in varied intonational environments, and the syllables of the most frequent phrases also tend to be regular in their stress relationships. We gradually learned something about the prosodic components of stress in words or phrases -- intensity, frequency and duration -- and decided to "program" our human speaker, if possible, to make the word recordings using prescribed stresses and intonations. Now the question was, what stress to prescribe for what. We turned to the study of probability of word occurrence in printed and spoken form.

The most frequent type of phrase in English texts is prepositional. Articles, prepositions and a connective are by far the most common English words. Most highly frequent words are monosyllabic. The stress of the most frequent words is usually very low. Most phrases begin with a preposition, as mentioned before, and most phrases and sentences end with a noun. Nouns carry much information and are generally prominent in the speech chain. And so on with the other grammatical classes...

We also examined so-called "intonation" and were soon convinced that it is acoustically reflected in durational shifts

as well as in frequency and intensity changes. Otherwise we knew very little except that syllables lengthen immediately before a pause, that pitch and intensity peaks usually start high and tend to decline toward the end of an utterance, and that pauses for punctuation are variable in length and of course have structural significance. (Some of these aspects of intonation almost certainly have a physiological base.)

Facts or observations of this sort, along with counts of form class sequence taken from texts in daily papers, books and periodicals, indicated that the spoken lexicon should be recorded on the basis of a word's grammatical function. Grammatical function is correlated to some extent with word stress category, and a trained speaker could, with effort, produce words at a prescribed relative pitch, length and loudness. An underlying assumption was that the stress prescriptions themselves would be valid, and that they would be consistently reproduceable by a human speaker on demand, and over a large span of time.

Hand-out One is a diagram of probable grammatical sequences in printed texts. The size of each circle indicates the projected relative prominence of a word as a given part of speech, used in writing the stress prescriptions. Although nearly all the form classes in English can be -- and often are -- preceded or followed by almost any of the form classes, the arrowed lines shown between any <u>two</u> classes stand for statistical likelihood of sequence.

Using the probability rationale for the manner in which word classes were to be spoken required that each of the 7200 words be classified as a member of a particular grammatical class, before it could be suitably recorded. This was a stumbling block.

Hand-out Two suggests the problem by showing a breakdown of the thousand most frequent words by <u>potential</u> membership in a

grammatical class. About half of the words can function in more than one role, and that role can be determined only by context. We therefore classified the multifunctional words by their most probable function, through educated guesswork and by intuition as native speakers. Impossible-to-classify words were put into the most neutral stress group, along with main verbs, whose stress in a sentence seems to be unpredictable. The prescriptions were written, and the recordings were made.

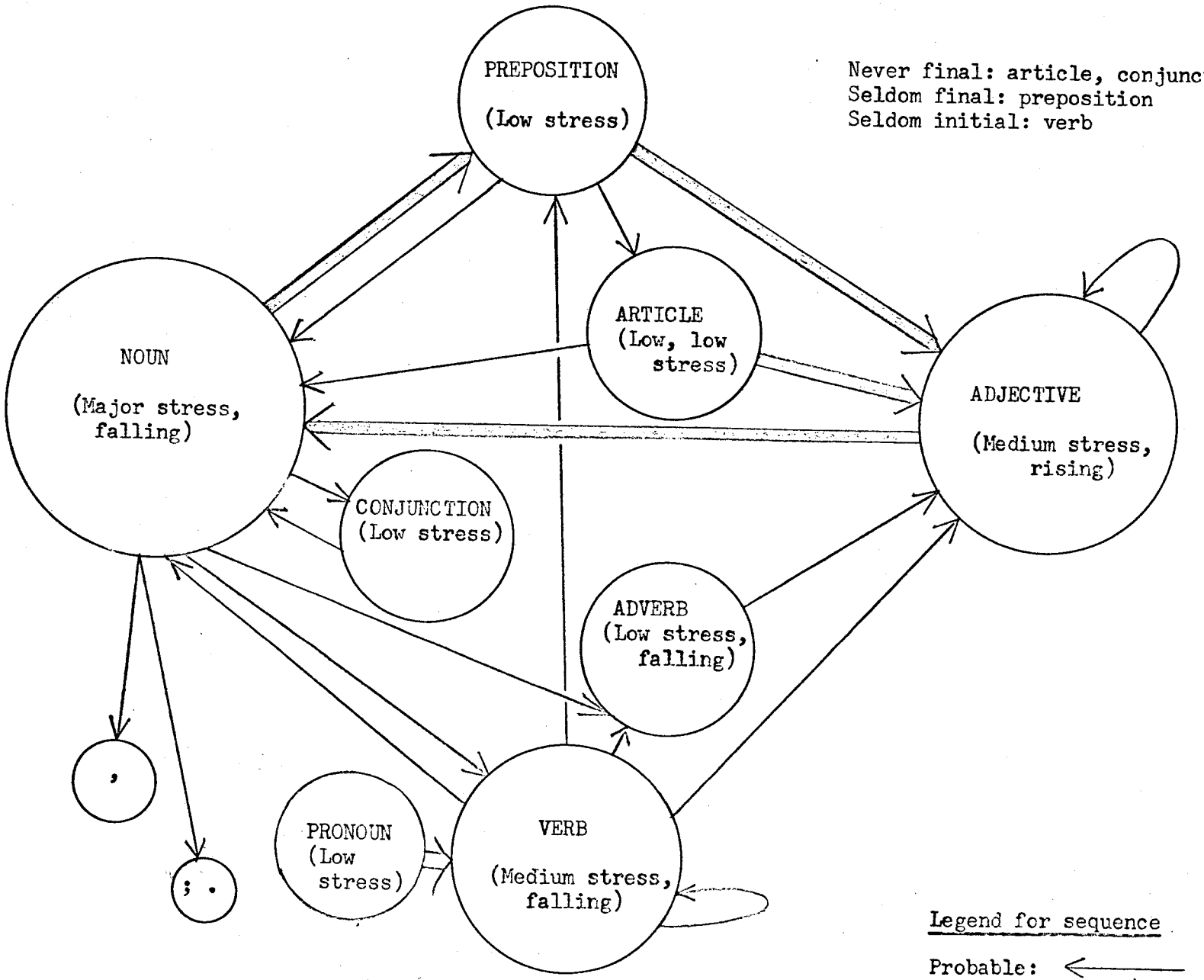
Again you may judge the output for yourselves, now bearing in mind that the sentences should be listened to not only to test the intelligibility, but also to note the words as stress types.

The recording speaker did follow the directions explicitly about 99% of the time, and the texts used have actually been a random test of the stored 7200 word vocabulary. The questions to think about when listening to the tapes are: To what degree is normal intonation approximated in these tapes, and in what kinds of cases is it least normal, and finally, what is normal intonation?

[Play Tape: Part II "Every city has its old guard.." (Current book sample, from Zinsser's The City Dwellers.)

Part III "The North Wind and the Sun.." (well-known phonetic exercize)]

HYPOTHESIZED MOST PROBABLE GRAMMATICAL AND INTONATIONAL SEQUENCES IN ENGLISH TEXTS



Grammatical Functions of the Most Frequent Words.

A preliminary breakdown was made of the commonplace functions possible for the 1027 most frequent words (i.e., written forms), as given by Dewey. The parts of speech used here as categories were Noun, Verb, Adjective, Adverb, Preposition, Connective, and Other (such as exclamatory word):

A) Words that have a single function,

Nouns	196
Verbs	140
Adj.	81
Adv.	48
Prep.	17
Conn.	8
Other	1
	491

B) Words that have two potential functions,

Noun or Verb	215	e.g., "being"
Noun or Adj.	120	"three"
Verb or Adj.	50	"open"
Adj. or Adv.	26	"better"
Adv. or Prep.	18	"to"
Adv. or Conn.	8	"since"
Conn. or Other	6	"though"
Adv. or Other	4	"really"
	447	

C) Words that have three potential functions,

Noun, Verb, Adj.	46	e.g., "present"
Noun, Adj., Adv.	7	"first"
Noun, Adj., Conn.	3	"either"
Verb, Adj., Prep.	3	"near"
	59	

D) Words that have four potential functions,

Noun, Verb, Adj., Adv.	6	"last, left, set, front, further, back"