Lexical Choice Criteria in Language Generation

Manfred Stede
Department of Computer Science
University of Toronto
Toronto M5S 1A4, Canada
mstede@cs.toronto.edu

1 Introduction

In natural language generation (NLG), a semantic representation of some kind — possibly enriched with pragmatic attributes — is successively transformed into one or more linguistic utterances. No matter what particular architecture is chosen to organize this process, one of the crucial decisions to be made is *lexicalization*: selecting words that adequately express the content that is to be communicated and, if represented, the intentions and attitudes of the speaker. Nirenburg and Nirenburg [1988] give this example to illustrate the lexical choice problem: If we want to express the meaning "a person whose sex is male and whose age is between 13 and 15 years", then candidate realizations include: *boy, kid, teenager, youth, child, young man, schoolboy, adolescent, man.* The criteria influencing such choices remain largely in the dark, however.

As it happens, the problem of lexical choice has not been a particularly popular one in NLG. For instance, Marcus [1987] complained that most generators don't really choose words at all; McDonald [1991], amongst others, lamented that lexical choice has attracted only very little attention in the research community. Implemented generators tend to provide a one-to-one mapping from semantic units to lexical items, and their producers occasionally acknowledge this as a shortcoming (e.g., [Novak, 1991, p. 666]); thereby the task of lexical choice becomes a non-issue. For many applications, this is indeed a feasible scheme, because the sub-language under consideration can be sufficiently restricted such that a direct mapping from content to words does not present a drawback — the generator is implicitly tailored towards the type of situation (or *register*) in which it operates. But in general, with an eye on more expressive and versatile generators, this state of affairs calls for improvement.

Why is lexical choice difficult? Unlike many other decisions in generation (e.g., whether to express an attribute of an object as a relative clause or an adjective) the choice of a word very often carries implicatures that can change the overall message significantly — if in some sentence the word *boy* is replaced with one of the alternatives above, the meaning shifts considerably. Also, often there are quite a few similar lexical options available to a speaker, whereas the number of possible syntactic sentence constructions is more limited. To solve the choice problem, first of all the differences between similar words have to be represented in the lexicon, and the criteria for choosing among them have to be established. In the following, I give a tentative list of choice criteria, classify them into *constraints* and *preferences*, and outline a (partly implemented) model of lexicalization that can be incorporated into language generators.

2 Word Choice Criteria

Only few contributions have been made towards establishing word choice criteria in NLG.[1] Hovy's [1988] generator PAULINE selected lexical items according to *pragmatic* aspects of the situation (rhetorical goals of the speaker giving rise to stylistic goals, which in turn lead to certain lexical choices). Also looking at the pragmatic level, Elhadad [1991] examined the influence of a speaker's *argumentative intent* on the choice of adjectives. Wanner and Bateman [1990] viewed lexical choice from a situation-dependent perspective: the various aspects of the message to be expressed by the generator can have different degrees of salience, which may give rise to certain thematizations and also influence lexical choice. Reiter [1990] demonstrated the importance of *basic-level categories* (as used by Rosch [1978]) for generation, overriding the popular heuristic of always choosing the most specific word available.

Generally speaking, the point of "interesting" language generation (that is, more than merely mapping semantic elements one-to-one onto words) is to tailor the output to the situation at hand, where 'situation' is to be taken in the widest sense, including the regional setting, the topic of the discourse, the social relationships between discourse participants, etc. There is, however, no straightforward one-to-one mapping from linguistic features to the parameters that characterize a situation, as, for example, stylisticians point out [Crystal and Davy, 1969]. Various levels of description are needed to account for the complex relationships between the intentions of the speaker and the variety of situational parameters, which together determine the (higher-level) rhetorical means for accomplishing the speaker's goal(s) and then on lower levels their stylistic realizations.

Here we are interested in the descriptional level of *lexis*: we want to identify linguistic features that

[1] Considerable work has been done on the construction of *referring expressions*, but this is just one specific sub-problem of lexical choice, and moreover a context-sensitive one. In this paper, we restrict ourselves to choice criteria that apply independently of the linguistic context.

serve as a basis for choosing a particular lexical item from a set of synonyms. Not all these features are equally interesting, however; as Crystal and Davy [1969] noted, the relation between situational features and linguistic features is on a scale from total predictability to considerable freedom of choice. Among the less interesting dimensions are *dialect* and *genre* (sub-languages pertaining to particular domains, for example legal language or sports talk), because they tend to merely fix a subset of the vocabulary instead of allowing for variation: the fact that what Americans call a *lightning rod* is a *lightning conductor* in British English does not imply a meaningful (in particular, not a goal-directed) choice for a speaker; one rarely switches to some dialect for a particular purpose. More interesting is the degree of *semantic specificity* of lexical items. An example from Cruse [1986]: *see* is a general term for having a visual experience, but there is a wide range of more specific verbs that convey additional meaning; for instance, *watch* is used when one pays attention to a changing or a potentially changing visual stimulus, whereas *look at* implies that the stimulus is static. Such subtle semantic distinctions demand a fine-grained knowledge representation if a generator is expected to make these choices [DiMarco *et al.*, 1993].

An important factor in lexical choice are *collocational constraints* stating that certain words can co-occur whereas others cannot. For instance, we find *rancid* butter, *putrid* fish, and *addled* eggs, but no alternative combination, although the adjectives mean very much the same thing.[2] Collocations hold among lexemes, as opposed to underlying semantic concepts, and hence have to be represented as *lexical* relations. They create the problem that individual lexical choices for parts of the semantic representation may not be independent: roughly speaking, the choice of word x for concept a can enforce the choice of word y for concept b.

Finally, a highly influential, though not yet very well-understood, factor in lexical choice is *style*.

3 Lexical Style

The notion of style is most commonly associated with literary theory, but that perspective is not suitable for our purposes here. Style has also been investigated from a linguistic perspective (e.g., Sanders [1973]), and recently a computational treatment has been proposed by DiMarco and Hirst [1993]. What, then, is *style*? Like Sanders, we view it broadly as the choice between the various ways of expressing the same message. Linguists interested in style, as, for instance, Crystal and Davy [1969], have analyzed the relationships between situational parameters (in particular, different genres) and stylistic choice, and work in artificial intelligence has added the important aspect of (indirectly) linking linguistic choices to the intentions of a speaker [Hovy, 1988]. Clearly, the difficult part of the definition given above is to draw the line between *message* and *style*: what parts of an utterance are to be attributed to its invariant content, and what belongs to the chosen mode of expressing that content?

In order to approach this question for the level of lexis, hence to investigate *lexical style*, it helps to turn the question "What criteria do we employ for word choice?" around and to start by analyzing what different words the language provides to say roughly the same thing, for example with the help of thesauri. By contrastively comparing similar words, their differences can be pinned down, and appropriate features can be chosen to characterize them. A second resource besides the thesaurus are guidebooks on "how to write" (especially in foreign-language teaching), which occasionally attempt to explain differences between similar words or propose categories of words with a certain "colour" (cf. [DiMarco *et al.*, 1993]). One problem here is to determine when different suggested categories are in fact the same (e.g., what one text calls a 'vivid' word is labelled 'concrete' in another).

An investigation of lexical style should therefore look for sufficiently general features: those that can be found again and again when analyzing different sets of synonymous words. It is important to separate stylistic features from *semantic* ones, cf. the choice criterion of semantic specificity mentioned above. The whole range of phenomena that have been labelled as *associative meaning* (or as one aspect under the even more fuzzy heading *connotation*) has to be excluded from this search for features. For example, the different overtones of the largely synonymous words *smile, grin* (showing teeth), *simper* (silly, affected), *smirk* (conceit, self-satisfaction) do not qualify as recurring stylistic features. Similarly, a sentence like *Be a man, my son!* alludes to aspects of meaning that are clearly beyond the standard 'definition' of *man* (human being of male sex) but again should not be classified as stylistic. And as a final illustration, lexical style should not be put in charge to explain the anomaly in *The lady held a white lily in her delicate fist*, which from a 'purely' semantic viewpoint should be all right (with *fist* being defined as *closed hand*).

Stylistic features can be isolated by carefully comparing words within a set of synonyms, from which a generator is supposed to make a lexical choice. Once a feature has been selected, the words can be ranked on a corresponding numerical scale; the experiments so far have shown that a range from 0 to 3 is sufficient to represent the differences. Several features, however, have an 'opposite end' and a neutral position in the middle; here, the scale is $-3 \ldots 3$.

[2] In NLG, collocation knowledge has been employed by, *inter alia*, Smadja and McKeown [1991] and Iordanskaja, Kittredge and Polguère [1991].

Ranking words is best being done by constructing a "minimal" context for a paradigm of synonyms so that the semantic influence exerted by the surrounding words is as small as possible (e.g.: *They destroyed/annihilated/ruined/razed/... the building*). Words can hardly be compared with no context at all — when informants are asked to rate words on a particular scale, they typically respond with a question like "In what sentence?" immediately. If, on the other hand, the context is too specific, i.e., semantically loaded, it becomes more difficult to get access to the inherent qualities of the particular word in question.

These are the stylistic features that have been determined by investigating various guides on good writing and by analyzing a dozen synonym-sets that were compiled from thesauri:

- FORMALITY: −3...3
  This is the only stylistic dimension that linguists have thoroughly investigated and that is well-known to dictionary users. Words can be rated on a scale from 'very formal' via 'colloquial' to 'vulgar' or something similar (e.g., *motion picture–movie–flick*).
- EUPHEMISM: 0...3
  The euphemism is used in order to avoid the "real" word in certain social situations. They are frequently found when the topic is strongly connected to emotions (death, for example) or social taboos (in a *washroom*, the indicated activity is merely a secondary function of the installation).
- SLANT: −3...3
  A speaker can convey a high or low opinion on the subject by using a slanted word: a favourable or a pejorative one. Often this involves metaphor: a word is used that in fact denotes a different concept, for example when an extremely disliked person is called a *rat*. But the distinction can also be found within sets of synonyms, e.g., *gentleman* vs. *jerk*.
- ARCHAIC ...TRENDY: −3...3
  The archaic word is sometimes called 'obsolete', but it is not: old words can be exhumed on purpose to achieve specific effects, for example by calling the pharmacist *apothecary*. This stylistic dimension holds not only for content words: *albeit* is the archaic variant of *even though*. At the opposite end is the trendy word that has only recently been coined to denote some modern concept or to replace an existent word that is worn out.
- FLORIDITY: −3...3
  This is one of the dimensions suggested by Hovy [1988]. A more flowery expression for *consider* is *entertain the thought*. At the opposite end of the scale is the *trite* word. Floridity is occasionally identified with high formality, but the two should be distinguished: The flowery word is used when the speaker wants to sound impressively "bookish", whereas the formal word is "very correct". Thus, the trite *house* can be called *habitation* to add sophistication, but that would not be merely 'formal'. Another reason for keeping the two distinct is the opposite end of the scale: a non-flowery word is not the same as a slang term.
- ABSTRACTNESS: −3...3
  Writing-guidebooks often recommend to replace the abstract with the concrete word that evokes a more vivid mental image in the hearer. But what most examples found in the literature really do is to recommend *semantically more specific* words (e.g., replace *to fly* with *to float* or *to glide*), which add traits of meaning and are therefore not always interchangeable; thus the choice is not merely stylistic. A more suitable example is to characterize an *unemployed* person (abstract) as *out of work* (concrete).
- FORCE: 0...3
  Some words are more forceful, or "stronger" than others, for instance *destroy* vs. *annihilate*, or *big* vs. *monstrous*.

There is an interesting relationship (that should be investigated more thoroughly) between these features and the notion of *core vocabulary* as it is known in applied linguistics. Carter [1987] characterizes core words as having the following properties: they often have clear antonyms (*big—small*); they have a wide collocational range (*fat cheque, fat salary* but **corpulent cheque, *chubby salary*); they often serve to define other words in the same lexical set (*to beam = to smile happily, to smirk = to smile knowingly*); they do not indicate the genre of discourse to which they belong; they do not carry marked connotations or associations. This last criterion, the connotational neutrality of core words could be measured using our stylistic features, with the hypothesis being that core words tend to assume the value 0 on the scales. However, the coreness of a word is not *only* a matter of style, but also of semantic specificity: Carter notes that they are often superordinates, and this is also the reason for their role in defining similar words, which are, of course, semantically more specific. It seems that the notion of core words corresponds with *basic-level categories*, which have been employed in NLG by Reiter [1990], but which had originated not in linguistics but in cognitive psychology [Rosch, 1978].

4 Towards a Model for Lexicalization

When the input to the generator is some sort of a semantic net (and possibly additional pragmatic parameters), lexical items are sought that express all the parts of that net and that can be combined into a grammatical sentence. The hard constraint on which

(content) words can participate in the sentence is that they have the right meaning, i.e., they correctly express some aspect of the semantic specification. The second constraint is that collocations are not to be violated, to avoid the production of a phrase like *addled butter*. The other factors mentioned above enter the game as *preferences*, because their complete achievement cannot be guaranteed — if we want to speak 'formally', we can try to find particularly formal words for the concepts to be expressed; but if the dictionary does not offer any, we have to be content with more 'standard' words, at least for some of the concepts underlying the sentence. We can maximize the achievement of lexical-stylistic goals, but not strive to fully achieve them.

To arrive at this kind of elaborate lexical choice, I first employ a *lexical option finder* (following ideas by Miezitis [1988]) that scans the input semantic net and produces all the lexical items that are semantically (or truth-conditionally) appropriate for expressing parts of the net. If the set of options contains more than one item for the same sub-net, these items can differ either semantically (be more or less specific) or connotationally (have different stylistic features associated with them).

The second task is to choose from this pool a set of lexical items that together express the complete net, respect collocational constraints (if any are involved), and are maximal under a preference function that determines the degree of appropriateness of items in terms of their stylistic and other connotational features. Finally, the choice process has to be integrated with the other decisions to be made in generation (sentence scope and structure, theme control, use of conjunctions and cue words, etc.), such that syntactic constraints are respected.

Two parts of the overall system have been realized so far. First, a lexical option finder was built with LOOM, a KL-ONE dialect. Lexical items correspond to configurations of concepts and roles (not just to single concepts, as it is usually done in generation), and the option finder determines the set of all items that can cover a part of the input proposition (represented as LOOM instances). Using inheritance, the most specific as well as the appropriate more general items are retrieved (e.g., if the event in the proposition is darning a sock, the items *darn, mend, fix* are produced for expressing the action).

5 Stylistic Lexical Choice in PENMAN

At the 'front end' of the overall system, a lexical choice process based on the stylistic features listed in section 3 has been implemented using the PENMAN sentence generator [Penman-Group, 1989]. Its systemic-functional grammar has been extended with systems that determine the desired stylistic "colour" and, with the help of a distance metric (see below), determine the most appropriate lexical items that fit the target specification.

Figure 1 shows a sample run of the system, where the **:lexstyle** keyword is in charge of the variation; its filler (here, **slang** or **newspaper**) is being translated into a configuration of values for the stylistic features. This is handled by the standard mechanism in PENMAN that associates keyword-fillers with answers to inquiries posed by the grammatical systems. In the example, the keyword governs the selection from the synonym-sets for *evict, destroy,* and *building* (stored in Penman's lexicon with their stylistic features). The chosen transformation of the **:lexstyle** filler into feature values is merely a first step towards providing a link from low-level features to more abstract parameters; a thorough specification of these parameters and their correspondence with lexical features has not been done yet.

More specifically, for every stylistic dimension one system is in charge to determine its numeric target value (on the scale −3 to 3). Therefore, the particular **:lexstyle** filler translates into a set of feature/value pairs. When all the value-inquiries have been made, the subsequent system in the grammar looks up the words associated with the concept to be expressed and determines the one that best matches the desired feature/value-specification. For every word, the distance metric adds the squares of the differences between the target feature value (tf) and the value found in the lexical entry (wf) for each of the n features: $\sum_{i=1}^{n}(tf_i - wf_i)^2$

The fine-tuning of the distance-metric is subject to experimentation; in the version shown, the motivation for taking the square of the difference is to, for example, favour a word that differs in two dimensions by one point over another one that differs in one dimension by two points (they would otherwise be equivalent). The word with the lowest total difference is chosen; in case of conflict, a random choice is made.

6 Summary and Future Work

An important task in language generation is to choose the words that most adequately fit into the utterance situation and serve to express the intentions of the speaker. I have listed a number of criteria for lexical choice and then explored *stylistic* dimensions in more detail: Arguing in favour of a 'data-driven' approach, sets of synonyms have been extracted from thesauri and dictionaries; comparing them led to a proposed set of features that can discriminate synonyms on stylistic grounds. The features chosen in the implementation have been selected solely on the basis of the author's intuitions (albeit using a systematic method) — clearly, these findings have to be validated through psychological experiments (asking subjects to compare words and rate them on appropriate scales). Also, it needs to be explored in more detail whether different parts of speech should be

```
(say-spl '(rr / rst-sequence
  :domain (d / EVICT  :actor (p / PERSON :name tom)
                      :actee (t / TENANT :determiner the :number plural)
                      :tense past :lexstyle slang)
  :range (e / DESTROY :actor p
                      :actee (b / BUILDING :determiner the)
                      :tense past :lexstyle slang)))
"Tom threw the tenants out, then he pulverized the shed."

(say-spl '(rr / rst-sequence
     < same as above >
                       :tense past :lexstyle newspaper)))
"Tom evicted the tenants, then he tore the building down."
```

Figure 1: Sample run of style-enhanced PENMAN

characterized by different feature sets.

An overall model of lexicalization in the generation process has been sketched that first determines all candidate lexical items for expressing parts of a message (including all synonyms and less-specific items), and a preferential choice process is supposed to make the selections. The front-end of this system has been implemented by extending the PENMAN sentence generator so that it can choose words on the basis of a distance function that compares the feature/value pairs of lexical entries (of synonyms) with a target specification. This target specification has so far only been postulated as corresponding to various stereotypical genres, the name of which is a part of the input specification to PENMAN. In future work, the stylistic features need to be linked more systematically to rhetorical goals of the speaker and to parameters characterizing the utterance situation. One of the tasks here is to determine which features should be valid for the whole text to be generated (e.g., formality), or only for single sentences, or only for single constituents (e.g., slant).

Besides, ultimately the work on lexical style has to be integrated with efforts on syntactic style [DiMarco and Hirst, 1993]. Other criteria for lexical choice, like those mentioned in section two, have to be incorporated into the choice process. And finally, it has to be examined how lexical decisions interact with the other decisions to be made in the generation process (sentence scope and structure, theme control, use of conjunctions and cue words, etc.).

Acknowledgements

Financial support from the Natural Sciences and Engineering Research Council of Canada and the Information Technology Research Centre of Ontario is acknowledged. Part of the work reported in this paper originated during a visit to the Information Sciences Institute (ISI) at the University of Southern California; thanks to Eduard Hovy for hospitality and inspiration. For helpful comments on earlier versions of this paper, I thank Graeme Hirst and two anonymous reviewers.

References

[Carter, 1987] Ronald Carter. *Vocabulary: Applied Linguistic Perspectives.* Allen & Unwin, London, 1987.

[Cruse, 1986] D. Alan Cruse. *Lexical Semantics.* Cambridge University Press, 1986.

[Crystal and Davy, 1969] David Crystal and Derek Davy. *Investigating English Style.* Edward Arnold, London, 1969.

[DiMarco and Hirst, 1993] Chrysanne DiMarco and Graeme Hirst. A Computational Theory of Goal-Directed Style in Syntax. *Computational Linguistics*, 19(??), 1993. Forthcoming.

[DiMarco *et al.*, 1993] Chrysanne DiMarco, Graeme Hirst, and Manfred Stede. The Semantic and Stylistic Differentiation of Synonyms and Near-Synonyms. In *Working Notes of the AAAI Spring Symposium on Building Lexicons for Machine Translation.* Stanford University, 1993. Forthcoming.

[Elhadad, 1991] Michael Elhadad. Generating Adjectives to Express the Speaker's Argumentative Intent. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-91)*, pages 98–104, 1991.

[Hovy, 1988] Eduard H. Hovy. *Generating Natural Language Under Pragmatic Constraints.* Lawrence Erlbaum, Hillsdale, NJ, 1988.

[Iordanskaja *et al.*, 1991] Lidija Iordanskaja, Richard Kittredge, and Alain Polguère. Lexical Selection and Paraphrase in a Meaning-Text Generation Model. In C. L. Paris, W. R. Swartout, and W. C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, chapter 11, pages 293–312. Kluwer, Dordrecht, 1991.

[Marcus, 1987] Mitchell Marcus. Generation Systems Should Choose Their Words. In Yorick Wilks, editor, *Theoretical Issues in Natural Language Processing*, pages 211–214. New Mexico State University, Las Cruces, 1987.

[McDonald, 1991] David D. McDonald. On the Place of Words in the Generation Process. In C. L. Paris, W. R. Swartout, and W. C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 227–248. Kluwer, Dordrecht, 1991.

[Miezitis, 1988] Mara Anita Miezitis. Generating Lexical Options by Matching in a Knowledge Base. Technical Report CSRI-217, University of Toronto, 1988.

[Nirenburg and Nirenburg, 1988] Sergei Nirenburg and Irene Nirenburg. A Framework for Lexical Selection in Natural Language Generation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 471–475, Budapest, 1988.

[Novak, 1991] Hans-Joachim Novak. Integrating a Generation Component into a Natural Language Understanding System. In O. Herzog and C. R. Rollinger, editors, *Text Understanding in LILOG*, pages 659–669. Springer, Berlin/Heidelberg, 1991.

[Penman-Group, 1989] Penman-Group. The Penman Documentation. Unpublished documentation for the Penman system, 1989.

[Reiter, 1990] Ehud Reiter. Generating Descriptions that Exploit a User's Domain Knowledge. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*. Academic Press, 1990.

[Rosch, 1978] Eleanor Rosch. Principles of Categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*. Lawrence Erlbaum, Hillsdale, NJ, 1978.

[Sanders, 1973] Willy Sanders. *Linguistische Stiltheorie*. Vandenhoeck & Ruprecht, Göttingen, 1973.

[Smadja and McKeown, 1991] Frank Smadja and Kathleen R. McKeown. Using Collocations for Language Generation. *Computational Intelligence*, 7:229–239, 1991.

[Wanner and Bateman, 1990] Leo Wanner and John A. Bateman. A Collocational Based Approach to Salience-Sensitive Lexical Selection. In *Proceedings of the Fifth International Natural Language Generation Workshop*, pages 31–38, Dawson, PA, 1990.