Implementation and Evaluation of PAROLE PoS in a National Context

Tilly Dutilh and Truus Kruyt

Institute for Dutch Lexicology
P.O. Box 9515
2300 RA Leiden
The Netherlands
dutilh@inl.nl; kruyt@inl.nl

Abstract

We are annotating the complete 20 million Dutch PAROLE corpus with PoS and lemma. The morphosyntactic tagging of 250,000 words during the PAROLE project was the first confrontation of the fine-grained Dutch PAROLE tagset and its 'functional' mode of application, with real corpus data. The correction of the manual tagging and the compilation of a 100,000 words training corpus for the automatic tagger initiated the evaluation of the suitability of the tagset and the methodology of tag assignment, which topics will both be discussed in this paper. The reality of corpus data brought about a number of adaptations, linguistic restrictions and generalisations. The most salient tagger results will be presented.

Our experience is relevant for a new project: the Integrated Language Database of 8th - 21st Century Dutch (ILD), which will contain a text corpus covering all these centuries. The corpus will be annotated with lemma and PoS, in which process historical lexica will be used. Obviously, we will have to tailor tagset and methodology of tag assignment optimally to these purposes.

1. Introduction

In the nineties, a number of linguistic departments, among which the Language Database department of the Institute for Dutch Lexicology (INL), participated in a series of European standardisation projects, investigating, among other things, the national linguistic resources for their reusability. Scientific and technical specifications were set for the harmonised compilation of fourteen lexica and text corpora out of these resources, with much attention paid to feasibility.

Within this framework, a *Comparative Report on Morphosyntactic Categories in Dutch* (Dutilh, 1994) was written as a contribution to the Corpus/Lexicon Morphosyntactic Subgroup of the EAGLES Project. In the PP-PAROLE project (1994-1996), the EAGLES recommendations on morphosyntactic encoding were evaluated and subsequently presented as specifications in a common generic tagset with addition of some “non-EAGLES” values: *the PAROLE Multilingual Corpus Tagset* (Volz & Lenz, 1996; Flores, 1996). In the ensuing LE-PAROLE project (1996-1998), the *Tagset for Dutch Morphosyntactic Corpus Annotation* (Dutilh, Raaijmakers & Kruyt, 1996) was developed on the basis of this standard and thereupon effectively applied to 250,000 out of the 20 million words of the Dutch PAROLE text corpus. 50,000 tags were manually corrected for all the features of the tag (fine-grained) and 200,000 were only corrected for the first two features: part of speech and type.

As our department intends to have the Dutch PAROLE text corpus on-line for linguistic research, we are currently annotating the complete corpus for lemma and PoS, using a PAROLEX-lexicon of ca. 245,000 entries and a tagger (De Does, de & Van der Voort van der Kleij, 2002). In the process of tagging, the lexicon is used twice: for checking the output of the tagger and for lemmatising. The lexicon is our former coarse-grained DutchTale-lexicon (Van der Voort van der Kleij & Kruyt, 1997), which has been converted to the PAROLE tagset and extended with lexical entries from the Dutch PAROLE-lexicon. The tagger is a combination of statistically-based (including memory-based) taggers, which makes use of a training corpus. This corpus of present-day Dutch texts contains ca. 100,000 words, tagged according to the fine-grained PAROLE tagset. The development of the tagger and the tagging of the training corpus have initiated the evaluation of the suitability of the tagset and the methodology of tag assignment, which topics will both be discussed in this paper. It stands to reason that the evaluation will have to be pursued in the near future, as we intend to tag historical texts with PoS and lemma as well (cf. 5).

2. Methodology of tag assignment and the form function alternance

Before turning to the tagset itself, we will first discuss the methodology of tag assignment. In 1994, the following statement was made: “In practice, tagging schemes up to the present have tended to give priority to one criterion over another - i.e. giving priority to function over form, or *vice versa*. The annotation scheme for a given tagged corpus should clearly state the use of such criteria.” (EAGLES. Morphosyntactic Annotation DRAFT**,** Oct 1994, p.19.) So, apart from choosing their annotation scheme (tagset), the countries involved in the PAROLE project had to make a methodological decision about the application mode of their tagset to the corpus. The INL Language Database Department opts for a functional approach, giving priority to functional over formal criteria.

2.1. Why priority of function over form?

In the design of the Dutch PAROLE corpus, no syntactic layer explicating the function of the lexical item in the sentence had been foreseen on top of the morphosyntactic corpus tagging. Therefore, we adopted the assumption that it would be best for linguistic researchers to be able to derive as much functional information as possible from the tagging. Various other reasons have contributed to this assumption, among which linguistic reasons and reasons of feasibility.

In the case of morphologically rich languages, formal tagging effectively contributes to a certain level of syntactic information. The Dutch language, however, lost a number of formal characteristics (and, consequently, a certain amount of functional information) during its evolution. For example, instances of the subgroup of adjectives ending on 'lik' formerly changed into 'like' when they were used as adverbs. Infinitives formerly got flexion when they were used as nouns. In present-day Dutch these formal differences no longer exist, which causes systematic class ambiguity. Another systematically ambiguous group of words are participles, which are either verb or adjective (and thus adverb).

There is also a difference from a crosslinguistic perspective. Contrary to English and French, there is no formal difference in Dutch between a basic adjective used as an adjective and a basic adjective used as an adverb; cf. French: tranquil <> tranquillement, English: quiet <-> quietly with Dutch: rustig <-> rustig.

Another reason for our assumption was inspired by the PAROLE tagset itself. One of the Part of Speeches, 'Determiner', is actually based on function, being the attributively used counterpart of Pronoun.

dit boek	Determiner,demonstrative
this book	
wat is dit ?	Pronoun,demonstrative
what is this?	

As a matter of fact, the PAROLE multilingual tagset provided for many functional features. Filling these functional slots would certainly solve some of the class ambiguity problems. However, these features were not obligatory and, for reasons of feasibility, the Dutch corpus tagset left out a number of them, among which the attributive, predicative and adverbal use.

2.2. Transcategorisation: descriptive lacuna

Another solution to systematic class ambiguity is to assume that words have a default or primary lexical PoS from which they 'transcategorise' into another PoS, dependent on their function in the sentence. Transcategorisation therefore, brings the functional perspective from feature level (cf 2.1) to PoS-level. We decided to adopt this approach for writing our lexicographer’s manual and for tagging our corpus.

In practice, this was going to bring about a lot of difficulties. The crux is that grammars have never been written from the perspective of corpus tagging. Although the phenomenon of transcategorisation is mentioned (mostly cases of nominalisation), it is not treated systematically. For example, we did not find answers to the following questions:

1. Can *any* PoS turn into another PoS?
2. If not, which PoS are 'allowed' to transcategorise and which are not?
3. If so, which PoS is 'open' for other PoS’s to transcategorise into and which criteria are decisive for membership to that particular word class?

To be more specific, here follow some examples:

Is it allowed for a noun to transcategorise into an adverb, when it is used in an adverbal function?

aan het eind van de week	Noun
at the end of the week	
eind deze week	Adv
end this week	

And is it allowed for a noun to transcategorise into an adjective, when used predicatively without a determiner or article?

hij is meer mens dan vis,	Noun/Adj? Noun/Adj?
he is more human than fish	

Can a cardinal or ordinal numeral transcategorise into a noun, an adjective or a determiner?

hij is de zevende vandaag	Noun
he is the seventh today	
hij is zevende geworden	Adj
he has seventh become	
hij is zes jaar	Det?/Adj?
he is six year	
hij is nu zes	Num/Adj?
he is now six	

Incidentally, the class of numerals is a problematic one and is not always supported crosslinguistically.

If transcategorisation is allowed, which criteria are then decisive for a word to be called, for instance, a noun: the nominal function in itself (being the head of the nominal phrase) or also the fact that the PoS is preceded by an article or a determiner?

hij is kandidaat Adj/Noun?
he is candidate
hij is onze kandidaat Noun
he is our candidate

And which criterion is decisive and overrules other characteristics? For example: does an old genitive ending ‘s’ to an adjective overrule its function as a noun?

iets moois Adj/Noun
something beautiful

Historically, 'moois' is an adjective with genitive casus, but nowadays it is commonly considered to be a noun. Some grammars, however, analyse 'moois' as a postdeterminer (and therefore as an adjective). And German (which capitalises nouns) considers it a noun : etwas Schönes.

A similar question applies to adjectives: which are the criteria for a word to be called an adjective ?

hij komt als advocaat/geroepen Adj?
he comes as advocate/called
hij is iemand/iets Adj?
he is somebody/something

When the function is adjectival (predicate or complement of the subject or object), does the functional criterion overrule the nominal phrase criterion? In other words: should every PoS in that function be tagged as adjective?

2.3. Subcategorisation: descriptive lacuna

The functional approach is not restricted to top level phenomena. In PAROLE, twelve out of thirteen word classes are subcategorised. Subcategorisation is giving a type to word class members according to their function and their meaning. For example, the subdivision of nouns into common and proper; of pronouns and determiners into interrogative, relative, indefinite, etc.; of articles into definite and indefinite; and so on. Subcategorisation is more commonly accepted than transcategorisation and is treated regularly in grammars. However, criteria for subcategorial membership are not always described clearly either. For example, every Dutch grammar consulted suggested a different list of auxiliary verbs. Copula are also either longlisted or shortlisted or somewhere in between. Nor is it clear whether indefinite quantifiers are numerals or indefinite pronouns (and thus indefinite determiners if they are used attributively).

2.4. Functional approach in practice

We limited our functional approach to the commonly accepted cases of transcategorisation. These are instances of nominalisation in the first place.
A criterion for adjectives, infinitives, numerals and determiners to become a member of word class noun is that they must be the head of a nominal phrase (with or without a determiner/article).

1. adjective -> noun
wij zagen mooie en lelijke bloemen Adj
we saw beautiful and ugly flowers
wij zagen mooie bloemen en lelijke Nou
we saw beautiful flowers and ugly

2. verb(infinitive) -> noun
ze gaan de schoorsteen afbreken Verb(inf)
they will pull down the chimney
wat zij zien als het afbreken van rechten Nou
what they consider as the pull down of rights

3. numeral -> noun
ik heb er drie Num
I have () three
ik prefereer die drie van gisteren Nou
I prefer those three of yesterday
ik kies voor de derde optie Num
I choose for the third option
de derde van links werkt beter Nou
the third from left works better

4. determiner, possessive -> noun
ik zag jouw moeder Det
ik saw your mother
geef me de jouwe ! Nou
give me the your !

Nouns derived from determiners are formally distinct because of their flexion-e.
Apart from nominalisations, we opted for a few other transcategorisations:

5. adjective -> adverb
het boek is mooi Adj
the book is beautiful
de pianist speelt mooi Adv
the pianist plays beautiful

6.verb, participle -> adjective
John heeft hard gewerkt Vpart
John has hard worked
de gewerkte uren Adj
the worked hours
ik tel die uren als gewerkt Adj
I count those hours as worked

Apart from transcategorisation, a lot of functional information can be derived from subcategorial information and information from the other tag features.

3. The Dutch PAROLE Tagset and its Application: an Evaluation

3.1 Introduction

In paragraph 2.1, we explained why the functional approach was adopted. The implementation of this approach into the lexicographer's instructions as well as the confrontations with the corpus data (see below) revealed some tough, but not prohibitive, problems (2.2, 2.3) and was to finally bring about an evaluation of our method and tagset (3.1-3.3) and a number of adaptations to the tagset (3.4).

The PAROLE tagset consists of tags for 13 PoS categories such as the traditional word classes 'noun', 'verb', 'adjective' etc. and the 'new' categories determiner, infinitive marker and residual. Every tag is specified by a type such as 'common' versus 'proper' noun or 'main' versus 'auxiliary' versus 'copula' verb. Further specifications are made by means of a number of features such as 'gender', 'number', 'degree', 'function', 'case', etc. Whenever a feature or its value is not relevant for a particular language or does not apply to a token in a specific context, the slot can be left empty. This results in corpus tags such as 'Ncms - -' (noun common, masculine singular, no case, no semantic gender) or 'A q p - - - i' (adjective, qualitative, positive, no gender, no number, no case, inflected). As said before, the Dutch PAROLE corpus tagset (Dutilh, Raaijmakers & Kruyt, 1996) was established on the basis of the multilingual PAROLE corpus tagset (see www.inl.nl/corp/parole-tagset.html for an overview of the Dutch instance of the PAROLE tagset). Linguistic decisions and decisions of feasibility had been based on grammatical knowledge present in the team and had been checked in grammatical reference works (with the ANS as the most prominent). Due to the restricted time schedule of the PAROLE project, the tag set and the lexicographer's manual could not be tested on corpus data before the actual correction of the 250,000 words. As a consequence, many particular instances of language use had not been foreseen and had to be analysed ad hoc. Reference works failed us many a time (2.2. and 2.3.). At the end of the PAROLE project, we updated the lexicographer's manual with the results of the correction. However, we had a similar experience when we started working on the training corpus: new instances of sometimes onorthodox language use cropped up and had to be defined and described in the manual. As a consequence, tagging consistency in the training corpus had to be checked because of the augmented instructions. It goes without saying that this repeated experience of analysing, improvement of instructions and consistency checking involved a thorough evaluation of our tagset and tag method.

This evaluation revealed that the tagset and its application had to be customised. Another reason for this was that some relevant grammatical specifications could not be discriminated by the automatic tagger. We'll describe here the main problems encountered.

3.2. Insufficient discriminating power of taggers

As said above, reference works are not always explicit about the exact criteria to define membership of a class or subclass of words. But on top of that, many criteria, however clear, can not be easily detected by a tagger because they are not formally expressed. A tagger, for example, does not 'see' subtle usage differences mentioned in grammars to distinguish between proper and common nouns.

Honda doet 't goed op de Hollandse markt
Honda is doing well on the Dutch market
hij reed zijn Honda de stad in
he drove his Honda into town
die Honda-circulaire is heel mooi uitgevoerd
that Honda brochure is beautifully made

Grammars say that Honda in sentence 2 and 3 is a common noun. However, the only criterion for a tagger to distinguish between proper and common nouns in Dutch is capitalisation. As Honda is three times written with a capital, the output is three times proper name.

3.3. Inapplicabilaty of values and non-observed linguistic restrictions

In the reality of corpus tagging some 'theoretically sound' decisions turned out to be inapplicable and some had to be adjusted or restricted because of non-human tagging.

a. Dutch language specific gender value '*contextual*' has been deleted from the tagset. 'Contextual' means that the gender value (masculine or feminine) actually has to be decided on in the context.

de getuige zette *zijn* hoed af N,c,masculine
the witness dropped *his* hat

de getuige zette *haar* hoed af N,c,feminine
the witness dropped *her* hat

This value turned out not to be feasible for an automatic tagger, because it presupposes careful reading of the context in order to find the reference to a female or masculine person.

b. We added a linguistic restriction on the feature *'degree'* for two groups of adverbs: those who are not derived from an adjective and pronominal adverbs. We implemented the generalisation that every 'general' adverb which does not have an adjectival counterpart in the lexicon, is never gradable (with one exeption: 'vaak' (often), 'vaker' (more often), 'vaakst' (most often'). Therefore, 'true' general adverbs have an empty slot for gradability. The same restriction applies to the complete subclass of 'pronominal' adverbs. They are not gradable either.

kunnen we daarover praten?
can we thereabout talk ?

For the group of deadjectival adverbs, however, it was not feasible to investigate their gradability. Contrary to the two categories of adverbs just mentioned, the deadjectival adverbs can be gradable or not, like their corresponding adjectives. For non-gradable adjectives, degree values 'positive, comparative and superlative' are not relevant and the actual tag slot should remain empty. This should apply, for example, to 'een gouden horloge' (a golden watch) and 'de volgende keer' (the next time). However, it is a huge amount of work to examine the 17,581 adjectives in our lexicon for gradability. This should preferably be attested on very large corpora because gradability is a productive (and not always predictable) process. So the adjectives have kept their value for gradability. And logically, deadjectival adverbs kept their gradability value too.

c. We added a linguistic restriction on digits. In the lexicon, every cardinal numeral above one has 'p' (plural) as default value for number.

acht schoenen Nc-p--
eight shoes

In the practice of corpus tagging, however, many cardinal numerals are expressed in digits and, more relevantly, they do not have contextual number implications in many different situations such as dates, currencies and weight:

date:	20 mei 1949	Numc---; Numc---
currency:	fl 20,00	Numc---
weight:	6 kilo	Numc---

Therefore, we decided to consider the number value not relevant for digits. This overgeneralisation brings about some incorrect tagging:

hij is nummer acht Numc-plural-
he is number eight
Nederland telt 19.356.598 inwoners Numc---
the Netherlands have 19,356,598 inhabitants

In the first sentence number is not relevant and in the second sentence the digit is followed by a plural noun.

d. We added a linguistic restriction on feature '*gender*' in surnames. Originally, the gender slot in noun tags had to be filled without restriction. However, gender is not relevant to certain subclasses of proper nouns, such as surnames (family names).

Jan Jansen Npms-- Np-s--
John Johnson
de Clintons Np-p--
the Clintons

The family name tag has an empty slot for gender.

e. As a result of the general problem that an automatic tagger does not syntactically analyse sentences as a human tagger would do, it is really difficult to tag main verb forms for function value 'transitive' or 'intransitive' (see De Does & Van der Voort van der Kleij, 2002). To give just one problematic example: Dutch verbs can be prefixed with a preposition. However, when prefixed, intransitive verbs turn can into transitive verbs and vice versa. This would not be a problem if the prefix remains stuck to the verb form, but in Dutch the prefix is separable. As a result, a tagger cannot see whether the verb is prefixed or not and therefore cannot easily decide between transitive or intransitive use (see table 2). To give an example, compare. the infinitive 'staan' (intransitive) with the infinitive 'voorstaan' (transitive). The latter is prefixed with 'voor'.

de partij staat strenge principes voor trans
the party stands for strong principles
de partij staat morgen voor grote problemen intrans
the party stands tomorrow for (is in front of) big problems.

In the first sentence we have a direct object ('strenge principes') depending on a transitive separable verb ('staat …voor') and in the second sentence we have a normal intransitive verb ('staan') followed by a prepositional phrase ('voor grote problemen').

3.4. Tag set insufficiency: missing types and values

Initially we had not chosen values for truncated words, foreign words and enumeration characters, which in PAROLE are types of the word class ‘residual’. These types have been added after all. The same applies to the verb mood values 'conjunctive' and 'imperative' (which are quite infrequent in a written present-day corpus) and the main verb function value 'reflexive'. We added these values after all. The empty slot for these features singled the tags out from the big heap of verb tags anyway and could therefore easily be filled with the specific value.

4. Tagger Results

The functional method in itself is not a serious problem for human taggers. Once the lexicographer's manual is clear about criteria of sub- and transcategorisation, correction is mainly a matter of consistency of analysis and application. However, manual correction of a 20 million words corpus is not feasible.

Because a functional application of the tagset leads to more possible taggings of a single word form, it was to be expected this was going to be more difficult for an automatic tagger. But it was not sure in advance to what extend this was going to be prohibitive.

The automatic tagging is done by means of a combination of statistical techniques (De Does & Van der Voort van der Kleij, 2002). We will here present the most salient results. On the basis of the training corpus, the tagger accuracy can be estimated at 97.6 % on Part of Speech, and at 92% on the full tagset. Analysis of the tagger output shows that on the PoS level, the most difficult distinctions are those between residual and noun, adjective and adverb and also between adjective and verb, which confusion is caused by the mood feature ‘participle’.

	ADJ	ADV	NOU	RES	VRB
ADJ	5560	276	185	12	167
ADV	257	7661	100	4	36
NOU	117	56	23448	116	179
RES	30	12	634	1569	17
VRB	79	22	149	2	12464

Table 1: confusion matrix for 5 important PoS. Rows correspond to actual values and columns to tagger assignments.

But the most problematic is defining the function for main verbs: transitive, intransitive, reflexive or impersonal use.

	Not Main	Imp	Intr	Refl	Trs
Not Main	4386	31	289	15	198
Imp	39	27	35	0	5
Intr	304	22	1846	12	476
Refl	3	0	12	87	93
Trs	215	2	368	42	3950

Table 2: verb function, accuracy 82.66 %

5. The Future: PAROLE and Historical Dutch

Our experience with tagging the PAROLE-corpus is relevant for a new, long term project at our institute: the Integrated Language Database of 8th - 21st Century Dutch (ILD) (Kruyt, 2000; www.inl.nl). Apart from dictionary and lexicon data, the ILD will contain a text corpus covering all these centuries. The corpus will be annotated for lemma and PoS, in which process historical lexica will be used. Obviously, we will have to tailor tagset and methodology of tag assignment optimally to these purposes.

From the perspective of standardisation, the question will be whether the PAROLE tagset can be applied to historical Dutch as well. In a pilot study, the PAROLE tagset was compared with the fine-grained PoS tagging scheme of the corpus of Early Middle Dutch (Van Dalen-Oskam & Depuydt, 2000). The conclusion was that the PAROLE tagset is less exhaustive, but can be used provided that it will be extended with some extra features from the PAROLE multilingual tagset and with some provisions for phenomena that are characteristic for medieval Dutch. In the short term, we will start research into the suitability of the PAROLE tagset for younger but still historical Dutch.

As demonstrated above, the methodology of tag assignment is a more complex issue. Questions to be answered include:

(1) which approach ('formal' or 'functional') is more appropriate in a diachronic framework (cf. 2.1)?

(2) is it feasible to develop a methodology and a tag representation which is compatible with both approaches (e.g. the addition of an extra slot to a 'functional' tag in order to preserve the initial lexical information or vice versa)?

(3) is it possible to make up for the loss of functional information throughout the centuries (2.1) and if so what are the implications for the tagset?

(4) to what extent is automatic PoS tagging feasible and how can we use the 1,6 million words corpus of Early Middle Dutch as a training corpus?

Research into these questions will start in the near future. Having learnt from the PAROLE-experience, we will base our decisions on substantial amounts of corpus data. Final decisions will also be related to an ongoing Dutch-Flemish language project, the Spoken Dutch Corpus, in order to harmonise Dutch resources.

6. Acknowledgements

We thank J. de Does (tagger) and J. van der Voort van der Kleij (lexicon) for their critical comments and for their contribution to the discussion.

7. References

ANS (1984). Algemene Nederlandse Spraakkunst. Onder redactie van G. Geerts et al. Wolters-Noordhoff, Groningen.

ANS (1997). Algemene Nederlandse Spraakkunst. Second edition. Haeseryn et al. Nijhoff, Groningen.

Brants, T. (2000). TnT - a statistical part-of-speech tagger. In Proceedings of the 6th Applied NLP Conference, ANLP-2000, April 29 - May 3. Seattle, WA..

Daelemans, W. et al. (2001). TiMBL: Tilburg Memory Based Learner, version 4.0, Reference Guide. ILK Technical Report 01-04.

De Does, J. & J. van der Voort van der Kleij (2002). Tagging the Dutch Parole Corpus. Submitted to Proceedings CLIN 2002.

Dutilh-Ruitenberg, M.W.F., (1994). A Comparative Report on Morphosyntactic Categories in Dutch as encoded in the CELEX Dutch Lexical Databases. Augmented with ten proposals for Dutch. INL Working Papers 1994-01.

Dutilh, T., S. Raaijmakers & T. Kruyt (1996). Tagset for Dutch Morphosyntactic Corpus Annotation. Parole Task 4.1.4a . INL Working Papers 96-02.

Flores, S. (1996) Synthesis for the parametrisable information for morphology. (ID: P-WP1.1-MEMO-ERLI-3) LE-PAROLE, October 1996.

Kruyt, J.G. (2000). Towards the Integrated Language Database of 8th-21st Century Dutch. In Revue française de linguistique appliquée V-2, 33-44. At www.inl.nl

Van Dalen-Oskam, K. & K. Depuydt (2000). Lemmatisering en codering in het VMNW-corpus. Internal paper.

Volz, N. & S. Lenz (1996): Multilingual Corpus Tagset Specifications, MLAP PAROLE 63-386 WP 4.1.4. IDS, Mannheim.

Van der Voort van der Kleij, J. & J.G. Kruyt (1997). Lexicon for a linguistic annotation of Dutch text. In: TELRI newsletter 5, 1997, 32—35.