Lexical Options in Multilingual Generation from a Knowledge Base

Manfred Stede

1 Dept. of Computer Science, University of Toronto, Canada
2 Research Center for Applied Knowledge Processing (FAW)
Helmholtzstr. 16, D-89081 Ulm, Germany

Abstract. From the viewpoint of multilingual generation, the common underlying knowledge base should be kept clear of language-specific concepts. This goal presupposes that lexical items of various languages cannot map one-to-one onto concepts all the time. We propose a more flexible way of attaching lexical items to configurations of concepts and roles, and a lexical option finder that determines the set of content words that cover pieces of the message to be expressed, thereby performing the first half of the “chunking” task (dividing the message into separately verbalizable parts). This pool of lexical options will also include synonyms and near-synonyms: items with identical denotation, that is semantic representation in the KB, but different connotational characteristics. From this set, the subsequent steps of the generation process can select the most preferred subset for expressing the message.

1 Introduction

When language is to be generated from an underlying knowledge base, lexical items need to be linked somehow to the representational units in the KB. In the following, we assume a KL-ONE style representation (LOOM [MacGregor, Bates 1987], in this case) and draw in particular on two distinctions made in this family of languages. First, on the division between concepts and relations (or roles) holding among concepts; second, on the division between terminological knowledge, which are concept and relation definitions, and assertional knowledge, which are instances of concepts and relations, representing entities of the world.

Previous language generators have typically employed a one-to-one mapping from KB units to lexical items, with their producers occasionally acknowledging this as a simplification (e.g., [Novak 1991, p. 666]). The direct association between concepts and words (or phrases) evades the problem of genuine lexical choice, which has prompted several researchers to point out the lack of work on this subject (e.g., [Marcus 1987], [Nirenburg, Nirenburg 1988], [McDonald 1991]). Choice means selecting from similar words and thus requires these to be adequately represented as synonyms or near-synonyms. While in principle one

³ For a general overview of research on lexicalization in language generation, see [Stede 1995].
could do so by creating an individual concept for each of them in the KB, this “so-

sition” would merely shift the responsibility for lexical choice to the application

program that produces the representation of the content to be expressed—after

all, it has to instantiate the ‘right’ set of concepts and relations that are given
to the generator. But many lexical choices are clearly linguistic matters, and no

application ought to bother with them (e.g., the decision to use either die or

kick the bucket), although it has to supply the general parameters like pragmatic

factors, a user model, etc., that will direct the generator in decision-making.

Moreover, we are interested in multilingual generation, and a one-to-one cor-

respondence between concepts and lexical items would require the presence of

language-specific concepts in the KB, which is not desirable from the knowledge

engineer’s perspective. Even when the target languages are closely related, we

have to deal with incongruities like this one from a bilingual automobile manual:

Disconnect the spark plug wire and . . .

Das Zündkabel abziehen und . . .

The closest translation of disconnect is trennen, and that of abziehen is pull off. Hence, the German sentence describes the nature of the physical movement and some property of the connection between the two parts, whereas the English version focuses on the effect that the action will have on the connection of the parts. Either way of using literal translations would be awkward in the given context; but again, we want to be able to generate these sentences from the same representation, not using a concept DISCONNECT for English and a separate concept ABZIEHEN for German. Later on we will examine this example more closely.

In this paper, we attempt to make progress on lexical choice in the following way. First, we argue in favor of separating denotation from connotation. We

treat the former as the conceptual part of word meaning, including selectional restrictions, that is represented in the knowledge base and in the KB–lexicon

interface, and the latter as features pertaining to emphasis, style, and the like,

that do not change the truth-conditions of the utterance, and yet contribute to

what Cruse [1986] calls evoked meaning, or what systemic-functional grammar

labels as interpersonal meaning. While the denotation amounts to the necessary

conditions for using a word, the connotational features establish a preference

ranking: a stylistic goal of, say, FORMALITY cannot always be achieved fully

(if the lexicon does not provide any formal words for the particular proposition at

hand); instead it is a matter of maximizing the fulfillment of the various stylistic

goals.

We represent the denotation of similar words by mapping them onto the same

semantic predicate, but with different thematic roles, selectional restrictions,
or distinguishing semantic traits. Therefore, we associate lexical items not to

concepts only, but to entire configurations of a concept and various roles and

4 This distinction has been made in semantic theory since medieval times, but in a

variety of ways. For a comprehensive historical overview, see [Garza-Cuarón 1991].