Human Variation and Lexical Choice

Ehud Reiter*
University of Aberdeen

Somayajulu Sripada†
University of Aberdeen

Much NLP research implicitly assumes that word meanings are fixed in a language community, but in fact there is good evidence that different people probably associate slightly different meanings with words. We summarise some evidence for this claim from the literature and from an ongoing research project, and discuss its implications for Natural Language Generation, especially for lexical choice, that is choosing appropriate words for a generated text.

1 Introduction

A major task in Natural Language Generation (NLG) is lexical choice, that is choosing lexemes (words) to communicate to the reader the information selected by the system's content-determination module. From a semantic perspective lexical-choice algorithms are based on models of word meanings, which state when a word can and cannot be used; of course lexical-choice algorithms may also consider syntactic constraints and pragmatic features when choosing words.

Such models assume that it is possible to specify what a word means to a user. However, both the cognitive science literature and recent experiments carried out in the SumTime project at Aberdeen suggest that this may be difficult to do because of variations between people, that is because the same word may mean different things to different people. More precisely, while people may agree at a rough level about what a word means, they may disagree about its precise definition, and in particular what objects or events a word can be applied to. This means that it may be impossible even in principle to specify precise word meanings for texts with multiple readers, and indeed for texts with a single reader unless the system has access to an extremely detailed user model. A corpus study in our project also showed that there were differences in which words individuals used (in the sense that some words were only used by a subset of the authors), and also in how words were orthographically realised (spelled).

This suggests that it may be risky for NLG systems (and indeed human authors) to depend for communicative success on the human reader interpreting words exactly as the system intends. This in turn suggests that perhaps NLG systems should be cautious in using very detailed lexical models and also that it may be useful to add some redundancy to texts in case the reader does not interpret a word as expected. This is especially true in applications where each user only reads one generated text; if users read many generated texts, then perhaps over time they will learn about and adapt to the NLG system's lexical usage. Human variability also needs to be taken into account by NLP researchers performing corpora analyses; such analyses should not assume that everyone uses identical rules when making linguistic decisions.

* Dept of Computing Science, Aberdeen AB24 3UE, UK. Email: ereiter@csd.abdn.ac.uk
† Dept of Computing Science, Aberdeen AB24 3UE, UK. Email: ssripada@csd.abdn.ac.uk

2 Evidence for Human Lexical Variation

2.1 Previous Research

Linguists have acknowledged that people may associate different meanings with words. Nunberg (1978, p81), for example, writes

> There is considerable variation among speakers in beliefs about what does and does not constitute a member of the category ... Take *jazz*. I may believe that the category includes ragtime, but not blues; you may believe the exact opposite. After all, we will have been exposed to a very different set of exemplars. And absent a commonly accepted authority, we must construct our own theories of categories, most probably in the light of varying degrees of musical sophistication.

Many modern theories of mental categorisation (Rosch, 1978; Smith and Medin, 1981) assume that mental categories are represented by prototypes or exemplars. Therefore, if different people are exposed to different category prototypes and exemplars, they are likely to have different rules for evaluating category membership.

Parikh (1994) makes a similar point, and backs it up with some simple experimentation. For example, he showed squares from the Munsell chart to subjects and asked them to characterise them as *red* or *blue*; different individuals characterised the squares in different ways. In another experiment he showed that differences remained even if subjects were allowed to associate fuzzy-logic type truth values to statements.

In the psychological community, Malt et al. (1999) investigated what names subjects gave to real-world objects. For example, they wished to know whether subjects would describe a pump-top hand-lotion dispenser as a *bottle* or a *container*. They were primarily interested in variations across linguistic communities, but they also discovered that even within a linguistic community there were differences in how subjects named objects. They state (p242) that only 2 of the 60 objects in their study were given the same name by all of their 76 native-English-speaker subjects.

In the lexicographic community, fieldworkers for the Dictionary of American Regional English (DARE) (Cassidy and Hall, 1996) asked a representative set of Americans to respond to 'fill-in-the-blank' questionnaires. This revealed substantial differences. For example,[1] there were 228 different responses to question B12, *When the wind begins to increase, you say its* ____, the most common of which were *getting windy* and *blowing up*; and 201 different responses to question B13, *When the wind begins to decrease, you say its* ____, the most common of which were *calming down* and *dying down*.

2.2 SumTime project

The SumTime project at the University of Aberdeen is researching techniques for generating summaries of time-series data.[2] Much of the project focuses on content determination (see, for example, (Sripada et al., 2001)), but it is also examining lexical choice algorithms for time-series summaries, which is where the work described here originated. To date, SumTime has primarily focused on two domains, weather forecasts and summaries of gas-turbine sensors, although we have recently started work in a third domain as well, summaries of sensor readings in neonatal intensive care units.

2.2.1 Gas Turbine domain In order to develop a lexicon for describing patterns in gas-turbine sensor data, we asked two experts to write short descriptions of 38 signal

1 Our thanks to Joan Hall, DARE editor, for providing us with data from the DARE fieldwork.
2 See `http://www.csd.abdn.ac.uk/research/sumtime` for general information about SumTime.

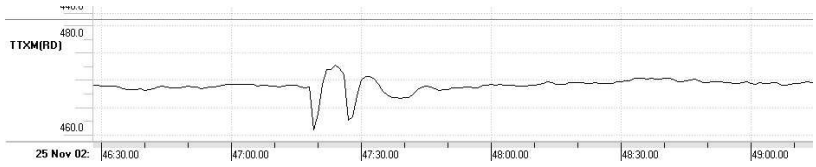

Figure 1
Signal fragment (gas-turbine exhaust temperature): Is this an *oscillation*?

day	hour	wind dir	wind speed
4-10-00	0	WSW	22
4-10-00	3	WSW	20
4-10-00	6	SW	16
4-10-00	9	SW	14
4-10-00	12	SSW	12
4-10-00	15	SSW	18
4-10-00	18	SSW	22
4-10-00	21	SSW	24
5-10-00	0	SW	26

Figure 2
Wind (at 10m) extract from 2-Oct-2000 data file (output of numerical weather model)

FORECAST 00-24 GMT, WEDNESDAY, 04-Oct 2000
WIND(10M): WSW 20-24 BACKING SSW 10-14 BY MIDDAY THEN
VEERING SW 24-28 BY EVENING

Figure 3
Wind (at 10m) extract from 5 day weather forecast issued on 2-Oct-2000

fragments. The descriptors were small, with an average size of 8.3 words. In no case did the experts produce exactly the same descriptor for a fragment. Many of the differences simply reflected usage of different words to express the same underlying concept; for example one expert typically used *rise* to describe what the other expert called *increase*. In other cases the differences reflected different levels of detail. For example, one fragment was described by expert A as *Generally steady with a slow rising trend. Lots of noise and a few small noise steps*, while expert B used the shorter phrase *Rising trend, with noise*. Both experts also had personal vocabulary; for example, the terms *bathtub* and *dome* were used only by expert A, while the terms *square wave* and *transient* were used only be expert B.

Most importantly from the perspective of this paper, there were cases where the differences between the experts reflected a difference in the meanings associated with words. For example, both experts used the word *oscillation*. 6 signals were described by both experts as oscillations, but two signals, including the one shown in Figure 1, were only described as oscillations by expert B. We do not have enough examples to solidly support hypotheses about why the experts agreed on some signals and disagreed on others, but one explanation that seems to fit the available data is that the experts agreed on signals which were very similar to a sine wave (which presumably is the prototype (Rosch, 1978) of an oscillation), but sometimes disagreed on signals which were less similar to a sine wave, such as Figure 1.

2.2.2 Meteorology domain In this domain we accumulated and analysed a corpus of 1099 human-written weather forecasts for offshore oil rigs, together with the data files (produced by a numerical weather simulation) that the forecasters examined when writing

hour	F1	F2	F3	F4	F5	total
0		**5**	**35**	1	3	44
3			1			1
6					1	1
9						0
12		1				1
15	5		2	3		10
18	**19**	3	1	**22**	4	**49**
21	7	**5**	22	3	**6**	43
total	31	14	61	29	14	149

Figure 4
How often *by evening* was used to refer to each time, for each forecaster (mode in bold font)

the forecasts. The forecasts were written by 5 different forecasters. A short extract from a typical forecast is shown in Figure 3; this text describes predicted changes in wind speed and direction two days after the forecast is issued. An extract from the corresponding data file is shown in Figure 2;it describes the predicted wind speed and direction from the numerical weather simulation, at 3-hourly intervals.

As in the gas-turbine domain, our corpus analysis showed that individual forecasters had idiosyncratic vocabulary that only they used. For example, one forecaster used the verb *freshening* to indicate a moderate increase in wind speed from a low or moderate initial value, but no other forecaster used this verb. There were also differences in orthography. For example, some forecasters lexicalised the four basic directions as *N*, *E*, *S*, and *W*, while others used the lexicalisations *N'LY*, *E'LY*, *S'LY*, and *W'LY*.

We performed a number of semantic analyses on when different forecasters used different words; these invariably showed differences between authors. For example, we attempted to infer the meaning of time phrases such as *by evening* by searching for the first data file record that matched the corresponding wind descriptor. The forecast in Figure 3, for example, says that the wind will change to *SSW 10-14* at the time suggested by *BY MIDDAY*. In the corresponding data shown in Figure 2, the first entry with a direction of SSW and a speed in the 10-14 range is 1200; hence in this example the time phrase *by midday* is associated with the time 1200. A similar analysis suggests that in this example the time phrase *by evening* is associated with the time 0000 (on 5-10-00).

We repeated this procedure for every forecast in our corpus, and statistically analysed the results to determine how individual forecasters used time phrases. More details about the analysis procedure are given by Reiter and Sripada (2002). As reported in that paper, the forecasters seemed to agree on the meaning of some time phrases; for example, all forecasters predominantly used *by midday* to mean 1200. However, they disagreed on other terms, including *by evening*. The use of *by evening* is shown in Figure 4; in particular, while forecaster F3 (the author of the text in Figure 3) most often used this phrase to mean 0000, forecasters F1 and F4 most often used this phrase to mean 1800. The differences between forecasters in their usage of *by evening* is significant at $p < .001$ under both a chi-square test (which treats time as a categorical variable) and a One-Way ANOVA (which compares the mean time for each forecaster; for this test we recoded the hour 0 as 24).

2.2.3 Knows Java experiment Some colleagues pointed out to us that meteorology in particular was a domain with an established sublanguage and usage conventions, whose words might correspond to technical terms, and wondered what would happen in a domain where there was no established sublanguage and technical terminology. We

therefore performed a small experiment in the University of Aberdeen, where we asked 21 postgraduate students and academic staff members to fill out a questionnaire asking which of the following individuals they would regard as *knowing Java*:[3]

- A cannot program in Java, but knows that Java is a popular programming language.
- B cannot write a Java program from scratch, but can make very simple changes to an existing Java program (such as changing a string constant that specifies a URL).
- C can use a tool such as JBuilder to write a very simple Java program, but cannot use control flow constructs such as while loops.
- D can write Java programs that use while loops, arrays, and the Java class libraries, but only within one class, she cannot write a program that consists of several classes.
- E can create complex Java programs and classes, but needs to occasionally refer to documentation for details of the Java language and class libraries.

Respondents could tick *Yes*, *Unsure*, or *No*. All 21 respondents ticked *No* for A and *Yes* for E. They disagreed about whether B, C, and D could be considered to know Java; 3 ticked *Yes* for B, 5 ticked *Yes* for C, and 13 ticked *Yes* for D.

In other words, even among this relatively homogeneous group, there was considerable disagreement over what the phrase *knows Java* meant in terms of actual knowledge of the Java programming language.

3 Implications for NLG

3.1 Lexical Choice

The previous section has argued that people in many cases do associate different meanings with lexemes and phrases such as *oscillation*, *by evening*, and *knows*; and that some words, such as *bathtub* and *freshening*, are only used by a subset of authors in a domain. What impact does this have on lexical choice?

In applications where users read only one generated text, it may be necessary to restrict lexeme definitions to those that we expect all users to share. Indeed, essentially this advice was given to us by a domain expert in an earlier project on generating personalised smoking-cessation letters (Reiter, Robertson, and Osman, 2000). In applications where users read many generated texts over a period of time, however, an argument could be made for using a richer vocabulary and set of lexeme definitions, and expecting users to adapt to and learn the system's vocabulary and usage over the course of time; it may be appropriate to add extra 'redundant' information to texts (Section 3.2) while the user is still learning the system's lexical usage. This strategy has some risks, but if successful can lead to shorter and less awkward descriptions. Consistency is essential if this strategy is followed; the system should not, for example, sometimes use *by evening* to mean 1800 and sometimes use *by evening* to mean 0000.

3 The questionnaire in fact contained a sixth item: "F can create complex Java libraries and almost never needs to refer to documentation because she has memorised most of it." However, a few subjects were unsure whether *create complex Java libraries* meant programming or meant assembling compiled object files into a single archive file using a tool such as *tar* or *jartool*, so we dropped F from our study.

We are not aware of previous research in lexical choice which focuses on dealing with differences in the meanings that different human readers associate with words. Perhaps the closest research strand is tailoring word choice and phrasing according to the expertise of the user (Bateman and Paris, 1989; Reiter, 1991; McKeown, Robin, and Tanenblatt, 1993). For example, COMET (McKeown, Robin, and Tanenblatt, 1993) could generate *Check the polarity* for skilled users and *Make sure the plus on the battery lines up with the plus on the battery compartment* for unskilled users. General reviews of previous lexical choice research in NLG are given by Stede (1995) and Wanner (1996); Zukerman and Litman (2001) review research in user-modelling and NLP.

In the linguistic community, Parikh (1994) has suggested that utility theory be applied to word choice. In other words, if we know (A) the probability of a word being correctly interpreted or misinterpreted, and (B) the benefit to the user of correct interpretation and the cost of misinterpretation, then we can compute an overall utility to the user of using the word. This seems like an interesting theoretical model, but in practice (at least in the applications we have looked at) while it may be just about possible to get data on the likelihood of correct interpretation of a word, it is probably impossible to calculate the cost of misinterpretation. This is because we do not have accurate task models that specify exactly how the user will use the generated texts (and our domain experts have told us that it is probably impossible to construct such models).

3.1.1 Near Synonyms A related problem is choosing between near synonyms, that is words with similar meanings. For example, choosing between *easing* and *decreasing* when describing changes in wind speed, or *saw-tooth transient* and *shark-tooth transient* when describing gas-turbine signals. The most in-depth examination of choosing between near synonyms was done by Edmonds (1999), who essentially suggested using rules based on lexicographic work such as *Webster's New Dictionary of Synonyms* (Gove, 1984).

Edmonds was working on machine translation, not generating texts from non-linguistic data, and hence was looking at larger differences than the ones we are concerned with. Indeed, dictionaries do not in general give definitions at the level of detail required by SumTime; for example, we are not aware of any dictionary which defines *oscillation* in enough detail to specify whether it is appropriate to describe the signal in Figure 1 as an *oscillation*. However, we also have some doubts as to whether the definitions given in synonym dictionaries such as Gove (1984) do indeed accurately represent how all members of a language community use near-synonyms. For example, when describing synonyms and near-synonyms of *error*, Gove (page 298) states that *faux pax* is 'most frequently applied to a mistake in etiquette'. This seems to be a fair match to DARE's fieldwork question (see Section 2.1) JJ41, *An embarrassing mistake: Last night she made an awful* ____, and indeed DARE (vol 2, page 372) states that *faux pax* was a frequent response to this question. However, DARE adds that *faux pax* was less often used in the South Midland region of the US (the states of Kentucky and Tennessee and some adjacent regions of neighbouring states), and also was less often used by people who lacked a college education. So, while *faux pax* might be a good lexicalisation of a 'mistake in etiquette' for most Americans, it might not be appropriate for all Americans; and for example if an NLG system knew that its user was a non-college educated man from Kentucky, than perhaps it should consider using another word for this concept.

3.2 Redundancy

Another implication of human variation is that if there is a chance that people may not interpret words as expected, it may be useful for an NLG system to include extra information in the texts it generates, beyond what is needed if users could be expected to interpret words exactly as the system intended. Indeed, unexpected word interpretations

could perhaps be considered to be a type of 'semantic' noise; and as with all noise, redundancy in the signal (text) can help the reader recover the intended meaning.

For example, the referring-expression generation model of Dale and Reiter (1995) selects attributes to identify referents based on the assumption that the hearer will interpret the attributes as the system expects. For instance, assume there are two books in focus, B1 and B2, and the system's knowledge base records that B1 has colour red and B2 has colour blue. Then, according to Dale and Reiter, *the red book* is a distinguishing description which uniquely identifies B1.

However, Parikh has shown that people can in fact disagree about which objects are red and which are blue; if this is the case, then it is possible that the hearer will not in fact be able to identify B1 as the referent after hearing *the red book*. To guard against this eventuality, it might be useful to add additional information about a different attribute to the referring expression; for example, if B1 is 100 pages and B2 is 1000 pages, the system could generate *the thin red book*. This is longer than *the red book* and thus perhaps may take longer to utter and comprehend, but its redundancy provides protection against unexpected lexical interpretations.

3.3 Corpus analysis

A final point is that differences between individuals should perhaps be considered in general by people performing corpus analyses to derive rules for NLG systems (Reiter and Sripada, 2002). To take one randomly-chosen example, Hardt and Rambow (2001) suggest a set of rules for deciding on VP-ellipsis which are based on machine learning techniques applied to the Penn Treebank corpus. These achieve a 35% reduction in error rate against a baseline rule. They did not consider variation in author, and we wonder if a considerable amount of the remaining error is due to individual variation in deciding when to ellide a VP. It would be interested to perform a similar analysis with author specified as one of the features given to the machine learning algorithm, and see if this improved performance.

4 NLG vs Human-Written Texts

To finish on a more positive note, human variation may potentially be an opportunity for NLG systems, because they can guarantee consistency. In the weather forecasting domain, for example, users receive texts written by all five forecasters, which means that they may have problems reliably interpreting phrases such as *by evening*; an NLG system, in contrast, could be programmed to always use this phrase consistently. An NLG system could also be programmed to avoid idiosyncratic terms which users might not be familiar with (*bathtub*, for example), and not to use terms in cases where people disagree about its applicability (eg, *oscillation* for Figure 1). Our corpus analyses and discussions with domain experts suggest that it is not always easy for human writers to follow such consistency rules, especially of they have limited amounts of time.

Being very speculative, interaction seems to be a key aspect of the process of humans agreeing on word usage (Garrod and Anderson, 1987). Perhaps a small group of people who constantly communicate with each other over a long time period (presumably the circumstances under which language evolved) will agree on word meanings. But in the modern world its common for human writers to write documents for people they have never met or otherwise interacted with, which reduces the effectiveness of the natural interaction mechanism for agreeing on word meanings.

In summary, dealing with lexical variation among human readers is a challenge for NLG systems, and will undoubtably require a considerable amount of thought, research, and data collection. But if NLG systems could do a good job of this, they might end

up producing superior texts to many human writers, which would greatly enhance the appeal of NLG technology.

Acknowledgments

Our thanks to the many individuals who have discussed this work with us (not all of whom agree with our analysis!), including Regina Barzilay, Ann Copestake, Robert Dale, Phil Edmonds, Jim Hunter, Adam Kilgarriff, Owen Rambow, Graeme Ritchie, Rosemary Stevenson, Sandra Williams, and Jin Yu. We are also grateful to the anonymous reviewers for their helpful comments. Special thanks to Joan Hall for providing us with the DARE fieldwork data. Last but certainly not least, this work would not have been possible without the help of our industrial collaborators at Intelligent Applications and WNI/Oceanroutes. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC), under grant GR/M76681.

References

Bateman, John and Cecile Paris. 1989. Phrasing a text in terms the user can understand. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI-89)*, volume 2, pages 1511–1517.

Cassidy, Frederick and Joan Hall, editors. 1996. *Dictionary of American Regional English*. Belknap.

Dale, Robert and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19:233–263.

Edmonds, Philip. 1999. *Semantic Representations of Near-Synonyms for Automatic Lexical Choice*. Ph.D. thesis, Computer Science Dept, University of Toronto.

Garrod, Simon and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27:181–218.

Gove, Philip, editor. 1984. *Webster's New Dictionary of Synonyms*. Merriam-Webster.

Hardt, Daniel and Owen Rambow. 2001. Generation of VP-ellipsis: A corpus-based approach. In *Proceedings of the 39th Meeting of the Association for Computation Linguistics (ACL-01)*, pages 282–289.

Malt, Barbara, Steven Sloman, Silvia Gennari, Meiyi Shi, and Yuan Wang. 1999. Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40:230–262.

McKeown, Kathleen, Jacques Robin, and Michael Tanenblatt. 1993. Tailoring lexical choice to the user's vocabulary in multimedia explanation generation. In *Proceedings of ACL93, Thirty-First Annual Meeting of the Association for Computational Linguistics*, pages 226–234.

Nunberg, Geoffrey. 1978. *The Pragmatics of Reference*. University of Indiana Linguistics Club, Bloomington, Indiana.

Parikh, Rohit. 1994. Vagueness and utility: The semantics of common nouns. *Linguistics and Philosophy*, 17:521–535.

Reiter, Ehud. 1991. A new model of lexical choice for nouns. *Computational Intelligence*, 7(4):240–251.

Reiter, Ehud, Roma Robertson, and Liesl Osman. 2000. Knowledge acquisition for natural language generation. In *Proceedings of the First International Conference on Natural Language Generation*, pages 217–215.

Reiter, Ehud and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of the Second International Conference on Natural Language Generation*.

Rosch, Eleanor. 1978. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*. Lawrence Erlbaum, Hillsdale, NJ, pages 27–48.

Smith, Edward and Douglas Medin. 1981. *Categories and Concepts*. Harvard University Press, Cambridge, Mass.

Sripada, Somayajulu, Ehud Reiter, Jim Hunter, and Jin Yu. 2001. A two-stage model for content determination. In *Proceedings of ENLGW-2001*, pages 3–10.

Stede, Manfred. 1995. Lexicalization in natural language generation: a survey. *Artificial Intelligence Review*, 8:309–336.

Wanner, Leo. 1996. Lexical choice in text generation and machine translation. *Machine Translation*, 11:3–35.

Zukerman, Ingrid and Diane Litman. 2001. Natural language processing and user modeling: Synergies and limitations. *User Modeling and User-Adapted Interaction*, 11:129–158.