How Many Multiword Expressions do People Know?

Kenneth Church
HLT COE
Johns Hopkins University
Kenneth.Church@jhu.edu

Abstract

What is a multiword expression (MWE) and how many are there? What is a MWE? What is many? Mark Liberman gave a great invited talk at ACL-89 titled "how many words do people know?" where he spent the entire hour questioning the question. Many of these same questions apply to multiword expressions. What is a word? What is many? What is a person? What does it mean to know? Rather than answer these questions, this paper will use these questions as Liberman did, as an excuse for surveying how such issues are addressed in a variety of fields: computer science, web search, linguistics, lexicography, educational testing, psychology, statistics, etc.

1 How many words do people know?

One can find all sorts of answers on the web:

- **Very low**: Apparently I only knew 7,000 words when I was seven and 14,000 when I was fourteen. I learned from exposure. Now things are not that easy in a second language, but it just shows that the brain can absorb information from sheer input.[1]
- **Low**: 12,000 – 20,000 words[2]
- **Higher**: 988,968[3]
- **Even higher**: 13,588,391[4]

[1] http://thelinguist.blogs.com/how_to_learn_english_and/2009/02/how-many-words-do-you-know-how-many-have-you-looked-up-in-a-dictionary.html

[2] http://answers.yahoo.com/question/index?qid=20061105205054AA5YL0B

[3] http://www.independent.co.uk/news/world/americas/english-language-nears-the-one-millionword-milestone-473935.html

2 Motivation

As mentioned in the abstract, Liberman used his ACL-89 invited talk to survey how various fields approach these issues. He started his ACL-89 invited talk by questioning every word in the title of his talk: *How many words do people know?*

1. What is a word? Is a word defined in terms of meaning? Sound? Syntax? Spelling? White space? Distribution? Etymology? Learnability?
2. What is a person? Child? Adult? Native speaker? Language Learner?
3. What does it mean to know something? Active knowledge is different from passive knowledge. What is (Artificial) Intelligence? Is vocabulary size a measure of intelligence? (Terman, 1918)
4. What do we mean by many? Is there a limit like 20,000 or 1M or 13.6M or does vocabulary size (V) keep growing with experience (larger corpora → larger V)?

The original motivation for Liberman's talk came from a very practical business concern. At the time, Liberman was running a speech synthesis effort at AT&T Bell Labs. As the manager of this effort, Liberman would receive questions from the business asking how large the synthesizer's dictionary would have to be for such and such commercial application.

Vocabulary size was also a hot topic in many other engineering applications. How big does the dictionary have to be for X? X can be anything from parsing, part of speech tagging, spelling correction, machine translation, word breaking for Chinese and Japanese (and English), speech recog-

[4] Franz and Brants (2006)

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 137–144,
Portland, Oregon, USA, 23 June 2011.

nition, speech synthesis, web search or some other application.

3 Dictionary Estimates

These questions reminded Liberman of similar questions that his colleagues in lexicography were receiving from their marketing departments. Many dictionaries and other reference books lead with a marketing pitch such as: "Most comprehensive: more than 330,000 words and phrases [MWEs]…" (Kipfer, 2001).

The very smallest dictionaries are called "gems." They typically contain 20,000 words. Unabridged collegiate dictionaries have about 500,000 words.[5] The Oxford English Dictionary (OED) has 600,000 entries.[6]

All of these dictionaries limit themselves to what is known as general vocabulary, words that would be expected to be understood by a general audience. General vocabulary is typically contrasted with technical terminology, words that would only be understood by domain experts in a particular topic. There are reference books that specialize on place names (Gazetteers), surnames, technical terminology, quotations, etc., but standard dictionaries of general vocabulary tend to avoid proper nouns, complex nominals (e.g., "staff meeting"), abbreviations, acronyms, technical terminology, digit sequences, street addresses, trademarks, product numbers, etc.[7] Even the largest dictionaries may not have all that much coverage because in practice, one often runs into texts that go well beyond general vocabulary.

4 Broder's Taxonomy of Web Queries

Obviously, the web goes well beyond general vocabulary. Web queries tend to be short phrases (MWEs), often a word or two such as a product number. Broder (2002) introduced a taxonomy of queries that has become widely accepted. His percentage estimates were estimated from AltaVista query logs and could use updating.

- Naviational (20%)
- Informational (48%)
- Transactional (30%)

Navigational queries are extremely common these days, perhaps even more common than 20%. The user intent is to navigate to a particular url:

- google → www.google.com
- Greyhound Bus → www.greyhound.com
- American Airlines → www.aa.com

Broder's examples of informational queries are: *cars, San Francisco, normocytic anemia, Scoville heat units*. The user intent is to research a particular information need. The user expects to read one or more *static* web pages in order to address the information need. Broder italicized "static" to distinguish informational queries from transactional queries. Transactional queries are intended to reach a site where further (non-static) action will take place: shopping, directions, web-mediated services, medical advice, gaming, downloading music, pictures, videos, etc.

5 User Intent & One Sense Per Query

I prefer a two-way distinction between

1. Navigational queries: user knows where she wants to go, and
2. Non-navigational queries: user is open to suggestions.

Google, for example, offers the following "related search" suggestions for "camera:" *digital camera, video camera, history of the camera, sony camera, ritz camera, Nikon camera, camera brands, camera reviews, camera store, beach camera, canon, photography, bestbuy, camara, cannon, circuit city. camero. Olympus, camcorder, b&h.* These kinds of suggestions can be very successful when the user is open to suggestions, but not for navigational queries. There are a number of other mechanisms for making suggestions such as ads and did-you-mean spelling suggestions.

[5] http://www.collinslanguage.com/shop/english-dictionary-landing.aspx
[6] http://www.oed.com/public/about
[7] See Sproat (1994) and references therein for more on complex nominals. See Coker et al (1990) for coverage statistics on surnames. See Liberman and Church (1991) for more on abbreviations, acronyms, digit sequences and more. See Dagan and Church (1994) for more on technical terminology.

Pitler and Church (2009) used click logs to classify queries by intent (CQI). Consider five types of clicks. Some types of clicks are evidence that the user knows where she wants to go, and some are evidence that the user is open to suggestions.

1. Algo: clicks on the so-called 10 blue links
2. Paid: clicks on commercial ads
3. Wikipedia: clicks on Wikipedia entries
4. Spelling Corrections: did you mean …?
5. Other suggestions from search engine

Many queries are strongly associated with one type of click (more than others).

- Commercial queries → clicks on ads
- Non-commercial queries → Wikipedia.

There is a one-sense-per-X constraint (Gale et al, 1992; Yarowsky, 1993). It is unlikely that the same query will be ambiguous with both commercial and non-commercial senses. Indeed, the click logs show that both ads and Wikipedia are effective, but they have complementary distributions. There are few queries with both clicks on ads and clicks on Wikipedia entries. For a commercial query like, "JC Penney," it is ok for Google to return an ad and a store locator map, but Google shouldn't return a Wikipedia discussion of the history of the company.

Although the click logs are very large, they are never large enough. How do we resolve the user intention when the click logs are too sparse to resolve the ambiguity directly? Pitler suggested using word sense disambiguation methods. For example, her method labels the ambiguous query "designer trench" as commercial because it is closer (in random walk distance) to a couple of stores than to a Wikipedia discussion of trench warfare during World War I.

More generally, random walk methods (like word sense disambiguation) can be used to resolve all sorts of hidden variables such as gender, age, location, political orientation, user intent, etc. Did the user mean X? Does the user know what she wants, or is she open to suggestions?

5.1 User Intent & Spelling Correction

Spelling correction is an extreme case where it is often relatively easy for the system to determine user intent. On the web, spelling correction has become synonymous with did-you-mean. The synonymy makes it clear that the point of spelling correction is to get at what users mean as opposed to what they say.

> *Then you should say what you mean,' the March Hare went on.*
> *`I do,' Alice hastily replied; `at least--at least I mean what I say--that's the same thing, you know.'*
> *`Not the same thing a bit!' said the Hatter.* (Lewis Carroll, 1865)

See Kukich (1992) for a comprehensive survey on spelling correction. Boswell (2004) is a nice research exam; it is short and crisp and recent.

I've worked on Microsoft's spelling correction products in two different divisions: Office and Web Search. One might think that correcting documents in Microsoft Word would be similar to correcting web queries, but in fact, the two applications have remarkably little in common. A dictionary of general vocabulary is essential for correcting documents and nearly useless for correcting web queries. General vocabulary is more important in documents than web queries.

The surveys mentioned above are more appropriate for correcting documents than web queries. Cucerzan and Brill (2004) propose an iterative process that is more appropriate for web queries. In Table 1, they show a number of (mis)spellings of Albert Einstein's name from a query log, sorted by frequency: *albert einstein* (4834), *albert einstien* (525), *albert einstine* (149), *albert einsten* (27), *albert einsteins* (25), etc. Their method takes a web query that may or may not be misspelled and considers nearby corrections with higher frequencies. The method continues to iterate in this way until it converges at a fixed point. The iteration makes it possible to correct multiple errors. For example, *anol scwartegger* → *arnold schwartznegger* → *arnold schwarznegger* → *arnold schwarzenegger*. They find that context is often very helpful. In general, it is easier to correct

the combination of the first name and the last name together than separately. So too, it is probably easier to correct MWEs as a combination than to correct each of the parts separately.

5.2 User Intent & Spoken Queries

Queries often depend on context in complex and unexpected ways. It has been said that there is no there there on the web, but queries from cell phones are often looking for stuff that has a "there" (a location), and moreover the location is often near the user (e.g., restaurants, directions).

Users now have the option to enter queries by voice in addition to the keyboard. Kamvar and Beeferman (2010) found voice was relatively popular on mobile devices with "compressed" (hard-to-use) keyboards. They also found some topics were relatively more likely to be spoken:

- Food & Drink: *Starbucks, tuna fish, Mexican food*
- Business Listings: *Starbucks Holmdel NJ, Lake George*
- Properties relating to places: *weather Holmdel NJ, best gas prices*
- Shopping & Travel: *Rapids Water Park coupons, black Converse shoes, Costco, Walmart*

Other topics such as adult queries are relatively less likely to be spoken, presumably because users don't want to be overheard. Privacy is more of a concern for some topics and less for others.

6 What is "large"?

The term "large vocabulary" has been a moving target. Vocabulary sizes have been increasing with advances in technology. Around the time of Liberman's ACL-89 talk, the speech recognition community was working really hard on a 20,000-word task. Since it was so hard at the time to scale up recognizers to such large vocabularies, some researchers were desperately hoping that 20,000 words would be sufficient to achieve broad coverage of unrestricted language.

At that time, I gave a workshop talk that used a much larger vocabulary of 400,000 words (Church and Gale, 1989). A leading researcher pulled me aside and begged me to tell him that I had made a mistake and there was no need to go beyond 20,000 words.

Similar questions came up when Google released their ngram counts over a trillion word corpus (Franz and Brants, 2006). There was considerable pushback from the community over the size of the vocabulary (13,588,391). Norvig (personal communication) called me up to ask if their estimate of 13.6 million seemed unreasonable.

While I had no reason to question Google's estimate, I was reluctant to make a strong statement, given Efron and Thisted (1976). Efron and Thisted studied a similar question: How many words did Shakespeare know (but didn't use)? They conclude that one can extrapolate corpus size a little bit (e.g., a factor of two) but not too much (e.g., an order of magnitude). Since Google is working with corpora that are many orders of magnitude larger than what I had the opportunity to work with, it would require way too much extrapolation to answer Norvig's question based on my relatively limited experience.

7 Vocabulary Grows with Experience

Many people share the (mistaken) intuition that there is an upper bound on the size of the vocabulary. Marketing pitches such "330,000 words" (above) suggest that there is a reasonable upper bound that a person could hope to master (something considerably more manageable than Google's estimate of 13.6 million).

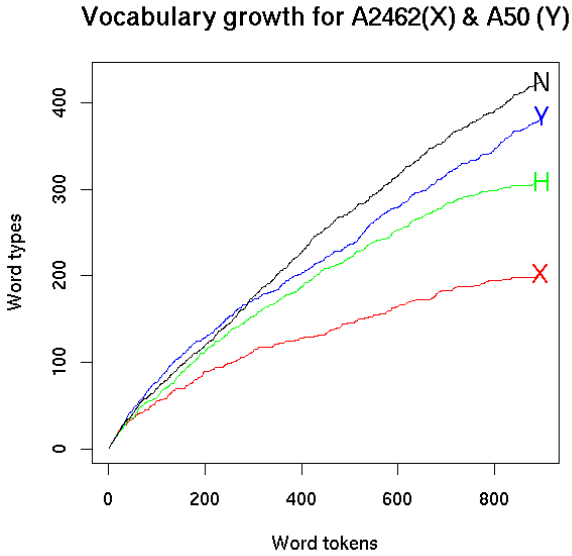
In fact, the story is probably worse than that. At ACL-1989, Liberman showed plots like those below.[8] These plots make it clear that vocabulary (V) is going up and up and up with corpus size (N). There appears to be no end in sight. It is unlikely that there is an upper bound. 20k isn't enough. Nor is 400k, or even 13.6 million…

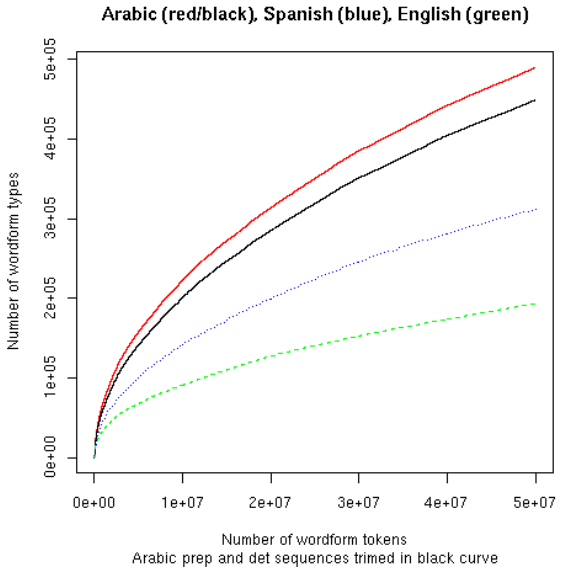
The different curves call out differences in what counts as a word. Do we consider morphologically related forms to be one word or two? How about

[8] Plots borrowed with permission from Language Log: http://itre.cis.upenn.edu/~myl/languagelog/archives/005514.html

upper and lower case? MWEs? The different curves correspond to different choices.

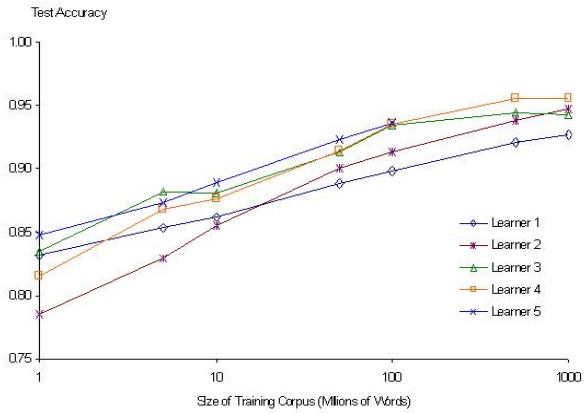
No matter how we define a word, we find that vocabulary grows (rapidly) with corpus size for as far as we can see. This observation appears to hold across a broad set of conditions (languages, definitions of word/ngram, etc.) Vocabulary becomes larger and larger with experience. Similar comments apply to ngrams and MWEs.





There is wide agreement that there's no data like more data (Mercer, 1985).[9] Google quoted Mercer in their announcement of ngram counts (Franz and Brants, 2006).

Banko and Brill (2001) observed that performance goes up and up and up with experience (data). In the plot below, they note that the differences between lines (learners) are small compared to the gains to be had by simply collecting more data. Based on this observation, Brill has suggested (probably in jest) that we should fire everyone and spend the money on collecting data.



Another interpretation is that experience improves performance on a host of tasks. This pattern might help account for the large correlation (0.91) in Terman (1918). Terman suggests that vocabulary size should not be viewed as a measure of intelligence but rather a measure of experience. He uses the term "mental age" for experience, and measures "mental age" by performance on a standardized test. After adjusting for the large correlation between vocabulary size and experience, there is little evidence of a connection between vocabulary size and intelligence (or much of anything else). Terman also considers a number of other factors such as gender and the language spoken at home (and testing errors), but ultimately concludes that experience dominates all of these alternative factors.

[9] Jelinek (2004) attributes this position to Mercer (1985) http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf.

8 What is a Word? MWE?

We tend to think that white space makes it pretty easy to tokenize English text into words. Obviously, white space makes the task much easier than it would be otherwise. There is a considerable literature on word breaking in Chinese and Japanese which is considerably more challenging than English largely because there is no white space in Chinese and Japanese. There are a number of popular dictionary-based solutions such as ChaSen[10] and Juman.[11] Sproat *et al* (1996) proposed an alternative solution based on distributional statistics such as mutual information.

The situation may not be all that different in English. English is full of multiword expressions. An obvious example involves words that look like prepositions: *up, in, on, with*. A great example is often attributed to Winston Churchill: *This is the sort of arrant nonsense up with which I will not put*.[12] One could argue that "put up with" is a phrasal verb and therefore it should be treated more like a fixed expression (or a word) than a stranded preposition.

8.1 Preventing Bloopers

Almost any high frequency verb (*go, make, do, have, give, call*) can form a phrase with almost any high frequency function word (*it, up, in, on, with, out, down, around, over*), often with non-compositional (taboo) semantics.

This fact led to a rather entertaining failure mode with a word sense program that was trained on a combination of Roget's Thesaurus and Grolier's Encyclopedia (Yarowsky, 1992). Yarowsky's program had a tendency to label high frequency words incorrectly with taboo senses due to a mismatch between Groliers and Roget's. Groliers was written for the parents of middle-American school children and therefore avoided taboo language, whereas Roget's was edited by Chapman, an authority on American Slang (taboo language). The mismatch was particularly nasty for high frequency words, which are very common in Groliers, but unlikely to be mentioned in Roget's, except when the semantics are non-compositional (taboo). Consequently, there was an embarrassingly high probability that Yarowsky's program would find embarrassing interpretations of benign texts.

While some of these mistakes are somewhat understandable and even somewhat amusing in a research prototype, such mistakes are no laughing matter in a commercial product. The testers at Microsoft worked really hard to make sure that their products don't make inappropriate suggestions. Despite their best efforts, there have been a few highly publicized mistakes[13] and there will probably be more unless we find better ways to prevent bloopers.

8.2 Complex Nominals and What is a Word?

Complex nominals are probably more common than phrasal verbs. Is "White House" one word or two? Is a word defined in terms of spelling? White space?

These days, among computational linguists, there would be considerable sympathy for using distributional statistics such as word frequency and mutual information to find MWEs. Following Firth (1957), we know a word by the company that it keeps. In Church and Hanks (1990), we suggested using pointwise mutual information as a heuristic to look for pairs of words that have non-compositional distributional statistics. That is, if the joint probability, P(x,y), of seeing two words together in a context (e.g., window of 5 words) is much higher than chance, P(x)P(y), then there is probably a hidden variable such as meaning that is causing the deviation from chance. In this way, we are able to discover lots of word associations (e.g., *doctor…nurse*), collocates, fixed expressions, etc.

If the list of MWEs becomes too large and too unmanageable, one could turn to a method like Stolcke pruning to cut back the list as necessary. Stolcke pruning is designed to prune ngram models so they fit in a manageable amount of memory. Suppose we have an ngram model with too many

[10] http://chasen.naist.jp/hiki/ChaSen/
[11] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html
[12] For a discussion of the source of this quotation, see http://itre.cis.upenn.edu/~myl/languagelog/archives/002670.html .
[13] http://www.zdnet.co.uk/news/it-strategy/1999/07/01/microsoft-sued-for-racist-application-2072468/

ngrams and we have to drop some of them. Which ones should we drop? Stolcke pruning computes a loss in terms of relative entropy for dropping each ngram in the model. The method drops the ngram that minimizes loss.

When an ngram is dropped from the model, that sequence is modeled with a backed off estimate from other ngrams. Stolcke pruning can be thought of as introducing compositionality assumptions. Suppose, for example, that "nice house" has more compositional statistics than "white house." That is, Pr(*nice house*) ≈ Pr(*nice*) Pr(*house*) whereas Pr(*white house*) >> Pr(*white*) Pr(*house*). In this case, Stolcke pruning would drop "nice house" before it drops "white house."

8.3 Linguistic Diagnostics

Linguists would feel more comfortable with defining word in terms of sound (phonology) and meaning (semantics). It is pretty clear that "White House" has non-compositional sound and meaning. The "White House" does not refer to a house that happens to be white, which is what would be expected under compositional semantics. It is accented on the left (the WHITE house) in contrast with the general pattern where adjective-noun complex nominals are typically accented on the right (a nice HOUSE), though there are many exceptions to this rule (Sproat 1994).[14]

Linguists would also feel comfortable with diagnostic tests based on paraphrases and transformations. Fixed expressions are fixed. One can't paraphrase a "red herring" as "*herring that is red." They resist regular inflection: "*two red herrings." In Bergsma *et al* (2011), we use a paraphrase diagnostic to distinguish [N & N] N from N & [N N]:

- [*dairy and meat*] *production*
    - *meat and dairy production*
    - *production of meat and dairy*
    - *production de produits* [*laitiers et de viand*] (French)
- *asbestos and* [*polyvinyl chloride*]
    - *polyvinyl chloride and asbestos*
    - *asbestos and chloride*
    - *l'asbesto e il* [*polivinilcloruro*] (Italian)

The first three paraphrases make it clear that "dairy and meat" is a constituent whereas the last three paraphrases make it clear that "polyvinyl chloride" is a constituent. Comparable corpora can be viewed as a rich source of paraphrase data, as indicated by the French and Italian examples above.

9 Conclusions

How many multiword expressions (MWEs) do people know? The question is related to how many words do people know. 20k? 400k? 1M? 13M? Is there a bound or does vocabulary size increase with experience (corpus size)? Is vocabulary size a measure of intelligence or just experience?

Dictionary sizes are just a lower bound because they focus on general vocabulary and avoid much of what matters on the web. Spelling correction is not the same for documents of general vocabulary and web queries.

One can use Stolcke pruning and other compositionality tricks to cut down on the number of the number of multiword units that people must know. But obviously, the number they must know is just a lower bound on the number they may know.

There are lots of engineering motivations for wanting to know how many words and MWEs people know. How big does the dictionary have to be for X (where X is parsing, tagging, spelling correction, machine translation, word breaking for Chinese and Japanese (and English), speech recognition, speech synthesis or some other application)?

Rather than answer these questions, this paper used these questions as Liberman did, as an excuse for surveying how such issues are addressed in a variety of fields: computer science, web search, linguistics, lexicography, educational testing, psychology, statistics, etc.

[14] Sproat has posted a list of 7831 English binary noun compounds with hand assigned accent labels at: http://www.cslu.ogi.edu/~sproatr/newindex/ap90nominals.txt

References

Harald Baayen (2001) *Word Frequency Distributions*. Kluwer, Dordrecht.

Michele Banko and Eric Brill. (2001) "Scaling to very very large corpora for natural language disambiguation," ACL.

Shane Bergsma, David Yarowsky and Kenneth Church (2011), "Using Large Monolingual and Bilingual Corpora to Improve Coordination Disambiguation," ACL.

Dustin Boswell (2004) "Speling Korecksion: A Survey of Techniques from Past to Present," http://dustwell.com/PastWork/SpellingCorrectionResearchExam.pdf

Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (September 2002), 3-10.

Lewis Carroll, 1865, Alice's Adventures in Wonderland.

Kenneth Church and William Gale (1989) "Enhanced Good-Turing and Cat-Cal: two new methods for estimating probabilities of English bigrams." HLT.

Kenneth Church and Patrick Hanks. (1990) "Word association norms, mutual information, and lexicography." CL.

Cecil Coker, Kenneth Church and Mark Liberman (1990) "Morphology and rhyming: two powerful alternatives to letter-to-sound rules for speech synthesis," In ESCA Workshop on Speech Synthesis *SSW1-1990*, 83-86.

Silviu Cucerzan and Eric Brill (2004) "Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users, EMNLP.

Ido Dagan and Kenneth Church. (1994) "Termight: identifying and translating technical terminology," ANLC.

William Gale, Kenneth Church and David Yarowsky. (1992) "One sense per discourse," HLT.

Bradley Efron and Ronald Thisted, (1976) "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*, 63, 3, pp. 435-447.

John Firth, (1957) "A Synopsis of Linguistic Theory 1930-1955," in Studies in Linguistic Analysis, Philological Society, Oxford; reprinted in Palmer, F. (ed.) 1968 Selected Papers of J. R. Firth, Longman, Harlow.

Alex Franz and Thorsten Brants (2006) http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html.

Fred Jelinek (2004) "Some of my Best Friends are Linguists," LREC.

Maryan Kamvar and Doug Beeferman, (2010) "Say What? Why users choose t speak their web queries," Interspeech.

Barbara Kipfer (ed.) (2001) Roget's Thesaurus, Sixth Edition, HarperCollins, NY, NY, USA.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Comput. Surv.* 24, 4.

Mark Liberman (1989) "How many words do people know?" *ACL*.

Mark Liberman and Kenneth Church (1991). "Text analysis and word pronunciation in text-to-speech synthesis." In *Advances in Speech Signal Processing*, edited by S. Furui and M. Sondhi.

Frederick Mosteller and David Wallace. *Inference and Disputed Authorship: the Federalist*, Addison-Wesley, 1964.

Emily Pitler and Kenneth Church. (2009) "Using word-sense disambiguation methods to classify web queries by intent," EMNLP.

Richard Sproat, William Gale, Chilin Shih, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese, CL.

Richard Sproat (1994) "English noun-phrase accent prediction for text-to-speech," Computer Speech and Language, 8, pp. 79-94.

Andreas Stolcke (1998) "Entropy-based Pruning of Backoff Language Models" Proc. DARPA News Transcription and Understanding Workshop.

Lewis Terman, (1918) "The vocabulary test as a measure of intelligence," *Journal of Educational Psychology*, Vol 9(8), pp. 452-466.

David Yarowsky. 1993. One sense per collocation, HLT.