Degand, Liesbeth (1993), Towards a Systemic Functional Grammar of Dutch for Multilingual Text Generation, In: Proceedings of the Fourth European Workshop on Natural Language Generation, Pisa, Italy, 28-30 April 1993, pp. 143 - 147.

Towards a Systemic Functional Grammar of Dutch for Multilingual Text Generation

Liesbeth Degand
UCL/Institut de Linguistique
Place B. Pascal, 1
1348 Louvain-la-Neuve, Belgium
e-mail:degand@etne.ucl.ac.be

January 29, 1993

Abstract: In this paper, we describe the results of a brief multilingual experiment in which a small text has been generated in English, German, and Dutch. Generation was on the basis of a multilingual architecture based on functional typology. The experiment has shown that this design allows the cumulative addition of other languages, in our case more particularly a Dutch language component, without revising the overall system and with much re-use of resources across languages.

1 Introduction

In this paper, we will describe the results of a brief multilingual experimen in which a small text about influenza, distributed by the European Union of Pharmacies in several European languages, has been generated in parallel in English, German and Dutch. The genre (text type) of the text is didactic, giving basic medical information to nonexpert users, the register (language style) is general, impersonal. Examples of generated fragments were the following:

English

The flu is primarily a seasonal disease. It occurs almost exclusively in autumn and winter. It is a very contagious viral infection that affects the respiratory system. You are infected with flu by direct contact with contaminated particles.

German

Die Grippe ist hauptsächlich eine saisonbedingte Krankheit. Sie tritt beinahe ausschliesslich in Herbst und Winter auf. Es ist eine sehr ansteckende Virusinfektion, die die Atemorgane befällt. Man steckt sich durch Kontakt mit ansteckenden Tröpfchen an.

Dutch

Griep is hoofdzakelijk een seizoensgebonden ziekte. Ze komt uitsluitend in de herfst en de winter voor. Het is een zeer besmettelijke virusinfektie die het ademhalingsstelsel aantast. Men krijgt griep door rechtstreeks kontakt met de besmette vochtpartikels.

Our findings about the possibility of having a common input to the generation process in three different languages are described elsewhere [Bateman *et al.*, 1993]. We will focus here on the possibility of adding a fully new linguistic component to the existing multilingual generation system, under development in collaboration between GMD/IPSI and the University of Sydney. It draws on a well-established English text generation system, the Penman system (under development at USC/ISI since 1980), which serves as the point of departure of the KOMET system for German (under development at GMD/IPSI since 1989) and the multilanguage generation project started at the

University of Sydney in 1990 for English, Japanese, and Chinese. The underlying linguistic theory of the three systems is systemic functional linguistics (see, e.g. [Halliday, 1978, Halliday, 1985, Matthiessen and Bateman, 1991]). The multilingual system has an architecture based on functional typology [Bateman *et al.*, 1991a]. In Section 2 of this paper, we will show that this design allows the cumulative addition of other languages without revising the overall system and with much re-use of resources across languages. In addition, the recently implemented word and morpheme ranks for English have also been extended to Dutch, commonalities and differences which appeared, and changes that had to be made are discussed in Section 3.

2 Adding a Dutch component to a multilingual framework

We will describe in this section how the multilingual framework has been extended with entirely new linguistic resources, in our case a Dutch grammar fragment. The set of goals of the experiment was to verify the theoretical foundations of the multilingual framework, as described in [Bateman *et al.*, 1991b]. The theoretical assumption underlying the multilingual system is that commonality across languages is functional in the first place. This functionality has to be preserved, but the structural realizations may very well differ. The functional typological foundation claims to make it possible to build a system where different languages can both converge and diverge in terms of knowledge source, semantics, grammar, etc. In this architecture, the linguistic system is composed of three strata : semantics, lexicogrammar, and phonology. Each of these levels is organized as a network of interrelated choice points (a system network). Each choice alternation is a system, a disjunction of two or more features together with the conditions under which these are available. Selections of features from this network are realized as items or structures. We will concentrate here on the lexicogrammatical stratum, more precisely we will show that the grammatical resources are shared across languages whenever this is possible, and we will show how divergences are treated. According to systemic theory, lexicogrammar is rank-based. English, German and Dutch grammar have four ranks: clause rank, group rank, word rank and morpheme rank. Units of a given rank consist of units of the rank next below in the ordering, i.e. clauses consist of groups, groups of words, and words of morphemes. Both last ranks have recently been implemented for English in the system, and will be discussed extensively in Section 3. Since the development of new components in the multilingual framework is pursued contrastively with respect to the English component (which is still the most developed one), the remainder of this paper will emphasize on differences and commonalities between Dutch and English, taking into account German elements only when necessary.

2.1 Dutch in contrast to English

The point of departure for achieving multilinguality is the English Nigel grammar [Matthiessen, 1990]. It is the shared grammatical resource for all the language components. Divergences to the English grammar are indicated by a language label, which permits language specific grammar specifications. The language specification can either concern a whole system (see Figure 1) or only a realization statement (see Figure 2). When no indication is provided, the system is valid for all languages.

In the following, we will concentrate on the divergences between English and Dutch that appeared during our experiment. Differences between both languages are reflected by diverging network traversals. In terms of systems, the changes to be made with respect to the English grammar are generally of 4 types.

One will have to:

- cut systems, i.e. the network is less delicate for the added language component. Dutch, for instance has no tagging, these systems will not be entered during the Dutch grammar traversal, but they will during the English grammar traversal. In order to achieve this, the system will be labeled with the English language specification.

```
(system
   :name (:dutch separable-verb)
   :inputs  lexical-verb-term-resolution
   :outputs ((0.5 separable-verb (insert prefix)
                                 (preselect prefix separable-prefix))
             (0.5 nonseparable-verb))
   :chooser  separable-verb-CHOOSER
   :selector CHOICE-MASTER
   :region prefixverbs
   :metafunction experiential)
```

Figure 1: Language specific system

```
(system
   :name EXPERIENTIAL-ADVERBIAL-MOOD
   :inputs EXPERIENTIAL-ADVERBIAL
   :outputs ((0.5 WH-ADVERBIAL)
             (0.5 NONWH-ADVERBIAL :dutch (inflectify CIRCUMSTANCE ADVERB)))
   :chooser EXPERIENTIAL-ADVERBIAL-MOOD-CHOOSER
   :selector CHOICE-MASTER
   :old-feature NONWH-ADVERBIAL
   :region ADVERBIAL
   :metafunction IDEATIONAL)
```

Figure 2: Language specific realization statements

```
(system
  :name (:dutch noun-gender)
  :inputs nominal-term-resolution
  :outputs (0.5 common-gender)
           (0.5 neuter-gender)
  :chooser noun-gender-chooser)
```

Figure 3: The Dutch noun-gender system

```
(gate
   :name POSITIVE-PRONOMINAL-TOTAL-NONPERSON
   :inputs (AND POSITIVE-PRONOMINAL-TOTAL NONPERSON)
   :outputs ((1.0 POSITIVE-PRONOMINAL-TOTAL-NONPERSON
              (:english lexify thing everything)
              (:dutch lexify thing alles)))
   :selector CHOICE-MASTER
   :old-feature POSITIVE-PRONOMINAL-TOTAL-NONPERSON
   :region PRONOUN
   :metafunction INTERPERSONAL)
```

Figure 4: Different lexicalization of closed class lexemes

- add systems, i.e. be more delicate. Dutch distinguishes between common and neuter grammatical gender for nouns, while English doesn't have any noun gender distinctions. This results in the addition of a noun-gender system, labeled with :dutch (see Figure 3).
- change realization statements, i.e. the commonality is functional, but the structural realization diverges, in this case the lexicalization of a closed class word is different for Dutch and English (see Figure 4).
- change entry conditions, sometimes languages seem to make the same paradigmatic distinctions, but they differ in that the entry conditions, i.e. the "paradigmatic context", to the systems are different.

Very often, divergences between languages will be reflected by a combination of these four possible changes.

As a general conclusion to the contrastive approach drawn from our experiences with Dutch, we can say that the principles announced in [Bateman *et al.*, 1991a] have been verified during the experiment, namely:

1. Within any given stratum, functional commonality is more likely to be maintained paradigmatically (in terms of systemic organization) rather than syntagmatically (in terms of structural organization). The lexicalization of the closed class words is an obvious illustration of this phenomenon.
2. Within any given stratum, functional commonality is more likely to be maintained at the lowest degree of delicacy than the highest. All included languages seem to distinguish between indicative and imperative, but English, for instance, distinguishes between different imperative forms while Dutch does not.

```
(LEXICAL-ITEM
  :NAME (:ENGLISH SAIL)
  :SPELLING "sail"
  :SAMPLE-SENTENCE "Knox sails tonight"
  :FEATURES (VERB INFLECTABLE UNITARYSPELLING S-ED LEXICAL
             NOT-CASEPREPOSITIONS NOT-TOCOMP NOT-QUESTIONCOMP
             NOT-MAKECOMP MIDDLE DOVERB EVENT NOT-ADJECTIVECOMP
             NOT-SUBJECTCOMP NOT-PARTICIPLECOMP
             NOT-BAREINFINITIVECOMP OBJECTNOTPERMITTED
             NOT-THATCOMP)
  :PROPERTIES NIL)
```

Figure 5: Lexical entry for *sail*

3. Within any given stratum, functional commonality is more likely to be maintained at the highest rank of organization than the lowest. For the grammar this means that clause systems are more likely to be shared than systems of groups or words or morphemes. This will become clear in Section 3.

Moreover, the contrastive approach saves a lot of time for the development of "new" grammars. Since functionality seems indeed to be rather common across languages, new components can substantially benefit from the work and research made for already developed languages. For the addition of a typologically close language such as Dutch with respect to English and German, we estimate the time needed to arrive at a similar level of "grammar coverage" to Nigel, which is one of the broadest computational grammars for any language, to be 1 person/5 months. In addition, we can expect that the more languages are added to the framework, the more likely it is for new languages to find a solution for a specific problem in one of the developed languages, this would then of course shorten the development time. The development of new language components in the multilingual system is also very benificiary to the languages in a further stage of development. The main "corrections" are a more functional motivation for some grammatical decisions, so that this functionality can be shared.

In the next section we will turn to the changes and additions that had to be made in order to add a word and morpheme rank to the system.

3 A word and morpheme rank for a multilingual environment

Until recently neither the PENMAN system nor the KOMET system had a word and morpheme rank. PENMAN had a hard code morphology in its lexical component which was invoked from the grammar with the classify and inflectify realization statements. Figure 5 shows the lexical entry for the verb *to sail*: the feature *S-ED* indicates that *sail* becomes *sails* in the third person singular presentform, and *sailed* in the pastform.

The KOMET grammar still works with a full-form lexicon, where the units are wordforms with lexico-semantic information and specifications of part of speech and inflectional properties in the form of feature lists [Teich, 1992a]. These features are accessed by means of the realization statements classify and inflectify, respectively. Figure 6 shows the lexical entry needed to generate the third person singular presentform of the verb *scheinen* (to shine).

Word and morpheme rank were first implemented for English in the PENMAN system, they were then extended to Dutch in the multilingual environment. This extension to another language brought

```
(lexical-item
  :name scheinen-3ps
  :spelling "scheint"
  :sample-sentence "Jetzt scheint die Sonne warm"
  :features  (LEXICALVERB INFLECTABLE UNITARYSPELLING S-ED LEXICAL
              EFFECTIVE DOLEXICALVERB DISPOSAL T
              G-3P g-sg ACTIVEFORM PRESENTFORM RELEXPLICIT)
  :comments  "process")
```

Figure 6: One of the lexical entries for the verb *scheinen*

```
(System:  PRIMARY-TENSE
 Input:
        TEMPORAL
 Output:
    (0.33333334 FUTURE)
      (CLASSIFY FINITE FUTUREAUXILIARY)
      (CONFLATE TEMPO0 FINITE)
    (0.33333334 PRESENT)
      (CONFLATE FINITE TEMPO0)
      (INFLECTIFY FINITE PRESENT-FORM)
    (0.33333334 PAST)
      (CONFLATE FINITE TEMPO0)
      (INFLECTIFY FINITE PAST-FORM)
 Chooser: PRIMARY-TENSE-CHOOSER
 LastChoice: PRESENT
 Selector: CHOICE-MASTER
 Region: TENSE
 Metafunction: LOGICAL)
```

Figure 7: Primary-tense system

also some implementational modifications for English first, which were then again tested on Dutch. Both ranks are still under further development for both English and Dutch.

For the moment, the word and morpheme rank traversals are still controlled by the classify and inflectify realization statements. In the case of irregular forms the realization statement will be fullfilled by a lexical feature in the lexical entry, and it will be applied straightforwardly, exactly as it was before. In the case of regular forms, there will be no lexical feature in the lexical entry, and the realization statement will be handled on word and morpheme rank during the further traversal.

So for instance, when *present* is chosen in the primary-tense system (see Figure 7), the inflectify realization statement sets a constraint on the network traversal to access the feature list in the given lexical entry: if the feature *present-form* is found, word and morpheme rank will not be traversed, if the feature is not found, the realization statement will be interpreted as a preselection statement that selects a lower rank, for further realization of the constituent. This traversal is developed in the next section.

een	zeer	besmettelijke	virusinfektie
		Attribute	

Figure 8: Functional structure on clause rank

een	zeer	besmettelijke	virusinfektie
Deictic	**Intensifier**	**Status**	**Thing/Head**

Figure 9: Structure on group rank

3.1 Realizing words and morphemes

In order to illustrate the word and morpheme rank traversal, let us consider the realization of the nominal group *een zeer besmettelijke virusinfektie* (a very contagious viral infection) appearing in the third sentence of our Dutch text.

After clause rank traversal, we have the functional structure as displayed in Figure 8 (we restrict ourselves to the structure obtained after having traversed the transitivity network, not taking into account the results of mood and theme).

The function Attribute is realized on group rank as a nominal group, with the functional constituents Deictic, Intensifier, Status and Thing, Deictic being realized immediately on this rank by the closed class lexeme *een* (see Figure 9).

On word rank, the right word classes are selected, Thing is realized as a noun, Status as an inflecting-adjective, and Intensifier as an adverb, they are all associated with the function Stem and inflecting-adjective also with the function Ending (Figure 10).

On morpheme rank, Stem associated with adverb and with noun will be realized by a Full-morpheme of the open class. At this point, the lexical entry is accessed and the feature *stem* of the lexical entry is realized (Figure 11).

Stem of the inflecting-adjective is realized on morpheme rank as a Regular-stem-morpheme, which is also an open class morpheme corresponding to the feature *stem* in the lexical entry "besmettelijk", but in addition, the function Ending is realized by the closed class Adjective-morpheme "e", which results in "besmettelijke". The structure after the morpheme rank traversal is displayed in Figure 12.

zeer	besmettelijke	virusinfektie
Stem	**Stem Ending**	**Stem**
adverb	**inflecting-adjective**	**noun**

Figure 10: Structure on word rank

```
(lexical-item
  :name (:dutch VIRUSINFEKTIE)
  :spelling "virusinfektie"
  :stem "virusinfektie"
  :sample-sentence  "Griep is een virusinfektie"
  :features  (noun common fem countable compound s-plural)
  :properties ())
```

Figure 11: Lexical entry of the noun "virusinfektie"

zeer	besmettelijke	virusinfektie
Full-morpheme	**Regular-stem-morpheme Adjective-morpheme**	**Full-morpheme**
"zeer"	"besmettelijk"+"e"	"virusinfektie"

Figure 12: Structure on morpheme rank

3.2 Word and Morpheme rank : Why you need them

Before the implementation of both word and morpheme rank, the organizing principle of SFL, namely choice through systems, was not applied. Our current implementation has shown though that it is possible to organize both ranks along the same principles, making paradigmatic generalizations possible here too. This favours a more coherent grammar, but more significantly the advantage of having a word and morpheme rank is the realization of the elements when and where they have to be, i.e. at the right time and place in the generation process. Let us take the example of the separable prefix in German and Dutch. The prefix is introduced at clause rank. In the KOMET grammar, this constituent will be realized immediately on clause rank with a lexify realization statement (see Figure 13). This means having a separate lexical entry for each separable prefix, and loosing every syntagmatic link between the verb and its separable prefix (see Figure 14).

In the case of a grammar with word and morpheme rank, the prefix is still inserted at clause rank as a verbal constituent (Figure 15), but it is realized linguistically on morpheme rank only (Figure 16). The appropriate prefix is chosen in the lexical entry of the verb being processed (Figure 17).

The traversal principles mentioned above are the same for the different languages, but the systems at this lower levels diverge rather strongly. This confirms the third principle of multilingual development, which says that commonality is more likely to be maintained at the highest rank of organization than the lowest.

4 Conclusions and future work

In the first part of this paper, we have shown that it is in fact possible as claimed in [Bateman *et al.*, 1991b] to add a totally new grammar fragment to the multilingual generation framework without revising the overall system. This is possible because the architecture of the system is designed along functional principles, rather than on structural realizations. The sharing of grammatical resources has shown that the already existing grammars are indeed reusable, reducing considerably the time necessary to develop a new component. The emphasis on the functional commonality across languages does not mean, however, that we consider languages only in relation to one another,

```
(system
   :name SEPARABLE-PREFIX-TYPES
   :inputs SEPARABLE-PREFIX-VERB
   :outputs ((0.20 AB-PREFIX (LEXIFY PREFIX AB))
             (0.20 ZURUECK-PREFIX (LEXIFY PREFIX ZURUECK))
             (0.20 VOR-PREFIX (LEXIFY PREFIX VOR))
             (0.20 AUF-PREFIX (LEXIFY PREFIX AUF))
             (0.20 AN-PREFIX (LEXIFY PREFIX AN)))
   :chooser SEPARABLE-PREFIX-TYPES-CHOOSER
   :selector CHOICE-MASTER
   :region VERB-MORPHOLOGY
   :metafunction EXPERIENTIAL)
```

Figure 13: Lexicalization of Prefix on clause rank

```
(lexical-item
  :name ab
  :spelling "ab"
  :sample-sentence "Er hat noch immer nicht abgenommen"
  :features  (prefix))
```

Figure 14: Lexical entry for the prefix "ab"

```
(System:  (:dutch SEPARABLE-VERB)
 Input:
        LEXICAL-VERB-TERM-RESOLUTION
 Output:
    (0.5 SEPARABLE-VERB)
           (INSERT PREFIX)
           (PRESELECT PREFIX SEPARABLE-PREFIX)
    (0.5 NONSEPARABLE-VERB)
 Chooser: SEPARABLE-VERB-CHOOSER
 LastChoice: NONSEPARABLE-VERB
 Selector: CHOICE-MASTER
 Region: PREFIXVERBS
 Metafunction: EXPERIENTIAL)
```

Figure 15: Insertion of Prefix on clause rank

```
(system
  :name  (:dutch closed-morpheme-type)
  :inputs
  closed
  :outputs
  ((0.1111 SEPARABLE-PREFIX  (preselect-substance-as-property head separable))
   (0.1111 plural-voiceless-morpheme (preselect-substance HEAD
                                                  plural-voiceless-morpheme))
   (0.1111 singular-voiceless-morpheme (preselect-substance HEAD
                                                 singular-voiceless-morpheme))
   (0.1111 plural-nonvoiceless-morpheme (preselect-substance HEAD
                                                plural-nonvoiceless-morpheme))
   (0.1111 singular-nonvoiceless-morpheme (preselect-substance HEAD
                                               singular-nonvoiceless-morpheme))
   (0.1111 voiceless-morpheme (preselect-substance HEAD voiceless-morpheme))
   (0.1111 nonvoiceless-morpheme (preselect-substance HEAD
                                                    nonvoiceless-morpheme))
   (0.1111 participle-morpheme (preselect-substance HEAD participle-morpheme))
   (0.1111 present-morpheme (preselect-substance HEAD present-morpheme))
   (0.1111 en-plural-morpheme (preselect-substance HEAD en-plural-morpheme))
   (0.1111 s-plural-morpheme (preselect-substance HEAD s-plural-morpheme))
   (0.1111 adjective-morpheme (preselect-substance HEAD adjective-morpheme)))
  :chooser nil; closed-morpheme-type-chooser -- this should be preselected.
  :region morphemes
  :metafunction logical)
```

Figure 16: The realization of Separable-prefix on morpheme rank

```
(lexical-item
  :name (:dutch AANTASTEN)
  :spelling "aantasten"
  :stem "tast"
  :sample-sentence  "Roest tast het ijzer aan"
  :features  (VERB  separable prefix doverb disposal effective
              weak hebbenaux  singularform voiceless t-stem)
  :properties  ((separable "aan")))
```

Figure 17: Lexical entry of a separable verb

minimalizing the differences. The aim of the multilingual framework is not to force languages into a common pattern - each language component taken individually should be linguistically and semantically motivated. To ensure this, we aim at developing a set of "heuristics" - "how do you build a new component into the system?" and an evaluation method of the stage of development of the added language component.

In the second part of the paper, we have focussed on the recently added word and morpheme ranks. We have shown that it is possible to organize both these ranks along the same systemic principles as on clause and group rank. The addition of both ranks now completes the grammar as a coherent whole; the main advantage being the right distribution of the realizations over the generation process. On the other hand, the implementation of the word and morpheme ranks made us aware of one of the weaknesses in the Penman-style implementation of SFL, namely the account of dependency relations. For the moment, a lot of information needs to be reduplicated over the grammar because information is not always accessible when it has to be. Accounting for such dependency relations and considering possible linguistic repercussions is a current topic of research [Teich, 1992b].

References

[Bateman *et al.*, 1991a] John A. Bateman, Christian M.I.M. Matthiessen, Keizo Nanri, and Licheng Zeng. Multilingual text generation: an architecture based on functional typology. In *International Conference on Current Issues in Computational Linguistics*, Penang, Malaysia, 1991. Also available as technical report of the department of Linguistics, University of Sydney.

[Bateman *et al.*, 1991b] John A. Bateman, Christian M.I.M. Matthiessen, Keizo Nanri, and Licheng Zeng. The re-use of linguistic resources across languages in multilingual generation components. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence, Sydney, Australia*, volume 2, pages 966 – 971. Morgan Kaufmann Publishers, 1991.

[Bateman *et al.*, 1993] John A. Bateman, Liesbeth Degand, and Elke Teich. Multilingual textuality: Some experiences from multilingual text generation. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, West Germany, 1993. submitted to The Fourth European Workshop on Natural Language Generation, Pisa, Italy, 28-30 April 1993.

[Halliday, 1978] Michael A.K. Halliday. *Language as social semiotic*. Edward Arnold, London, 1978.

[Halliday, 1985] Michael A.K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1985.

[Matthiessen and Bateman, 1991] Christian M.I.M. Matthiessen and John A. Bateman. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Frances Pinter Publishers and St. Martin's Press, London and New York, 1991.

[Matthiessen, 1990] Christian M.I.M. Matthiessen. Lexicogrammatical cartography: English systems. Technical report, University of Sydney, Linguistics Department, 1990. Ongoing expanding draft.

[Teich, 1992a] Elke Teich. Komet: grammar documentation. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, West Germany, 1992.

[Teich, 1992b] Elke Teich. Towards a notion of dependency for systemic functional grammar. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, West Germany, Darmstadt, Germany, 1992. Paper presented at the Nineteenth International Systemic Congress, Sydney, Australia.