Corpus-based metrics for assessing communal common ground

Roman Kutlak (r04rk9@abdn.ac.uk)

Kees van Deemter (k.vdeemter@abdn.ac.uk)

Chris Mellish (c.mellish@abdn.ac.uk)

Computing Science Department, University of Aberdeen
Aberdeen AB24 3UE, Scotland, UK

Abstract

This article presents the first attempt to construct a computational model of common ground. Four corpus-based metrics are presented that estimate what facts are likely to be in common ground. The proposed metrics were evaluated in an experiment with human participants, focussing on a domain of famous people. The results are encouraging: two of the proposed metrics achieved a large positive correlation between the estimates of how widely known a property of a famous person is and the percentage of participants who knew the corresponding property.

Keywords: Common Ground; Common Knowledge; Mutual Knowledge; Evaluation with human subjects; Web as corpus

Introduction

Assessing other people's knowledge is crucial in many situations. Teachers, for example, do well to highlight information that their pupils do not know. Examples in other areas abound. Suppose, for example, we want to persuade you to reduce your intake of butter. We might do this by telling you ”butter gives you high cholesterol”. This argument only works if you, the hearer, know that cholesterol is bad for you, as is often assumed, for instance because it raises the likelihood of heart disease. The (presumed) fact that cholesterol is bad for you happens to be well publicised, and this might be what lies behind our assumption that you know it. Similar examples obtain in advertising, where companies might persuade you to buy a toothpaste by saying it contains fluoride, because they assume that many viewers know that fluoride is good for your teeth. It is often important to distinguish between knowledge and belief, but we will focus on cases where the distinction is less than crucial.

The difference between information assumed to be “given” (i.e., known by the hearer) and “new” (i.e., privileged information of the speaker) is crucial to philosophers, logicians and linguists (Frege (1892 (1952)); Strawson (1952); Van Eijck (1993), to mention but a few) and it is highly relevant to computational linguists working on Natural Language Generation (NLG) programs (Reiter & Dale, 2000), whose output is meant to mimic human language use. A central example is the generation of referring expressions, which has been studied extensively over the last 20 years (Krahmer & van Deemter, 2012). For example, an NLG program that aims to identify a person would do well to express properties that are likely to be known by the reader. For example, the expression “the former member of Led Zeppelin” would not be very informative to a hearer who has never heard of Led Zeppelin.

To the best of our knowledge, no general computational models exist for assessing what knowledge is likely to be known. In this paper, we examine a corpus-based strategy for building such a computational model. But, before we go into the details of our approach, there are some terminological and conceptual issues to be clarified. Common and mutual knowledge have been defined in different ways. In this paper, we shall follow the terminology of Vanderschraaf and Sillari, which the authors clarified with the following example.

> Suppose each student arrives for a class meeting knowing that the instructor will be late. That the instructor will be late is mutual knowledge, but each student might think only she knows the instructor will be late. However, if one of the students says openly, “Peter told me he will be late again”, then the mutually known fact is now commonly known. Vanderschraaf and Sillari (2009)

Thus, mutual knowledge is knowledge shared by a group of people. Common knowledge might be informally characterised as knowledge that is *publicly* shared by a group of people. Slightly more precisely, A and B have *mutual knowledge* of p if and only if A knows p and B knows p. They have *common knowledge* of p if they have mutual knowledge of p, and A knows that B knows p, and B knows that A knows p, and A knows that B knows that A knows p, and so on, *ad infinitum* (Lewis, 1969). Logicians and game theorists have proposed various precise definitions of common knowledge (including cases with more than two knowers), typically cast in epistemic logic, which formalise the “ad infinitum” (above) in different ways (Vanderschraaf and Sillari (2009)). For reasons that will become clear later, we use a third term that is often used in this connection, *common ground*, in a loose sense, when the distinction between mutual and common knowledge is irrelevant.

The psychologists Clark and Marshall observed that, in simple situations, common knowledge is enforced by “triple co-presence”, where the speaker, the hearers and entities are physically present and the speaker believes that the hearers attend to the entities (Clark and Marshall (1981)). They contrast this simple situation (which they call *personal* common ground) with *communal* common ground, which arises not from physical co-presence but from being in a shared community (e.g., people living in Paris). Speakers are frequently able to distinguish between knowledge that is available to members of such communities or to outsiders (Jucks, Becker, & Bromme, 2008; Nickerson, Baddeley, & Freeman, 1987). Personal common ground comes from joint personal experience of the agents; communal common ground derives from a range of sources, including the likelihood of common experience. For example, suppose Paris residents see the Eiffel Tower as they travel around their city. As they have no reason to believe that others do not see the same Eiffel Tower, they can exploit the knowledge shared by the residents of Paris namely, that the Eiffel Tower is in Paris. They can use this mutual knowledge as a shared basis for the communal common ground. In this case, the community are the Paris residents.

Our purpose was not to design a new theory of common ground, but to estimate what atomic facts are likely to be in (communal) common ground, using a corpus-based method. The method is based on metrics that use the frequency of information in a corpus to predict how widely known the information is. If it is successful, the method could be used to model different communities by studying different corpora. In other words, we offer a *parametrised* model, that has a corpus (or, equivalently, a community) as its parameter.

We focus in this paper on reference to famous people, because famous people are a prime instance of something people actually have common knowledge about. The proposed metrics will be used as one of a number of heuristics for selecting the content of a description of a famous person.

Estimating Common Ground

The proposed heuristic simplistically hypothesises that facts that often co-occur in a document are more likely to be known because of frequency and repetition effects on memory (Atkinson & Shiffrin, 1968). Additionally, if a fact occurs frequently in a very large corpus (such as the world-wide web), which has many authors, then this implies that many people (i.e., many authors) know this fact. It seems plausible that people write about what they themselves know but, to our knowledge, no one attempted to examine how much can shallow corpus methods (i.e., methods not involving semantic analysis) tell us about mutual knowledge (Vanderschraaf & Sillari, 2009).

Assessing mutual knowledge is of substantial interest, and practical use, in its own right. So, how about common knowledge? If a fact occurs frequently in a a corpus, is this evidence for common knowledge (as opposed to just mutual knowledge)? Clark (1996) essentially answered this question in the affirmative. He suggested that instead of thinking directly in terms of the knowledge of others, people use evidence as a basis for common ground (as in our example of the Eiffel Tower). Nickerson et al. (1987) and a series of studies performed by R. Krauss and S. Fussell (Fussell & Krauss, 1991; Krauss & Fussell, 1991) showed that people often use their own knowledge to estimate what others know. Given that it would not make much sense for computers to use “their own” knowledge, an alternative source of knowledge has to be utilised, and we have chosen corpora as such a source of knowledge. Corpora in combination with measures of association were previously used in distributional models of semantic representation, where the main assumption is that if words appear in a similar context they have a similar meaning (Firth, 1957). Riordan and Jones (2011) examined in detail several distributional and feature-based models and concluded that they performed similarly well. Jurafsky (2003) argued that probabilistic modelling, as used by computational linguists, can effectively model some of the phenomena observed by psychologists. Given the past success of probabilistic models applied to various tasks including distributional semantics (Baroni & Lenci, 2010), we believed that the use of these techniques as a tool for estimating common ground was at least promising.

Measures of association

Below, we list some of the main metrics that have been proposed for measuring the strength of association between words. These metrics assume that a *context* for the words has been defined. The context is frequently defined as a limited number of words before or after the target word or a short frame such as a paragraph in which the target word occurs. These contexts are not suitable for our purpose, because a fact about a person can be mentioned further away from the person’s name, especially if the name is pronominalised in consequent paragraphs. Instead, we will use an article as a context for our search. This can be, for example, a news article or a Wikipedia article.

Frequency The simplest measure of association between a person and a property (a fact about a person) is the frequency of occurrence of the name and the property together in a corpus. Taking a collection of documents as a corpus, frequency corresponds to the count of articles that contain the name and the property. This association is then the value of $count(n,p)$ where n stands for the name of an entity and p is the property in question.

Conditional Probability A more sophisticated measure is conditional probability calculated as (1) and (2), where (1) measures the probability of the name given a property, and (2) measures the probability of a property given the name of the person. While the former measure normalises the results by the frequency of the property, the later measure takes into account how famous each person is.

$$assoc_{prob}(n,p) = P(n|p) = \frac{count(n,p)}{count(p)} \quad (1)$$

$$assoc_{prob}(p,n) = P(p|n) = \frac{count(p,n)}{count(n)} \quad (2)$$

Pointwise Mutual Information (PMI) (Fano, 1961) is a measure that compares how often two events x and y occur together. PMI exploits the fact that if two terms appear together often their joint probability ($P(n,p)$) will be higher than if they were independent ($P(n)P(p)$). The value of PMI is positive for terms that co-occur and negative otherwise.

$$assoc_{PMI}(n,p) = log_2 \frac{P(n,p)}{P(n)P(p)} \quad (3)$$

One problem with PMI is that infrequent words that only appear together achieve a disproportionately high score. In order for a property to be in common ground, it also has to be frequently mentioned. To mitigate the problem, (Hodges, Yie, Reighart, & Boggess, 1996) suggest multiplying each PMI score by $count(n,p)$. To reduce the big difference between the numbers of documents and to take into consideration the association as measured by PMI as opposed to the mere count, we multiply the PMI scores by the square root of the count. The final formula used for calculating the association is given by (4). Our pilot experiment showed better results with the adjusted PMI metric and any subsequent reference to PMI refers to (4).

$$assoc_{PMI}(n,p) = \sqrt{count(n,p)} * log_2 \frac{P(n,p)}{P(n)P(p)} \quad (4)$$

Search Engine as Corpus

How do we acquire the frequencies n and p in the metrics above? Turney (2001) successfully used the AltaVista search engine to measure association between words using a variation of PMI called PMI-IR, where he used numbers of hits returned by the search engine instead of real corpus probabilities. The number of hits corresponds to the number of documents on the Internet that contain the search term. The functionality of providing the number of hits is available from other search engines. Google does not respond to queries from programs other than web browsers but offers Google Custom Search which allows programmers to achieve the same functionality upon registration. Note that each of the search engines only searches a subset of all the documents available on the Internet and these subsets can differ substantially. This is also the case for Google accessed from a web browser and from a Google Custom Search.

There has been a debate as to whether to use search engines for research purposes (Kilgarriff, 2007; Pedersen, 2008). One of the arguments against using search engines was that the queries are optimised by performing morphological adjustments such as stemming and by looking up related words or synonyms. While these issues are pertinent to lexicography, they seem to be of less importance when it comes to establishing an association between a person and a property. A property can be described by different words and so such optimisations can in fact be very useful. There can be a problem with morphological changes to names but many search engines also offer search for exact phrases that are not morphologically or semantically manipulated and so we can avoid optimisations at places where they are undesirable. This is usually achieved by embedding the searched string in quotes.

Another problem that comes with the usage of search engines is the fact that we do not know the number of searched documents. This is necessary for calculating the probabilities used by the PMI metric. We have chosen a large constant $N = 1.0e12$ for normalising the counts (the number of search results for the word *the* is about 25 billion on Google and about 10 billion on AltaVista and Bing and 2.4 billion on Google custom search).

Experiment

We performed an experiment to evaluate how well the different heuristics perform. Given a person and a set of properties, our heuristics produce a set of $\langle property : score \rangle$ pairs, where the *score* for each property is calculated by one of the described measures of association. Our goal is to assign scores so that they reflect the commonality of a particular property with regard to the name. This means that properties that are often associated with a name (e.g., Isaac Newton was a physicist) should get a higher score than properties that are less frequently associated with the name (e.g., Isaac Newton was the warden of the Royal Mint).

We used hearers' individual knowledge to asses how well the proposed heuristics perform. More specifically, the participants viewed statements such as "*Andy Warhol was American*" and "*Ernest Hemingway is the author of For whom the bell tolls*" and were asked to select one of the following statements: *true*, *false* or *don't know*. Our hypothesis is that **when a metric assigned a property higher score, a higher proportion of participants should give an affirmative answer** (i.e., state that the sentence involving the property is true). The success of the metric is measured as a Spearman correlation between the output of the metric and the percentages of affirmative answers assigned to the individual statements by the participants.

Heuristic Options and Pilot

Aside from the choice of metric, several other choices had to be made. The first choice was which search engine to use. As we had no reason to believe that a particular search engine will perform better than others, our pilot tested the metrics on the three major search engines: AltaVista (Yahoo), Bing and Google.

The second choice is what search terms to choose. Most properties can be expressed as a combination of the attribute and a value extracted from sentences such as "Alfred Nobel was born in Stockholm." Choosing the value only would lead to a loss of information, because there would be no difference between properties such as ⟨bornIn : Stockholm⟩ and ⟨diedIn : Stockholm⟩ (since in both cases we would only search for Stockholm). On the other hand, attributes such as *actedIn* can be expressed by many similar expressions (e.g., starred). In such case, using both the attribute and the value might be too restrictive. As only empirical testing can show which option is better, we tested both. In the following tables, V stands for value only (e.g., "Stockholm") and AV stands for attribute and value (e.g., "born in Stockholm").

Thirdly, there is the question what to do with synonyms. While sometimes it might help to let a metric count all synonyms of a word, as people remember concepts rather than exact words, sometimes we would prefer to look for an exact phrase. This is especially the case when the value of the property is a proper name. This means that we had the option to quote the searched term to force the used search engine to look for an exact match. Again, our pilot tested both options (i.e., quoting and no quoting).

The choices described above left us with a large number of combinations. To minimise the likelihood of type II errors, we first performed a pilot experiment. The pilot uses a different set of stimuli than the real experiment. Based on the pilot, we then selected the most promising combinations. The setup and the procedure used in the pilot experiment were similar to the actual experiment (which is described in the following sections).

Table 1: Results of the pilot study: Spearman correlation between the heuristics and knowledge of hearers.

SE + Opt	Frequency	P(n | p)	P(p | n)	PMI
AltaVista V	0.27	0.25	0.30	0.32
Bing V	0.25	0.20	0.26	0.29
Google V	**0.47**	0.14	**0.37**	**0.51**
Google AV	**0.60**	0.23	**0.50**	**0.64**

Table 1 shows the Spearman correlations between the results of the individual metrics (unquoted option) and people's judgement. The options with quoted properties proved less useful so our final evaluation used unquoted properties. The best results were achieved by using Google and expressing properties as attribute and value. Field (2009) treats values around 0.1 as indicating small effects, values around 0.3 as medium effects and values around 0.5 as large effects. This standard terminology gives our PMI and Frequency based metrics a large (positive) correlation, and our P(p | n) metric a medium (positive) correlation. To validate our results, we selected **Frequency, P(p | n) and PMI** and evaluated them on a different set of properties using the Google search engine (table 1 shows the relevant numbers from the pilot study). The results of the final evaluation can be found in the section Results and Discussion.

Participants

71 English speakers participated in our main experiment. 5 participants were discarded because they have not finished the experiment and further 5 participants were removed because the number of errors they made was more than 4 (mean + 2 * std. dev). The total number of participants was 61; 30 females, 29 males and 2 unspecified.

Materials

Ten people were selected for the experiment, each of whom was famous enough that their names occurred on the BBC Historical Figures page [1]. Based on the pilot, we attempted to select the 10 in such a way that they varied maximally (i.e., spaced evenly) in terms of how well known they are. We created sentences of the appropriate form from facts concerning these people mentioned in Wikipedia and the BBC Historical Figures page. We also added properties that did not hold true of the person in question to keep our participants more focused and to make it less likely that a participant answered *true* to each statement without using their knowledge. Only the true statements were used in the analysis. The false statements were used as a measure of participant's effort. Participants who answered *true* to more than 4 false statements were discarded. We used 7 true properties and 5 false (control) properties for each person. This resulted in total of 120 statements. To make the task shorter, the statements were ordered alphabetically and then split into 5 groups of 24 statements (14 true, 10 false, 2 or 3 properties of each person in a group). Participants were randomly assigned to judge the statements in one of the groups. Figure 1 shows the names that were chosen for the evaluation and table 2 shows a sample of the properties that were judged by the participants along with the percentage of agreement answers.

- Admiral Nelson
- Alfred Nobel
- Andy Warhol
- Duke of Wellington
- Emperor Hirohito
- Ernest Hemingway
- Florence Nightingale
- Heinrich Himmler
- Louis Pasteur
- Plato

Figure 1: Famous people used in the evaluation experiment.

Procedure

In order to find a large number of participants, the experiment was conducted online using the Amazon Mechanical Turk (MTurk). The use of MTurk can have some drawbacks, because it lets participants work from home, which makes it difficult to ensure that they are fully dedicated to the task; even worse, computer programs have occasionally been known to perform the task (instead of real people). Responses collected during the pilot experiment showed a large variability in the participants' effort, the amount of time taken to complete the experiment, and a large proportion of participants from non-English speaking countries. To mitigate some of these problems, to ensure a reasonable level of proficiency in English, and to avoid automatic responses generated by computers, participants had to successfully pass a cloze test which amounted to a very strict test of their English proficiency (Stubbs & Tucker, 1974). (Only native or highly fluent speakers tend to pass.) Furthermore, the final evaluation was advertised only to the US and UK population of the MTurk. In this way, we focussed on a particular cultural-linguistic community; the choice seemed natural

[1] http://www.bbc.co.uk/history/historic_figures/

Table 2: List of properties of Ernest Hemingway, corresponding condition and the percentage of affirmative answers. Rank AV and Rank V show how the corresponding properties ranked according to the PMI metric using Google with unquoted properties.

Property	Condition	Percentage	Rank AV	Rank V
Ernest Hemingway was a writer.	true	100.0	1	2
Ernest Hemingway was American.	true	100.0	2	1
Ernest Hemingway received the Nobel Prize in Literature.	true	63.6	4	4
Ernest Hemingway is the author of For whom the bell tolls.	true	54.5	3	3
Ernest Hemingway committed a suicide.	true	50.0	6	5
Ernest Hemingway was British.	false	27.3	-	-
Ernest Hemingway was born in Oak Park.	true	25.0	5	6
Ernest Hemingway received the Italian Silver Medal of Bravery.	true	20.0	7	7
Ernest Hemingway is the author of A tale of two cities.	false	13.3	-	-
Ernest Hemingway invented dynamite.	false	0.0	-	-
Ernest Hemingway died in a plane crash.	false	0.0	-	-
Ernest Hemingway was born in Paris.	false	0.0	-	-

given that the searched pages must have been written in English in order to contain the searched terms. The inclusion of the pre-requisites (cloze test and country restrictions) greatly improved the results (e.g., less variation in the time taken to complete the experiment and fewer number of participants who made errors).

The first page showed the instructions on how to answer and how to navigate the website and also urged the participants to rely on their own knowledge and avoid using the Internet to answer the questions. The participants were then asked to fill in some information such as sex, age group and interests. The participants then viewed one statement at a time and were asked to select one of the three provided options (true, don't know, false). The participants could also provide a comment for each statement. After finishing the experiment they were given an opportunity to provide additional open comments.

The search engine queries were performed over December 2011 and January 2012. To ensure replicability of the experiment, we saved all the queries and the corresponding numbers of hits returned by the search engines. These files are available on our website [2].

Results and Discussion

Table 2 contains a sample of statements that were shown to the participants. Condition *true* means that it was a true statement and *false* means it was a false (control) statement. As previously mentioned, only the true statements were used in the analysis. The percentages of affirmative answers were correlated with the output of the metrics using Spearman correlation. All calculations were performed using the R statistical package (R Development Core Team, 2010).

Table 3 shows the final results of our experiment. We used the Google search engine and tested expressing properties as attribute and value (condition AV) and as value only (condition V). The properties were unquoted in both cases.

Our results show a large positive correlation between the PMI and the Frequency based metrics, and the knowledge of people and a medium positive correlation achieved by the P(p | n). This suggests that a heuristic for common ground that employs either the Frequency or the PMI metric, to large extent, agrees with the knowledge of general public.

Table 3: Spearman correlation between the heuristics and the knowledge of hearers. All correlations were significant at $p < 0.001$.

Option	Frequency	P(p | n)	PMI
Google AV	0.639	0.437	0.664
Google V	0.632	0.475	0.662

The presented heuristic seems to work relatively well with the kinds of facts that appear in natural language generation systems. A natural question is whether the heuristic can give good results for facts that are so widely known that they are not explicitly stated. There are two kinds of these facts.

The first kind are facts such as "a person has a stomach." While it seems improbable that such a fact would be explicitly mentioned in a corpus, there are ways of implying it by statements such as "a person can get a stomach flu" or "a person can increase the risk of getting stomach cancer..." These statements include the words "person" and "stomach" and the heuristic can pick up these words without doing any semantic analysis. We tested the heuristic with a few of such statements and it seems to place them in to the well known part of common ground (i.e., assigns high scores to such facts). A proper evaluation would be needed to confirm this trend.

The second kind of facts are facts such as "Einstein had a stomach." This kind of facts requires inference, e.g. Einstein is a person and people have a stomach therefore Einstein had a stomach. As our heuristic works on the surface level of the text, it will not produce the expected results for such facts.

Conclusion and further work

We set out to find a computation estimation of common ground, starting with mutual knowledge (in the sense of Vanderschraaf and Sillari (2009)). We hypothesised that standard co-occurrence measures could be used as an approximation of a solution to the problem and tested several of

[2] http://www.abdn.ac.uk/~r04rk9/cge.zip

these measures against the knowledge of people in a particular community (cf. Clark and Marshall (1981)), as acquired in a new experiment with human participants. We consider these results to be highly encouraging. They suggest that the proposed heuristic (based on either Frequency or PMI, combined with Google search) are on the right track, at least in terms of estimating how widely known an atomic fact is (i.e., mutual knowledge); in section Estimating Common Ground we argued that this also makes it plausible that these heuristics could offer a reasonable approximation of *common* knowledge (i.e., the facts of which everyone in the community knows that everyone in the community knows them), but this was not directly investigated.

The community investigated in our experiment was comprised of native speakers of English in the UK and the USA with access to a computer. We believe it would be interesting to test to what degree different communities be modelled by different knowledge sources.

Our current work focuses on combining the metrics investigated here with other heuristics (including a discriminatory power heuristic for assessing the usefulness of a fact) to improve content selection algorithms for Natural Language Generation (e.g., Reiter and Dale (2000)).

Acknowledgements

We would like to thank the members of the Natural Language Generation group of the University of Aberdeen and the anonymous reviewers for their valuable comments. This research is sponsored by the Scottish Informatics and Computer Science Alliance (SICSA).

References

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K. Spence & J. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). Academic Press, New York.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.*, *36*(4), 673–721.

Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.

Clark, H. H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). New York: Cambridge University Press.

Fano, R. M. (1961). *Transmission of information: A statistical theory of communications*. New York: Wiley.

Field, A. (2009). *Discovering statistics using spss*. SAGE publications Ltd.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *In studies in linguistic analysis.* Blackwell.

Frege, G. (1892 (1952)). On sense and reference. In P. T. Geach & M. Black (Eds.), *Translations from the Philosophical Writings of Gottlob Frege.* Oxford: Basil Blackwell.

Fussell, S. R., & Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology*, *21*(5), 445–454.

Hodges, J., Yie, S., Reighart, R., & Boggess, L. (1996). An automated system that assists in the generation of document indexes. *Natural Language Engineering*, *2*(02), 137-160.

Jucks, R., Becker, B.-M., & Bromme, R. (2008). Lexical entrainment in written discourse: Is experts' word use adapted to the addressee? *Discourse Processes*, *45*(6), 497-518.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In *Probabilistic linguistics* (pp. 39–96). MIT Press.

Kilgarriff, A. (2007, March). Googleology is bad science. *Comput. Linguist.*, *33*, 147–151.

Krahmer, E., & van Deemter, K. (2012, March). Computational Generation of Referring Expressions: A Survey. *Computational Linguistics*, *38*(1), 173-218.

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, *9*(1), 2–24.

Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, Massachusetts: Harvard University Press.

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, *64*(3), 245 - 259.

Pedersen, T. (2008, September). Empiricism is not a matter of faith. *Comput. Linguist.*, *34*, 465–470.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. New York, NY, USA: Cambridge University Press.

Riordan, B., & Jones, M. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345.

Strawson, P. (1952). Introduction to logical theory.

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of english proficiency. *The Modern Language Journal*, *58*(5/6), pp. 239-241.

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the twelfth european conference on machine learning (ecml-2001).*

Vanderschraaf, P., & Sillari, G. (2009). Common knowledge. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2009 ed.). http://plato.stanford.edu/archives/spr2009/entries/common-knowledge/.

Van Eijck, J. (1993). The dynamics of description. *Journal of Semantics*, *10*(3), 239-267.