Driving multilingual sentence generation with lexico-grammatical resources

Raymond Kozlowski
Department of Computer and Information Sciences
University of Delaware
Newark, DE 19716, USA
kozlowsk@cis.udel.edu

Abstract

It is desirable for a generation architecture to use diverse lexical and grammatical forms of expression, especially for multilingual generation and paraphrasing. This paper presents a generation architecture capable of generating such variety in a uniform manner. Central to our approach are lexico-grammatical resources which pair elementary semantic structures with their syntactic realization and all syntactic consequences. Since the resources, contained in the lexicon, encapsulate information necessary to produce the realizations of a semantic input, the generation mechanism itself is simple and free from exceptional processing.

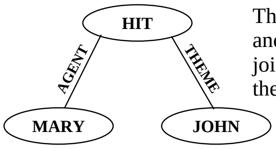
1 Introduction

Some aspects of multilingual sentence generation have caused difficulties for sentence generation architectures, particularly when realizations in different languages contain very different syntactic structures ((Dorr, 1993); (Stede, 1999); (Nicolov and Mellish, 2000); (Elhadad et al., 1997); (Bateman et al., 1991); (Lavoie et al., 2000); (Shieber and Schabes, 1990)). We note that similar challenges are encountered within a language in paraphrase generation ((Stede, 1999), (Nicolov and Mellish, 2000)).

In this paper, we present a simple sentence generation architecture that is flexible enough to handle variety both within a language and across languages in a uniform fashion. One tenet of our system is that the generation methodology should not be aware of, let alone explicitly handle, the divergent ways of expressing propositions. Our belief (like (Dorr, 1993)) is that the burden of handling language divergences should be entirely at the level of specifying the syntactic and lexical properties of a particular language. We differ significantly from Dorr, however, in that Dorr's handling of divergences amounts to encoding directives for where and how a priori assumptions about the syntax/semantics interface should be overridden. In contrast, our main concern is how the lexico-syntactic alternatives of a language should be stated so that the generation methodology can use them without exceptional processing. To this end, we have developed a set of principles for our lexico-grammatical resources, to allow variety in generation with a simple, uniform methodology.

2 Variety in surface expression

Sentence generation takes as input some semantic representation of the meaning to be conveyed in a sentence in some language. We make the assumption that the input is a hierarchical predicate/argument structure such as that shown in



The input consists of the predicate HIT and two arguments MARY and JOHN joined with the predicate HIT using the thematic roles AGENT and THEME.

Figure 1: The semantic input for *Mary hit John*

Fig. 1[1]. The output of this process should be (one of) a set of grammatical sentences whose meaning matches the original semantic input. We view the generation process as decomposing the semantic input into pieces, finding lexical resources in a language to realize the decomposed pieces, and then putting the realizations of the pieces together in a way that adheres to the syntactic principles of the language.

One of the challenges in multilingual sentence generation is that different languages seemingly place different constraints on each of these subprocesses. A number of researchers have identified the notion of cross-linguistic divergences and have noted the difficulty divergences play in machine translation (including one from an interlingua). We discuss here three divergences presented in (Dorr, 1993), selected to illustrate how divergences affect each stage of generation.

2.1 Conflational divergence

Conflation is the incorporation of some of the arguments of a predicate into the realization of the predicate itself. A conflational divergence occurs when the incorporated argument differs from one language to another, as in the English sentence (1) and its French translation (2).

(1) *Amy swam across the river.*
(2) *Amy a traversé la rivière à la nage.*
Amy crossed the river by swimming

The English verb *swim* incorporates the manner of motion, while the French verb *traverser* incorporates the path. This divergence illustrates that input decomposition cannot be done independently of the lexical choice the generator makes. This is problematic for any generator that makes a priori assumptions about input decomposition independent of particular words.

2.2 Demotional divergence

A demotional divergence occurs when a logical head is realized by a syntactic head in one language but is realized in an argument position in another, as in the English sentence (3) and its German translation (4).

(3) *Fred likes to dance.*
(4) *Fred tanzt gerne.*
Fred dances likingly

The same content of enjoying an activity is realized by the English verb *like* and the German adverb *gerne*. This divergence poses a considerable difficulty for most existing systems. The difficulty stems from the fact that the realization of a clause typically starts with the main verb which sets up a syntactic context into which other constituents are fit. It is assumed that this main verb realizes the predicate at the "top" of the input. Assuming that the enjoying is the top predicate, the English case (3) is standard. In German, on the other hand, the top predicate is realized by the adverb *gerne*, typically not seen as setting up an appropriate syntactic context into which the remaining arguments can be fit.

In handling this divergence, some existing systems use non-determinism as to which predicate syntactic processing starts with ((Stede, 1999); (Nicolov and Mellish, 2000)[2]), some use additional information (about the salient relation in (Stone and Doran, 1997) and perspective in (Elhadad et al., 1997)[3]), or exceptional processing (parameter `:DEMOTE` in (Dorr, 1993)).

2.3 Thematic divergence

A thematic divergence occurs when different languages place the argument realizations differently with respect to the head, as in the English sentence (5) and its Spanish translation (6).

(5) *I like Boston.*
(6) *Boston me gusta.*
Boston to me appeals

This divergence affects putting realizations together in that the mapping between semantic roles and syntactic positions depends on the words used. Any system that presupposes a consistent mapping from thematic roles to syntactic positions would require exceptional processing for the divergent cases. For instance, (Dorr, 1993) uses parameters `:EXT` and `:INT`.

[1] For our purposes, we keep the examples simple and include no pragmatic features. In general, the input may contain such features as long as it remains hierarchical.

[2] In (Nicolov and Mellish, 2000), the choice of a predicate imposes hierarchy on a non-hierarchical input.

[3] From personal communication, different perspectives may not be required, but then the lexical chooser must and some use exceptional processing that identifies the main verb at the same time the adverb is selected.

2.4 Paraphrases within a language

Cases similar to those of cross-linguistic divergences occur within a language in the form of paraphrases, e.g. a parallel of the demotional divergence within English occurs with excelling realized by a verb and an adverb, as in (7-8).

(7) *Barbara excels at teaching.*

(8) *Barbara teaches well.*

We contend that it is particularly difficult to justify the use of exceptional processing in the generation of paraphrases.

3 Our generation architecture

We overcome the challenges discussed in the previous section in our architecture by having each of the stages of the generation process be informed by individual lexico-grammatical resources stored in the lexicon. Generation is driven by the semantic input. The input is realized by selecting lexico-grammatical resources matching pieces of it, starting with the top predicate. The realization of a piece containing the top predicate provides the syntactic context into which the realizations of the remaining pieces of the input can be fit (the placement of these being determined by the resource). We make no assumptions about the syntactic rank or category of this realization. The key to our ability to handle the divergences in a uniform manner is that our processing is driven by our lexicon and thus we do not make any a priori assumptions about 1) how much information is realized by a lexical unit, 2) the mapping between semantic and syntactic types (and thus the syntactic rank or category of the realization of the top piece), and 3) the nature of the mapping between thematic roles and syntactic positions. Because this information is contained in each lexico-grammatical resource, generation can proceed no matter what choices are specified about these in each individual resource.

3.1 The algorithm

Our algorithm is a simple, recursive process.

1. given an unrealized input, find a lexico-grammatical resource that matches a piece containing the top predicate

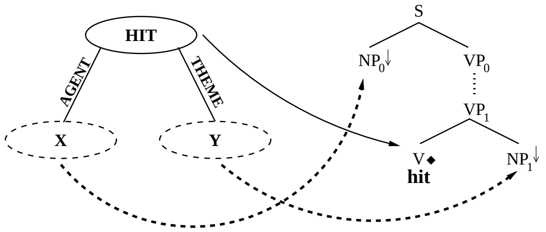


Figure 2: A resource for HIT

2. recursively realize arguments and modifiers, as determined by the resource in step 1
3. combine the realizations in step 2 with the resource in step 1, as determined by it

The details about the matching (and the satisfaction of selectional restrictions), putting resources together, and bookkeeping, not important here, can be found in (Kozlowski, 2001), (Kozlowski et al., 2002), and (Kozlowski, 2002).

3.2 Lexico-grammatical resources

The key to the simplicity of the algorithm lies in the lexico-grammatical resources, which encapsulate information necessary to carry through generation. A resource consists of three parts:

- the semantic side: the portion of semantics realized by the resource (including the predicate and any arguments; this part is matched against the input semantics)
- the syntactic side: either word(s) in a syntactic configuration or a grammatical form without words, and syntactic consequences
- a mapping between semantic and syntactic constituents indicating which constituent on the semantic side is realized by which constituent on the syntactic side

The syntactic side of the lexico-grammatical resources of a language bears some similarity to a lexicalized grammar for that language. Our resources, however, are designed to contain minimal complete semantic units with their syntactic realizations, whatever those might be.

The semantic side of the resource in Fig. 2 indicates that this resource realizes the predicate

HIT and the thematic roles AGENT and THEME. The arguments filling those roles (which must be realized separately, as indicated by dashed ovals) appear as variables X and Y to be matched against actual arguments. The syntactic side contains the verb *hit* in the active voice configuration[4]. The mapping includes a link between HIT and the anchor of the syntactic structure (*hit*), between the agent (X) and the subject, and between the theme (Y) and the complement.

3.3 Using resources in the algorithm

Step 1 of our algorithm requires matching the semantic side of a resource against the top of the input. One resource that matches the top of the input in Fig. 1 is the one in Fig. 2. In doing the matching, the arguments MARY and JOHN are unified with X and Y. The dashed ovals around X and Y indicate that this resource does not realize them. These arguments are realized recursively in step 2. In step 3, the realizations are put together with the syntactic side of the resource, as indicated by the mapping.

3.4 Principles for resources

Since all syntactic knowledge is contained in the resources, they are essential for the success of our algorithm. We have developed a number of principles that guide what they should include. The theme that runs through them is that a resource should be centered around a semantic unit and be minimal but complete.

3.4.1 Semantic motivation

In our architecture, a syntactic unit may appear in the realization only if it appears on the syntactic side of a resource whose semantic side matches a part of the input. No independent reasoning about syntax is done; it is semantics that drives generation.

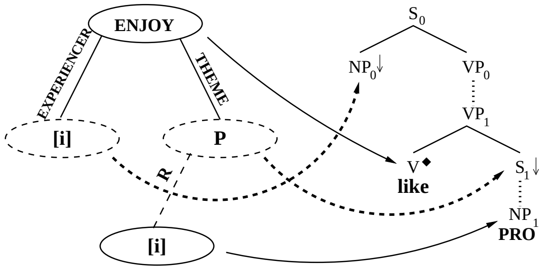


Figure 3: A resource for ENJOY

Principle 1 *A lexico-grammatical resource should contain the realization of some non-empty semantic piece.*

Principle 1 suggests that the semantically-vacuous *it* as in *it rains* should not be a separate resource and that an idiom, semantically non-decomposable, should be a single resource.

3.4.2 Syntactic consequences

Principle 2 *A lexico-grammatical resource should include all syntactic consequences of the lexico-grammatical unit.*

Examples of syntactic consequences include a semantically-vacuous *it* triggered by the verb *rain* and a subject *PRO* of the complement of the verb *like* (Fig. 3), as in (3). Principle 2 keeps the methodology modular and without additional mechanisms to keep track of syntactic requirements made by individual resources.

3.4.3 Thematic roles and arguments

We contend that some thematic roles are realized implicitly by the syntactic configuration captured by a resource.

Principle 3 *For all predicates realized by a lexico-grammatical resource, the semantic side should include as realized all thematic roles of those arguments whose realization is required in the given syntactic configuration.*

The notion of *required arguments* cannot be defined by considering semantics alone; it also depends on the syntactic configuration. For instance, for the predicate HIT realized by the verb *hit* in the active voice configuration, the agent and theme arguments are required. For HIT realized by *hit* in the *passive* voice configuration, however, only the theme is required.

[4] The syntactic side is an elementary structure of a formalism closely related to D-Tree Substitution Grammars (DSG, (Rambow et al., 2001)). The choice of formalism is not the focus of this paper - another formalism that allowed adherence to the principles discussed below would work in our architecture. To read Fig. 2, note that nodes marked with ↓ are substitution nodes corresponding to syntactic positions into which the realizations of arguments will be substituted. The dotted line indicates a domination of length zero or more where syntactic material (e.g. modifiers) may end up.

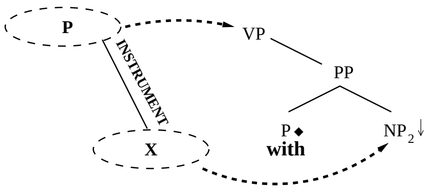


Figure 4: A resource for INSTRUMENT

Resources should capture the necessary syntax/semantics relationships in order to allow putting individual resources together.

Principle 4 *For all thematic roles realized by a lexico-grammatical resource such that the arguments filling those roles are unrealized, the semantic side should include those arguments as unrealized. The syntactic side should include the positions for all corresponding argument realizations. The mapping between the semantic and syntactic constituents should be set accordingly.*

For instance, for HIT realized by the verb *hit* in the active voice configuration, since the AGENT and THEME roles are considered realized (by Principle 3), the agent and theme *arguments* should be included in the resource as unrealized. The syntactic side should include the positions for the corresponding argument realizations, the subject and complement, with the mapping set.

3.4.4 Modifier resources

Predicative resources handle required arguments of a predicate and are complete in themselves in that they realize complete semantic subtrees when the required arguments are filled in. Other arguments can be realized by *modifier* resources which are not complete in themselves in that they never realize a complete semantic subtree, even with their own arguments filled in. They always realize a thematic role with respect to a generic predicate, itself unrealized. The issue for generation is that these resources must be combined in an appropriate fashion with the resource realizing the modified predicate.

Principle 5 *The semantic side of a modifier lexico-grammatical resource should include the uninstantiated predicate being modified and the realized thematic role. The syntactic side should include the position for the predicate realization. The mapping should be set accordingly.*

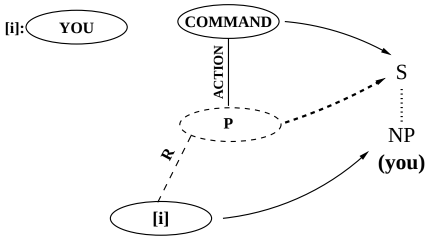


Figure 5: A resource for COMMAND

An example of a modifier resource is one for the INSTRUMENT role realized by the preposition *with*, shown in Fig. 4. On the semantic side, the realized INSTRUMENT role modifies an unrealized, uninstantiated predicate. The mapping indicates that this predicate is realized as a VP (which is where this modifier will be attached into the realization of the predicate).

3.4.5 Minimality of resources

Resources should be complete but elementary.

Principle 6 *A lexico-grammatical resource should not be decomposable into smaller resources that satisfy the other principles.*

Principle 6 yields the maximal use of compositionality in generation. While the other principles may cause a large number of resources, this one helps to keep it down. Examples where principle 6 applies include separate resources for the imperative and the wh-question forms. Because of Principle 6, it is possible for a resource not to contain words, e.g. the imperative (Fig. 5).

4 Conclusions

We have presented a uniform and flexible generation architecture. Processing is driven by the input matched against the semantic side of lexico-grammatical resources. Our resources, which follow the principles we have outlined, are the key to our methodology. They allow, for instance, a uniform treatment of cross-linguistic divergences. Incorporation is handled in a uniform manner because the matching against the semantic side of a resource determines the portion realized and the remaining semantics. The

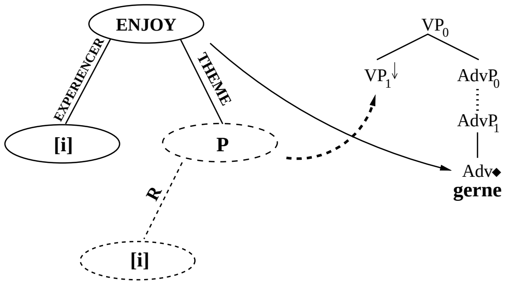


Figure 6: A resource for ENJOY in German

thematic divergence is handled in a uniform manner because, according to our principles, the mapping between semantic and syntactic constituents is defined within resources and determines where the realizations of arguments are placed with respect to that of a predicate.

No assumptions are made about the syntactic rank or category on the syntactic side, or what can set up a syntactic context. This is particularly helpful for the demotional divergence, as in (3-4). Consider the resource for the predicate ENJOY realized by the German adverb *gerne* (likingly) in Fig. 6. The uninstantiated theme appears on the semantic side, the position for its syntactic realization (the clause in which the adverb is to appear) appears on the syntactic side (the node VP$_1$), and the mapping is set accordingly. Thus, this resource sets up a syntactic context into which a clause realizing the theme argument can be fit in exactly the same way as into one set up by a main verb. Consequently, the mechanism proceeds in a uniform manner, regardless of the syntactic rank and category.

An example of the paraphrasing power of our architecture is the ability to generate (9-10) from the same input in a uniform manner, by properly specifying the resources for the wh-question and imperative forms.

(9) *Who invented calculus?*

(10) *Identify the inventor of calculus!*

We have developed a fully-operational prototype of our generation system, capable of generating, among others, all cases presented here.

References

John Bateman, Christian Matthiessen, Keizo Nanri, and Licheng Zeng. 1991. The re-use of linguistic resources across languages in multilingual generation components. In *International Joint Conference on Artificial Intelligence.*

Bonnie J. Dorr. 1993. Interlingual machine translation: a parametrized approach. *Artificial Intelligence*, 63.

Michael Elhadad, Kathleen McKeown, and Jacques Robin. 1997. Floating constraints in lexical choice. *Computational Intelligence.*

Raymond Kozlowski, Kathleen F. McCoy, and K. Vijay-Shanker. 2002. Selectional restrictions in natural language sentence generation. In *6th World Multi Conference on Systemics, Cybernetics, and Informatics (SCI'02).* To appear.

Raymond Kozlowski. 2001. Utilizing the variety of lexico-grammatical resources in uni- and multilingual sentence generation. Ph.D. dissertation proposal. Department of Computer and Information Sciences. University of Delaware.

Raymond Kozlowski. 2002. DSG/TAG - An appropriate grammatical formalism for flexible sentence generation. In *Student Research Workshop at the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02).* To appear.

Benoit Lavoie, Richard Kittredge, Tanya Korelsky, and Owen Rambow. 2000. A Framework for MT and Multilingual NLG Systems Based on Uniform Lexico-Structural Processing. In *6th Applied Natural Language Processing Conference.*

Nicolas Nicolov and Chris Mellish. 2000. PROTECTOR: Efficient Generation with Lexicalized Grammars. In *Recent Advances in Natural Language Processing.*

Owen Rambow, K. Vijay-Shanker, and David Weir. 2001. D-Tree Substitution Grammars. *Computational Linguistics*, 27(1):87–122.

Stuart M. Shieber and Yves Schabes. 1990. Synchronous Tree-Adjoining Grammars. In *Computational Linguistics.*

Manfred Stede. 1999. *Lexical semantics and knowledge representation in multilingual text generation.* Kluwer Academic Publishers, Boston.

Matthew Stone and Christine Doran. 1997. Sentence Planning as Description Using Tree Adjoining Grammar. In *Association for Computational Linguistics.*