

Listen, Hear!

by Geoff Plant

Learning From the 500 Most Commonly Occurring Words

Introduction

There are significant differences between the language forms we use when we write and when we speak. Crystal¹ pointed out that in spoken language, we use “looser construction, repetition, rephrasing, and comment clauses (e.g., “you know,” “mind you,” “as it were”),” and can rely on factors such as context and extralinguistic cues to resolve potential misunderstandings. In contrast, when we write we adopt a far more “correct” style, and pay far greater attention to producing grammatically correct sentences. If we spoke as we write, we would, almost certainly, be regarded as being overly formal and pedantic.

The Dahl Corpus

I took these differences into account when I started to search for lists of the most-frequently occurring English words². There are numerous lists of the most-frequently occurring words in written language, but, at first, I found it impossible to find a list that presented data from spoken language. Finally, in the mid-1990's, I found, quite by accident, a book by Hartvig Dahl³ entitled “Word frequencies of spoken American English.” Although its source was rather unusual, the transcripts of 255 psychoanalytic sessions conducted with 15 patients by 14 therapists, the more I looked at the lists, the more convinced I became that I had found a “gold mine” of useful information. As Dahl noted, the sessions provided the opportunity for the patients “to overcome inhibitions, to speak freely.. about the ordinary events and concerns of everyday life.”⁴ As a result, they reflect word usage patterns that reflect everyday communication.

I decided that I would concentrate my efforts on the first 500 words in Dahl's list, as these seemed to represent the “core” of spoken language. The importance of these words can be seen in the following analysis. The speakers in Dahl's corpus produced over a million words in total, of which there are 17,871 different words listed. A small number of these words contributed greatly to the overall pattern of use. For example, the personal pronoun “I” occurred on 61,586 separate occasions, and represented almost 6% of the total number of words produced. Further, the first 10 words in the list – I, and, the, to, that, you, it, of, a, know – represented 26% of the words used! When I looked at the first 100 words, I found that they represented almost 63% of the total, while the first 500 words formed 83.5% of the words used in the sessions. It was obvious to me that these words were of critical importance in spoken language, and, as a result, should be included in materials developed to test and train children with hearing loss. First of all, however, I wanted to conduct a detailed analysis of these 500 most-frequently-occurring words.

Syllabic structure

When I looked at the syllabic structure of the words, I found that there were 311 monosyllables, 151 two-syllable, 28 three-syllable, and 10 four-syllable words. I find that most deaf children have great

¹ Crystal, D. 1998. Speaking of writing and writing of speaking. **Longman Language Review**, 1, <http://www.awl-elt.com/dictionaries/> p. 1

² A list of the most-frequently-occurring words is usually called a corpus (plural corpora)

³ Dahl, H. 1979. **Word Frequencies of Spoken American English**, Verbatim, Essex, CT

⁴ Op cit, p. iv

trouble producing words of three or more syllables, and it's perhaps not surprising, given the paucity of such items in our "core" vocabulary.

Consonant distribution

In order to conduct this analysis, I transcribed each of these words into phonetic script, and then asked a native speaker of American English to check my transcriptions. Once we were agreed on the correct representations, I counted the total number of occurrences of each consonant, and expressed this as a percentage of the overall number of consonants in the list. These ranged from over 11% for [t] to no occurrences at all for the voiced fricative /zh/ found in words such as "treasure" and "pleasure." The order of occurrence from most frequently to least frequently occurring was /t, r, n, s, d, l, k, m, w, z, f, ng, p, b, h, g, v, th, th, sh, y, ch, j, zh/. Further analysis revealed that six items (/t, r, n, s, d, l/) represented 55% of all consonant occurrences. These involve a partial or complete constriction at, or near the alveolar ridge, by either the tongue tip or tongue body. There are few visual cues to these articulatory positions, and access to the acoustic cues accompanying their production, such as those which occur with cochlear implants, will greatly enhance acquisition of these vital consonants. It should also be noted that the consonants with the most visible articulatory positions, such as /f/, /p/, and /th/, occurred much less frequently.

Vowel distribution

Analysis of the distribution of the vowels was a little more difficult, as these vary from dialect to dialect. Again, however, it is sobering to reflect that the articulatory positions of the three most frequently occurring vowels (/i/ in "heed," /I/ in "hid," and /e/ in "head") all require an accurate raising and fronting of the tongue body. In order to produce these vowels accurately, speakers require access to the energy peaks (formants) that reflect tongue height (the first formant), and tongue place (the second formant). The second formants of these vowels lie in the range 1,500 – 3,000 Hz depending upon the age and gender of the speaker. This information may not be provided by hearing aids, but should be accessible to those using cochlear implants.

Monosyllabic structure

I also looked at the consonants (C) and vowels (V) making up the 311 one-syllable words in the list. I found that the most common form was CVC (143 occurrences), as in words such as "that," "was," and "but," followed by the CVCC ("don't," "think"), CV ("the," "to"), and VC ("it," "in") structures.

Contractions

One way that spoken communication differs from written communication involves the use of contractions. For example, while it is usual to write, "I am," we normally say, "I'm." In looking at first 500 items in the Dahl list, I found 35 contractions. The contractions "don't," "I'm," "it's," "that's," "didn't," "you're," and "I've," occurred in the first 100 items in the Dahl list.

Homophones

The Merriam-Webster Dictionary defines a homophone as "one of two or more words pronounced alike but different in meaning or derivation or spelling." Examples include words such as "two," "to," "too," and "four," "for," "fore." In looking at the Dahl list, I found more than 70 such words. These included the examples listed above, and other frequently occurring words such as "I," "knew," and "know." These words can be quite confusing for deaf children, and I always try to point them out to the children with whom I work.

Preparing testing and training materials

Once I had completed the analysis, I set out to prepare a series of testing and training materials using this set of words. I feel that deaf children need to become very familiar and comfortable with these items, as they represent such an important part of spoken English.

Speech Stuff

In 2001, I published a book⁵ containing the Dahl words used in a variety of ways. This included a listing of all 500 words with a pronunciation guide for each word. Where necessary, I included both the citation form of the word, and its more normal production in conversational speech. Other sections involved the use of the words for word and sentence testing, and a listing of the contractions and homophones.

Word lists

One possible use of these words included the preparation of word lists for testing and training. I selected 100 words from the CVC's in the list to form two 50-item test lists. I chose CVC's because they provide the opportunity for accurate scoring at both the word and phoneme level. When I present these words for identification, I score the subject for not only the number of words correctly identified, but also the number of phonemes. I have always believed that this is a much "fairer" method of scoring, because it recognizes that an error such as hearing "phone" as "foam," involves a minor confusion between two closely related phonemes. One of the lists is shown below.

1	did	11	head	21	feel	31	might	41	got
2	much	12	leave	22	one	32	came	42	will
3	come	13	less	23	him	33	yet	43	those
4	same	14	wife	24	been	34	look	44	done
5	lot	15	kid	25	with	35	said	45	has
6	bad	16	have	26	put	36	mean	46	that
7	like	17	had	27	home	37	then	47	was
8	which	18	such	28	guess	38	would	48	than
9	love	19	wrong	29	his	39	get	49	tell
10	type	20	but	30	down	40	them	50	this

When I present this list to one of my clients, I say each word in a short carrier phrase such as "Number one is did," "Number two is much," and score for the whole word (correct or incorrect), and the initial consonant, vowel, and final consonant. For example, if the client's response to the word "wrong" is "Ron," I score the word as incorrect, but note that the initial consonant and vowel were correct, while the final consonant involved substitution of /n/ for /ng/. At the completion of the list, I not only know how many words were correctly identified, but also have a record of the direction of error responses. This information is invaluable in determining areas of concern for an individual client, and provides cues as to the direction of future training.

The words can also be used for listening training, although, once this has occurred, they should not be used for test purposes. If a client is having difficulty hearing the words when they are produced in isolation, I put them in short sentences to see what effect context has on identification. For example, I say, "Number one is 'did.' Did you go for a walk today?" This sort of training can help develop an awareness of the importance of synthetic skills in speech understanding.



⁵ Plant, G. 2001. **Speech Stuff**, Hearing Rehabilitation Foundation, Somerville, MA

Sentence lists

I've also developed a number of sentence lists that contain only words from the first 500 items in the Dahl corpus. When I tell people about the development of these lists they often ask if this was a very difficult task. My response is always the same, "No, it's far more difficult to make up sentences that don't contain these words!" Although an exaggeration, it is quite close to the truth. These words are so important in English that it is difficult to imagine producing intelligible sentences that do not contain at least some of them. I've included some examples of these sentences below with the number of words in each sentence

1. Do you know where she went to high school?	9
2. I was asking them to help us with this work.	10
3. I should be able to see you some time tomorrow morning.	11
4. What time did you start?	5
5. I will see her in a couple of days.	9
6. I tried to read that book, but it was too hard for me.	13
7. That's the best thing to do.	6
8. I have to see the doctor.	6
9. What do you want to do next weekend?	8
10. I'm sure I put it in my room.	8

Each list consists of 25 sentences, and contains a total of 200 words. I've tried to include sentences of varying length, as deaf people often report that they have special difficulties with longer sentences. Readers who would like to obtain copies of these sentences can contact me at hearf@aol.com, and I will be happy to send them to you.

When I use these lists for training, I present each sentence in turn, and ask the client to repeat back as many words as possible. The presentation condition depends upon the skill level of the individual client, but can include auditory only, auditory-visual, in quiet, or in a noise background. The lists can also be used for speech training with deaf children, as the sheer ubiquity of these words means that they need to be within the production skills of deaf speakers. The child's production of these sentences can be recorded on either audio- or videotape, and then played back to listeners who are unfamiliar with the speaker. The listeners' scores when they attempt to repeat the child's sentences can be used to estimate her/his speech intelligibility, and can help pinpoint areas that require special attention. It is sometimes interesting to see what effect the presentation of the materials, both auditory only and auditory-visually, has on a listener's ability to understand the talker. Quite often, talkers who are difficult to understand via listening only, become much more intelligible when lipreading cues are also provided, and we should take this into account when measuring the skills of deaf speakers.

Conclusion

I've provided a brief overview of my analysis of the first 500 items in the Dahl corpus, but many readers may wish to find out more about this work. A more detailed paper⁶ is available, and I will send copies of it to anyone who contacts me at the above email address. Readers are also asked to send in topics that could be suitable for future issues.

⁶ Plant, G. 2000 An analysis of the most frequently occurring words in spoken American English. *Volta Review*, 101(2), 71 - 99