INTERNATIONAL JOURNAL OF TRANSLATION
Vol. 16, No, i, Jan-June 2004

Representing Dutch Morphology in a Machine Translation System

EDWARD A. KOOL

In order to create an efficient machine translation system for Dutch, it is necessary, first of all, to solve the following tasks:

- create an efficient morphological analyzer for Dutch, which will be able to recognize the word forms in the source text, when translating from Dutch into another language;
- create an equally efficient morphological synthesizer for generating correct Dutch word forms in the target text when translating into Dutch.

Dutch is one of the languages of the Germanic group, and, although it is mainly analytical, unlike Slavic languages, it, as our experience shows, poses a lot of challenges that make its formalized representation a task far from trivial.

The PARS family of MT systems embraces the following languages:

- Slavic: Polish, Russian, Ukrainian [1]; Bulgarian is under way;
- Germanic: English, German [2], Dutch.

As for any language in the PARS family, we begin with the set of parts of speech. For Dutch, we have introduced the following part of speech (POS) classes:

Noun
Verb
Adjective
Adverb
Preposition
Participle
Part of a composite word

Conjunction except “en/of”
Conjunction “en/of”
Article
Particle te
wat
welke
Particle niet

As we can see, some of the classes have one word only, such as ‘Particle *te’* The reason of introducing such classes are purely practical: easier homography and homonymy resolution.

Each POS has its own set of subclass features. Below are descriptions of subclass features for two POS classes: nouns and verbs.

1. NOUN FEATURES

Animate: no, yes.
Gender/Number:

Masculine
Plural
Feminine
Neuter
Pronoun “ik”
Pronoun “wij”
Pronoun “jij”
Pronoun “jullie”

Semantics:

none
Geography
Time
Quantity
Nationality
Name

There is a number of *noun declension.* Below are some examples:
Plural ending:

-s	bedelaar	bedelaars	beggar
-n	gemeente	gemeenten	municipality

-en	kracht	krachten	force
-een	genie	genieën	genius
-eren	kind	kinderen	child

Genitive ending: none, -s.

One of the most serious challenges is alterations, which are very typical of Dutch nouns as well as verbs (see below). We have systematized the Singular-Plural alterations, and here are a few examples:

Singular-Plural Alterations:

f-v	dief	dieven	thief
s-z	huis	huizen	house
ook-oke	strook	stroke	strip
aag-age	aanvraag	aanvragen	request

2. VERB FEATURES

Conjugation: Regular or Irregular

Used with auxiliary verb: hebben, zijn.

Special type:

modal verb
auxiliary verb “hebben”
auxiliary verb “zijn”
auxiliary verb “worden”
other auxiliary

Transitivity: intransitive, transitive.

Reflexive: yes/no.

Conjugation:

Infinitive ending: -en, -n, -an, -none.
Ending in the 2nd and 3rd persons, singular, Present Tense: -t, none.
Ending in the singular, Past Tense: -nd, -de, -te, -t, -d, none.
Ending in the plural, Past Tense: -ten, -den, -nden, -en, none.
Ending in Partizip II (Participle II): -t, -d, -en, -an, none.
Partizip II is formed with prefix ge-: yes/no.

Prefix:

We have compiled a list of Dutch separable and inseparable prefixes, which we consider complete or at least close to complete. Here are just a few examples:

Separable : aan, af, beeld, beet,... , steen, stijf, stil, stop...
Inseparable: be, ge...

Verb alterations are by far the greatest challenge among the Dutch morphological features. There are many dozens of alteration types for vowels and consonants, including single/double vowel alterations, depending on the verb form:

- Infinitive
- Present 1st, Singular
- Present 2nd and 3rd Singular
- Past Singular
- Past Plural
- Participle II.

We have managed to systematized them and come with the Verb Alteration Table, which is one of the most sophisticated tools underlying the Dutch morphological analyzer and synthesizer in the PARS family of MT systems,

Here is a portion of the Table:

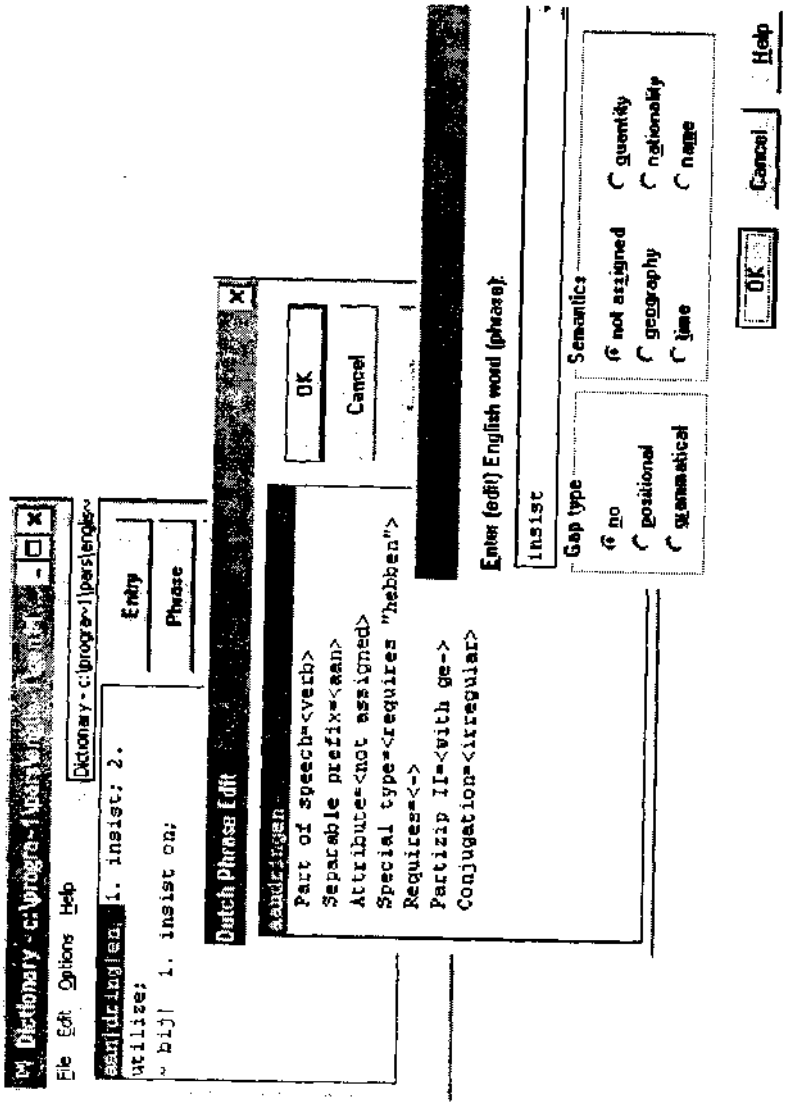
Infinitive	*Present Ist, Singular*	*Present 2nd & 3rd Sg.*	*Past Singular*	*Past Plural*	*Participle II*
A	a	a	i	i	a
a	a	aa	o	o	a
a	aa	aa	oe	oe	a
a	aa	aa	oe	oe	aa
op	oop	oop	och	och	och
ou	ou	ou	ie	ie	ou
ou	ou	ou	iel	iel	ou
t	t	t	st	st	t

Due to the alterations, the Dutch verbs have rather a sophisticated set of conjugation paradigms. Here are some of the complete verb conjugation paradigms recognized and synthesized by the PARS morphological engine:

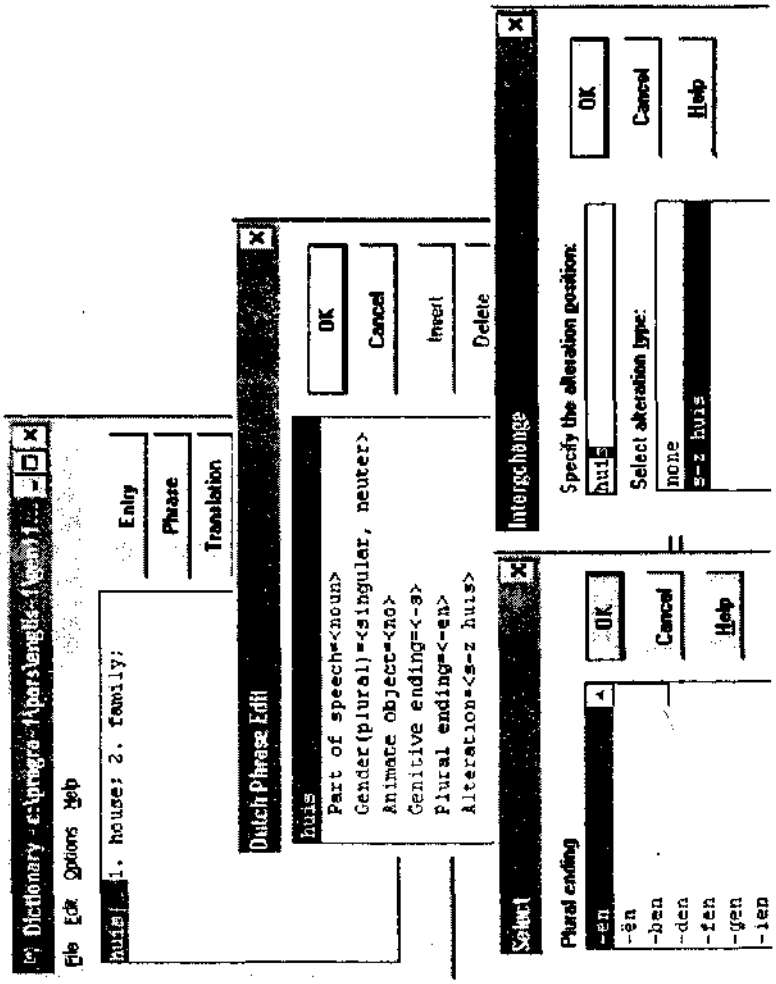
	Present 1st Singular	*Present* 2nd&3rd Singular	*Past* 1st,2nd,3rd Singular	*Past* 1st,2nd,3rd Plural	Participle I	Participle II	Auxiliary
Regular		"+t" "+de"	"+te"or "-den"	"-ten" or "ge—d"	"+end	"+ge—t" or_	
werken	werk	werk-t	werk-te	werk-ten	werk-end	ge-werk-t	hebben
fietsen	fiets	fiets-t	fiets-te	fiets-ten	fiets-end	ge-fiets-t	hebben
uiten	uit	Uit	uit-te	uit-ten	uit-end	geuit	hebben
antworden	antwoord	antwoord -t	antword -de	antwoord -den	antwoord -end	ge- antword	hebben
Irregular							
bakk-en	bak	bakt	bak-te	bakten	bakk-end	gebakken	hebben
be*denk*-en	bedenk	bedenk-t	bedach-t	bedachten	bedenkend	bedach-t	hebben
bederven	bederf	bederft	bedierf	bedierven	bedervend	bedorven	hebben; zijn
bedragen	bedraag	bedraag-t	bedroeg	bedroeg -en	bedragend	bedrag-en	hebben
bedriegen	bedrieg	bedrieg-t	bedroog	bedrog-en	bedriegend	bedrog-en	hebben: zijn
beginnen	begin	begint	begon	begonnen	beginnend	begonnen	hebben
begrijpen	begrijp	begrijpt	begreep	begrepen	begrijpend	begrepen	hebben
behouden	behoud	behoudt	behield	behielden	behoudend	behouden	hebben

PARS/H DICTIONARY EDITOR

The above features are used for tagging the Dutch words entered into the English⇔Dutch dictionaries in the PARS/H English⇔Dutch MT system. The dictionary editor is user friendly: in particular, it lets the lexicographer have a Dutch verb semi-automatically tagged. For example, after the Dutch verb *aandringen* has been entered into the dictionary, and the POS = Verb selected, the dictionary editor prompts that *aan* is a separable prefix, and the conjugation type is Regular.



When tagging Dutch nouns, the lexicographer selects the relevant values for the corresponding noun features, such as Gender, endings, and alteration type:



DICTIONARY FILES

A typical dictionary in the PARS/Dutch system consists of two halves: English-Dutch and Dutch-English.

Each half consists of 3 files:

- master file — main file comprising dictionary entries (.DIC);
- main index — (.NDI). used for translation and dictionary updating;
- alphabetic index — used for dictionary updating ('editing').NDX).

The first 1 -5 characters of the dictionary filenames are the same for all the files of this dictionary and are called 'the name of the dictionary'. Names of the files of the English-Dutch part end in _EH, while those of the Dutch-English part have _HE. *For example:*

GEN_HE.DIC	is the Dutch-English master-file of the general dictionary
GEN_EH.NDX	is the English-Dutch alphabetic index of that dictionary

A dictionary used for unidirectional translation may only consist of the corresponding half. A dictionary that is not used for updating may have no alphabetic index

The PARS/H morphological analyzer and synthesizer make use of the English⇔Dutch dictionaries based on the above-mentioned Dutch grammar features. This makes it possible for the analyzer to recognize any Dutch word-form in the source Dutch text, and synthesize the correct Dutch word-form in the target Dutch text. Besides, a special set of rules is aimed at decomposing the Dutch composite nouns.

This Dutch analysis and synthesis apparatus can and will be laid in the foundation of a family of Dutch MT systems, such as Dutch⇔German, Dutch⇔French (both under way), and others.

GRAMMAR RULES

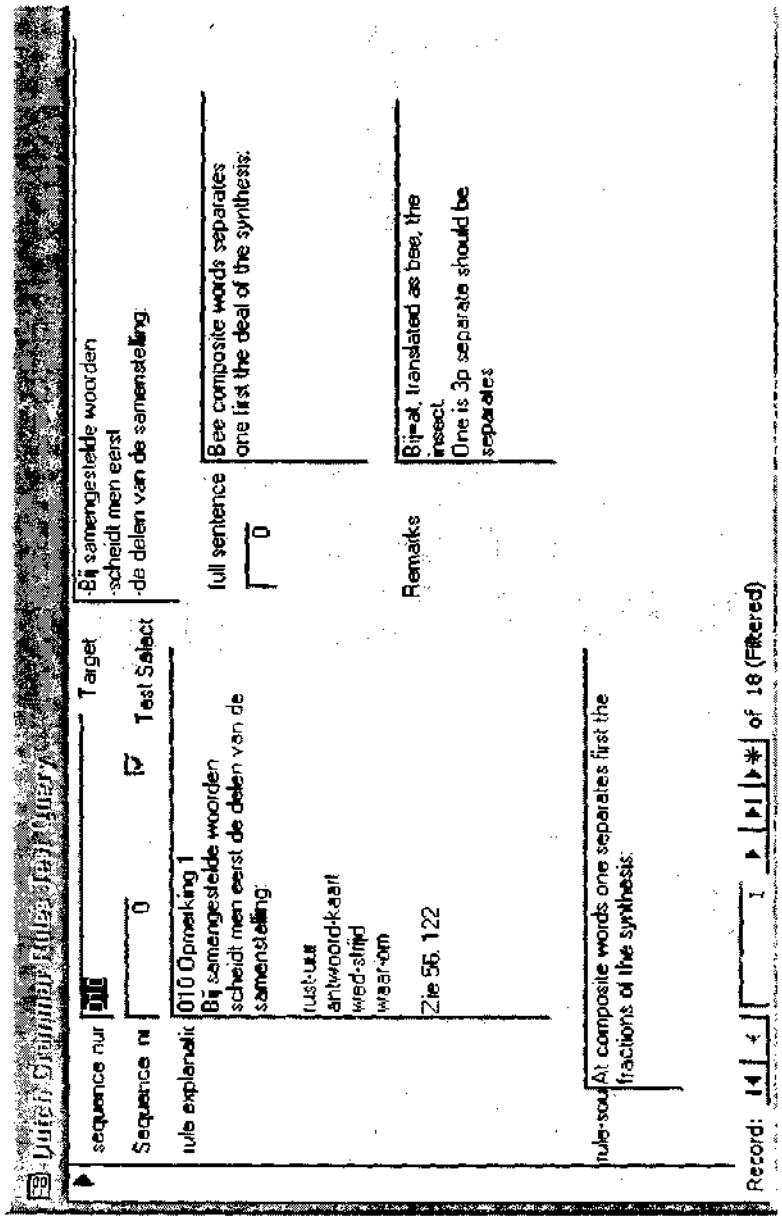
In addition to the above-mentioned morphological analyzer for translating to and from the Dutch language, two grammar books are used to specify and test the PARS English ⇔ Dutch MT program:

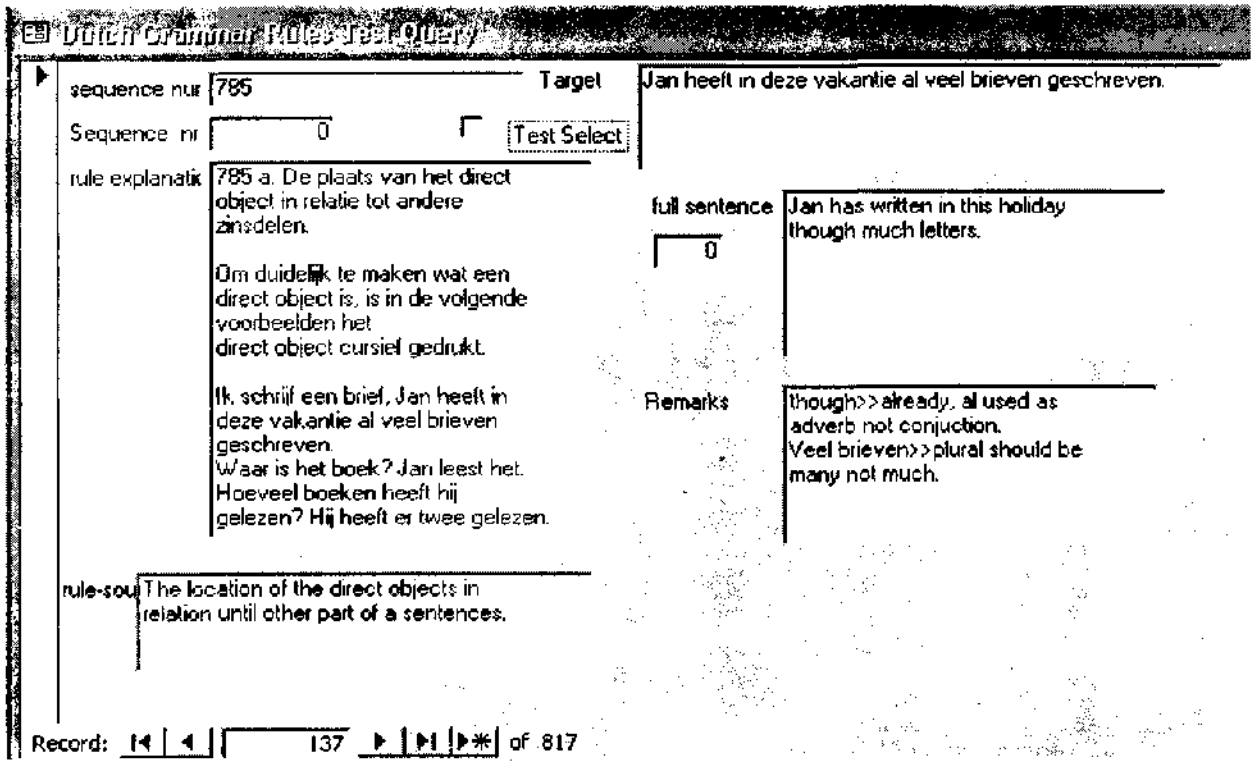
i. *Nederlandse grammatica voor anderstaligen*
(Dutch grammar for non-native speakers). ISBN 90 5517 014 3

This grammar book describes more than 800 rules. We are developing a translation program that would translate between English and Dutch according to those rules. The rules cover the issues ranging from spelling to irregular verbs.

ii. *English Grammar in Use*
Cambridge University Press. ISBN 0 521 43680 X

This grammar book describes more than 450 rules. Again, the translation program is supposed to translate between English and Dutch according to those rules. The rules pertain to subjects ranging from Present and Past to Phrasal Verbs.


Dutch Grammar Rules Test Query
sequence nur
Info
Target
Sequence nr
0
Test Select
rule explanatic
010 Opmerking 1
Bij samengestelde woorden
scheidt men eerst de delen van de
samenstelling:
rust-uur
artwoord-kaart
wied-stijd
waar-om
Zie 56. 122
Bij samengestelde woorden
scheidt men eerst
de delen van de samenstelling:
full sentence
0
Bee composite words separates
one first the deal of the synthesis:
Remarks
Bijaat, translated as bee, the
insect.
One is 3p separate should be
separates
rule-sou
At composite words one separates first the
fractions of the synthesis:
Record:
1
of 18 (Filtered)


Dutch Grammar Rules Test Query
sequence nur
785
Target
Jan heeft in deze vakantie al veel brieven geschreven.
Sequence nr
0
Test Select
rule explanatic
785 a. De plaats van het direct object in relatie tot andere zinsdelen.
Om duidelijk te maken wat een direct object is, is in de volgende voorbeelden het direct object cursief gedrukt.
Ik schrijf een brief, Jan heeft in deze vakantie al veel brieven geschreven.
Waar is het boek? Jan leest het.
Hoeveel boeken heeft hij gelezen? Hij heeft er twee gelezen.
full sentence
Jan has written in this holiday though much letters.
0
Remarks
though>>already, al used as adverb not conjuction.
Veel brieven>>plural should be many not much.
rule-sou
The location of the direct objects in relation until other part of a sentences.
Record:
137
of 817

11/13/2003
Present continuous (I am doing)
Study this example situation:
This means: she is driving now, at the time of speaking. The action is not finished.
Am/is/are -ing is the present continuous:
I am (I'm) driving
he/she/it is (= he's etc.) working
we/you/they are (= we're etc.)doing etc
I am doing something = I'm in the middle of doing something; I've started doing it and I haven't finished yet.
Often the action is happening at the time of speaking:
1. Please don't make so much noise. I'm working.
2. 'Where's Margaret?' 'She's having a bath.' (not 'she has a bath')
3. Let's go out now. It isn't raining any more. (not 'it doesn't rain')
4. Hello, Jane. Are you enjoying the party? (not 'do you enjoy')
5. I'm tired. I'm going to bed now. Goodnight!
But the action is not necessarily happening at the time of speaking.
For example:
6. Tom and Ann are talking in a café. Tom says
I'm reading an interesting book at the moment.
7. I lend it to you when I have finished it.
Record: 1 of 31

REFERENCES

1. Michael S. Blekhman. 2003. *Slavic Morphology and Machine Translation. Multilingual.* Volume 14, Issue 4, 28-31.
2. Michael Blekhman et al. 2002. A new family of the PARS translation systems. In: *Machine Translation: From Research to Real Users*, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, October 6-12, 2002, Proceedings. *Lecture Notes in Computer Science* 2499 Springer 2002, 232-236.

ACKNOWLEDGEMENTS

The author is extremely grateful to Mr. Andrei Kursin and Ms Alla Rakova of Lingvistica '98 Inc. for writing the PARS/H English⇔Dutch software. Michael Blekhman has taken part in discussing the PARS/H project.

EDWARD A. KOOL
LINGVISTICA B.V.
P.O.Box 311, 5100
AH DONGEN
THE NETHERLANDS
www.ling98.com. www.lingvistica.nl