Language variety at the Dutch Language Institute

1. Introduction

The *Instituut voor de Nederlandse Taal* (Dutch Language Institute) aims to be a widely accessible scholarly institute in the field of the Dutch language. In addition, the institute strives to take on a central position in the Dutch-speaking world (i.e. Belgium, the Netherlands, Suriname and the Dutch Caribbean) as a developer, keeper and distributor of corpora, lexica, dictionaries and grammars. With these sustainable language resources, all resulting from scholarly methods, the Dutch Language Institute provides the necessary building blocks for the study of Dutch. With regard to language variety, the institute pays attention to variation in language between the various Dutch-speaking countries as well as to regional language variety within these countries.

The *Nederlandse Taalunie* (the Union for the Dutch Language) is the international organisation in which the Netherlands, Belgium and Suriname combine their strengths to support Dutch at home and around the world in order to keep the language as dynamic and vigorous as it is today. The *Taalunie* stimulates people and parts of society to use Dutch in the most effective and efficient way possible by developing language policies and services.

Many of the language materials that are developed and/or hosted at the Dutch Language Institute are used by the *Taalunie* to facilitate the exchange of knowledge of Dutch all over the world. They include, amongst others, several historical and contemporary dictionaries of Dutch, a wordlist for spelling purposes, a series of bilingual dictionaries and a portal on the grammar of Dutch.

2. Language variety between the Dutch-speaking countries

Dutch is used in Belgium, the Netherlands, Suriname and the Dutch Caribbean. It is obvious that, although there is a common core to the language, all these countries have their own variety of Dutch characterised by pronunciation, word choice and grammatical peculiarities. The geographical location of the countries in which Dutch is spoken also has its influence on language variety. In recent years, for instance, the Dutch variant in the Netherlands and Belgium has gained many loanwords from English. The same applies to Dutch in Suriname, where there is also plenty of influence from the indigenous languages. In the Caribbean variant of Dutch, Papiamento often plays a major role.

To be considered Dutch and, as such, to get an entry in the wordlist, a loan-word has to comply with a few rules: it must be spelled and inflected according to Dutch rules (I) and it must have been found several times in an entirely Dutch context (II). It is not a rule but it certainly does help loanwords to be integrated when they also appear as part of one or more compounds and derivations (III). The English loanword *uploaden* 'to upload' is used as an illustration of this.

(I) Dutch inflection

	(present)	(past)			
1 sg	upload	uploadde			
2 sg	uploadt	uploadde			
3 sg	uploadt	uploadde			
1, 2, 3 pl	uploaden	uploadden			

infinitive uploaden
past participle geüpload
present participle uploadend

(II) Dutch context

- Google biedt grootgebruikers van opslag de mogelijkheid om harde schijven op te sturen. Het bedrijf uploadt die dan naar de cloudopslag van Google.
- Vaak is voor het **uploaden** een username en een password vereist.
- Bestanden van de server ophalen wordt downloaden genoemd en het verzenden van bestanden naar de server heet uploaden.
- Wanneer u een bestand probeert te sluiten terwijl dit wordt geüpload, wordt er een dialoogvenster weergegeven met de voortgang van de upload.

(III) Dutch compounds and derivations

- AFRC wil de video's toegankelijker maken en is een grote uploadactie naar YouTube gestart.
- Door de toenemende vraag naar data, snelheid en connectivity én uploadcapaciteit zijn er een aantal ontwikkelingen in kabelland.
- De uploadsnelheid is de snelheid waarmee een op een computernetwerk aangesloten apparaat data naar het netwerk kan sturen.
- Wij kunnen jouw (bedrijfs)presentatie door middel van een data-upload op de USB stick zetten, waardoor deze USB stick een perfect promotieartikel wordt.

At present the Dutch Language Institute is working on two projects that deal with language variety between the Dutch-speaking countries on a daily basis.

3. Woordenlijst der Nederlandse Taal

The Woordenlijst Nederlandse Taal (Wordlist of Dutch; www.woordenlijst.org), subsequently referred to as the Woordenlijst, provides the correct spelling of almost

190,000 Dutch words from the entire language area. The complete *Woordenlijst* is available online free of charge. The paper edition, traditionally called the *Groene Boekje* (Green Booklet), referring to the colour of the cover, contains a selection of the online *Woordenlijst*, namely the 5,000 most frequent words in Dutch as well as the 47,000 words that were most frequently looked up by users in the previous online edition of the *Woordenlijst*.



Fig. 1: Het Groene Boekje

The focus of this project is on the correct spelling of Dutch words, information like the part of speech, word gender, word sense and word geography are only of importance when they have consequences for the way in which the word is written. Whether a word is known in the complete language area or only in a small part of it does not influence its spelling. For this reason, the *Woordenlijst* does not give geographical information. What is important is that Dutch language users all over the area can rely on the wordlist to give them the correct spelling of the words they are looking for.

To cover the whole language area, corpus material is collected. The corpus for the *Woordenlijst* predominantly consists of newspaper and web material. Unfortunately, not all language areas provide the same amount of data, which results in a corpus that is still rather imbalanced in geographical terms. For the Netherlands and Belgium there is always a huge amount of data; for Suriname and the Dutch Caribbean, unfortunately, a lot less material is available. To enable the editors to make a representative choice from the material, subcorpora are employed for each country. This enables them to collect interesting words from underrepresented parts of the language area that would otherwise disappear in the wealth of words in the overall corpus. The most recent edition of the *Woordenlijst* contains about 2,500 words that are mainly found in Suriname and some 400 words from the Dutch Caribbean. Out of a total of almost 190,000 words in the online wordlist this might not look like much, but it must be borne in mind that the vast majority of Dutch words are common to the entire language area.

4. Algemeen Nederlands Woordenboek

The second project dealing with language variety between the Dutch-speaking countries is the *Algemeen Nederlands Woordenboek* (Dictionary of Contemporary Dutch; http://anw.ivdnt.org), subsequently referred to as the ANW. In this dictionary, extended lexicographic descriptions are given of more than 75,000 contemporary Dutch words, and this amount is growing daily. The dictionary is compiled on the basis of the *Corpus Hedendaags Nederlands* (Corpus of Contemporary Dutch), a collection of over 800,000 texts taken from magazines, news broadcasts and legal documents.

The ANW does not only explain word senses but also provides information on, for instance, usage, register, collocations and word relations. Lexical variety is one of the key elements of the dictionary. This includes words for concepts that are unique to one particular country and that therefore only occur in one part of the Dutch language area, such as *zwampvis*, a fish species that lives in swamps and can be found in Suriname, or *zeeklas*, the Dutch word in Belgium for a school camp at sea (see Fig. 2 and 3).

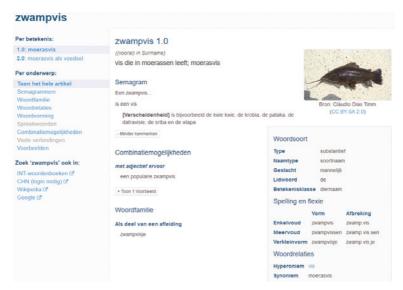


Fig. 2: The entry zwampvis (specific to Suriname) in the Algemeen Nederlands Woordenboek (ANW)

Another kind of variety is where there are different words for the same concept. This is, for instance, the case with a particular kind of fruit that is called *kaneelappel* in Suriname while the common word in the Dutch Caribbean is *schubappel* (see Fig. 4 and 5)

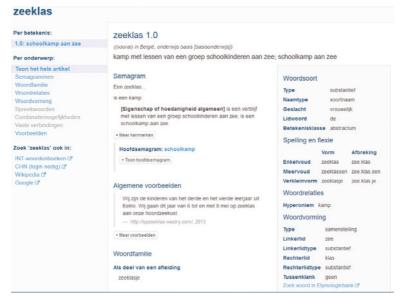


Fig. 3: The entry zeeklas (specific to Flanders) in the Algemeen Nederlands Woordenboek (ANW)

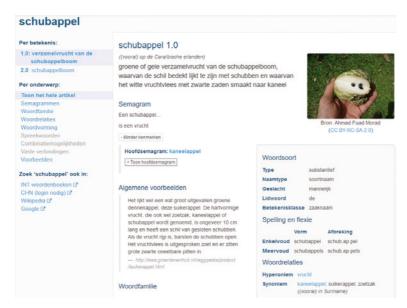


Fig. 4: The entry *schubappel* (specific to the Dutch Caribbean) in the *Algemeen Nederlands Woordenboek* (ANW)



Fig. 5: The entry kaneelappel (specific to Suriname) in the Algemeen Nederlands Woordenboek (ANW)

kleed 1.0

+ Meer voorbeelden

afgepast en afgewerkt stuk weefsel van stevige stof gemaakt om iets mee te bedekken, meestal de vloer, of om ergens overheen te leggen, meestal over een meubelstuk, bijvoorbeeld een tafel



Fig. 6: The entry kleed in the sense of 'carpet' (specific to the Netherlands) in the Algemeen Nederlands Woordenboek (ANW)

It becomes really confusing when the same Dutch word denotes different concepts in the various parts of the Dutch-speaking area. This is the case with *kleed*, for instance, where it is something put on the floor under a table or in front of the fireplace in the Netherlands and the word for a dress in Belgium (see Fig. 6 and 7).

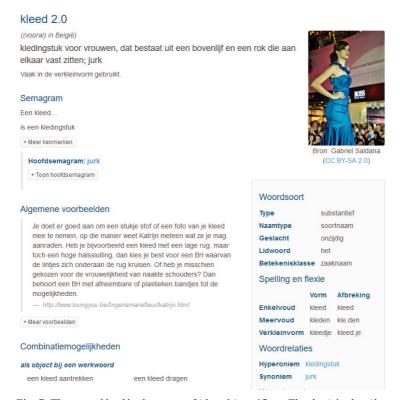


Fig. 7: The entry *kleed* in the sense of 'dress' (specific to Flanders) in the *Algemeen Nederlands Woordenboek* (ANW)

The search options in the ANW include the possibility to search for words from a certain geographical area. This enables users to find information on words and expressions that are typical for one or more parts of the Dutch language area (see Fig. 8).

Like the Woordenlijst, the ANW is available online free of charge: http://anw.ivdnt.org/.



Fig. 8: Search option for language variety in the Algemeen Nederlands Woordenboek (ANW)

5. Language variety within individual Dutch-speaking countries

A project that covers regional varieties of Dutch, i.e. the dialects that exist within a country's borders, is the *Dictionary of the Southern Dutch Dialects* (DSDD; https://research.flw.ugent.be/nl/projects/dsdd-dictionary-southern-dutch-dialects-integrated-lexicological-infrastructure-southern), that started in 2017. The project aims to build an infrastructure in which the three main dictionaries for the Brabantic, Limburgian and Flemish dialects will be available in one dictionary portal, and the data in these dictionaries can be processed for combined search questions. It is a collaboration between the Dutch Language Institute and Ghent University, with the former providing technical support and its experience in linking lexicographical data, and the latter providing the dictionary data and editorial staff. The institute will also host the portal once the project has been finished in 2019. At the moment the linking tool is being tested by processing about 1,500 entries from the three dictionaries (see Fig. 9 and 10).

This is a more complex process than it sounds because the dictionaries have not always worked in the same way and sometimes the relations can be found on different lexical levels. For instance, sometimes the diminutive is included in the main dictionary entry and sometimes it has an entry of its own. The institute helps the editors find solutions to this and other issues.

After solving all these issues, preparing the data and building the dialect portal, the next step will be to integrate other dictionaries of Dutch dialects.

Another option is the integration of the contemporary dialectal variants of Dutch and the historical corpora and dictionaries that are already available at the institute. A small pilot project has already been executed, showing promising results. Using the same tool as for linking concepts in the dialect dictionaries, we tried to link the 13th century words for animals described in the *Vroegmiddelnederlands Woordenboek* (Dictionary of Early Middle Dutch; http://vmnw.ivdnt. org) to corresponding entries in the Flemish dialect dictionary. In more than 75 per-

cent of the cases, a 13th century word could be linked to a modern Flemish concept. The pilot project resulted in a dataset of 524 concepts which covered 790 entries in the *Vroegmiddelnederlands Woordenboek* and 173 entries and 2,694 keywords in the Flemish dialect dictionary. And this was only in connection with the names of animals!



Fig. 9: Concepts and definitions of dialect data in the integration tool of the Dutch Language Institute

wbd III_4_3 akkermunt munt 14_Akkermunt meutha Toon munt wbd III_4_3 akkermunt paardsbloemheilige 14_Akkermunt Akkermunt Akkermunt Toon paardsbloemheilige 14_Akkermunt wbd III_4_3 akkermunt paardsbruinbeilige 14_Akkermunt Akkermunt Toon paardsbruinbeilige 14_Akkermunt wbd III_4_3 akkermunt knapperd 14_Akkermunt akkermunt Toon paardsbruinbeilige 14_Akkermunt Akkermunt Roon paardsbruinbeilige 14_Akkermunt Akkermunt Roon paardsbruinbeilige 14_Akkermunt Akkermunt Roon paardsbruinbeilige 14_Akkermunt Akkermunt Roon paardsbruinbeilige 14_Akkermunt Roon paardsbruinbe	Toon 5	∨ rijen							Eerste	Vorige	1 2	3	Volgende	Laatst
wbd III_4_3 akkermunt munt 14_Akkermunt mentha Toon munt winding munt wbd III_4_3 akkermunt paardsbloemheilige 14_Akkermunt Akkermunt Akkermunt Toon paardsbloemheilige wbd III_4_3 akkermunt paardsbruinbeilige 14_Akkermunt Akkermunt Akkermunt Toon paardsbruinbeilige wbd III_4_3 akkermunt knapperd 14_Akkermunt akkermunt Akkermunt Toon wafanten wbd III_4_3 akkermunt muntkruid 14_Akkermunt akkermunt Toon muntkruid Toon muntkruid Toon muntkruid Wbd III_4_3 akkermunt muntkruid 14_Akkermunt akkermunt Toon muntkruid Wbd III_4_3 akkermunt muntkruid 14_Akkermunt Akkermunt Wbd III_4_3 akkermunt muntkruid 14_Akkermunt Akkermunt Wbd III_4_3 akkermunt muntkruid Wbd III_4_3 akkermunt muntkruid Wbd III_4_3 akkermunt muntkruid Wbd III_4_3 akkermunt muntkruid Wbd III_4_3 akkermunt Wbd III_4_3 akkerm					14_Akkermunt									
wbd III_4_3 akkermunt paardabloemheilige 14_Akkermunt Akkermunt Toon paardabloemheilige wbd III_4_3 akkermunt paardabloemheilige 14_Akkermunt Akkermunt Toon paardabloemheilige wbd III_4_3 akkermunt paardabruinheilige 14_Akkermunt Akkermunt Toon paardabruinheilige wbd III_4_3 akkermunt knapperd 14_Akkermunt akkermunt Toon wafnamen wbd III_4_3 akkermunt munfkruid 14_Akkermunt akkermunt Toon munfkruid wafnamen munfkruid wafnamen munfkruid akkermunt Toon munfkruid wafnamen waf	dictionary (publication_id (lemma 0	keyword	matching_lemma (relation (definition (definition_question	comm	ent k	eyword	id	keywo	rd_org
web III_4_3 akkermunt paardsbruinbeilige 14_Akkermunt Akkermunt Akkermunt Toon paardsbruinbeilige wid III_4_3 akkermunt knapperd 14_Akkermunt akkermunt Akkermunt Toon windenting wid III_4_3 akkermunt montkruid 14_Akkermunt akkermunt Akkermunt Toon muntkruid windentin	wbd	III_4_3	akkermunt	munt	14_Akkermunt			mentha					munt	
wid III_4_3 akkermunt knapperd 14_Akkermunt akkermunt akkermunt Toon knapperd wid III_4_3 akkermunt muntkruid 14_Akkermunt akkermunt Akkermunt Toon muntkruid watenten	wbd	III_4_3	akkermunt	paardsbloemheilige	14_Akkermunt			Akkermunt					paardsbl	oemheili
wid III_4_3 akkermunt munkruid 14_Akkermunt akkermunt Toon munkruid walanten	wbd	III_4_3	akkermunt	paardsbruinheilige	14_Akkermunt			Akkermunt					paardsbe	uinheilig
varianten	wid	III_4_3	akkermunt	knapperd	14_Akkermunt			akkermunt					knapper	d
Eerste Vorige 1 2 3 Volgende Laatst	wid	III_4_3	akkermunt	muntkruid	14_Akkermunt			akkermunt					muntkru	id
									Eerste	Vorige	1 2	3	Volgende	Laatst

Fig. 10: Linked concepts in the integration tool of the Dutch Language Institute

Expanding the linking to all historical dictionaries of Dutch will give even better results but will also present new challenges. To mention one of them, how specific or generic is the word an informant has entered for a given concept? When we, for instance, find the word *bruwier* both for Dutch *kiekendief* 'harrier' and *wouw* 'kite', technically birds of two different genera, does this mean that the informant

considers these birds to be alike, or does he or she use a more generic word for both which could also be used for other birds of related genera, for instance Dutch *valk* 'falcon'?

Although this makes fully automated linking more complex, it certainly provides many new insights into the development of the Dutch dialects and in the relation between these dialects and standard Dutch.

dictionary	entry	variant
WVD (Flemish)	kiekendief	brawier, bruwier, kiekendief, klamper, valk, witte valk
WBD (Brabantic)	kiekendief	brouwier, grijskopvalk, hazenklamper, kiekendief, kiekenklamper, kiekenplot, kiekensperwer, kiekenvalk, kiekjesdief, kikkerdief, klamper, konijnenklamper, koop, kuikendief, moeruil, plot, witte klamper, zonnedraaier
WLD (Limburgian)	kiekendief	blotser, buizerd, duivenman, duivenpiet, hennendief, hennenschelm, kiekendief, klamper, kuikendief, kuikempiet, kuikenschelm, rietwouw, sperwer, stootkop, stootsvogel, stootvogel, takraaf, valk, zilvervalk
MNW (1250-1550)	kiekendief	kiekendief, kukendief
WNT (1550-1921)	kiekendief	kiecken-dief, kiekendief

Table 1: Representations of the concept/entry KIEKENDIEF 'harrier' in the dialect dictionaries and historical dictionaries of Dutch

dictionary	entry	variant
WVD (Flemish)	wouw	bruwier
WBD (Brabantic)	wouw	wije
ONW (< 1200)	*weio	uueippo
VMNW (1200-1300)	wiwe; wuwe	wywen; wië, wiwe, wouwe, wu, wuw
MNW (1250-1550)	wie (IV); wiwe; wuwe	wihe, wye; wywe; wië, wiwe, wouwe, wu, wuw
WNT (1550-1921)	wouw	wauwe, wije, wiwe, wou, wouw, wouwe, wu, wuw, wuwe

Table 2: Representations of the concept/entry WOUW 'kite' in the dialect dictionaries and historical dictionaries of Dutch

6. Conclusion

The Dutch Language Institute pays attention to all variants of Dutch, both between the countries that are part of the Union for the Dutch Language and within their borders. The Dutch used in Belgium differs from the Dutch used in the Netherlands, but the Dutch of Antwerp or Amsterdam also differs from the Dutch in Ghent or Utrecht. At the moment various projects are running at the institute that deal with this kind of variation, all with their own focus on language variety. The Woordenlijst Nederlandse Taal, the Algemeen Nederlands Woordenboek and the Database of the Southern Dutch Dialects all play their role in the description of Dutch. Future projects will build on these projects, collecting and describing new data, linking these data to existing datasets and preparing the data for the various groups of intended users, thus providing a solid basis and environment for many kinds of contemporary and historical linguistic research and an easily accessible platform on linguistic knowledge to be used by language users in the entire Dutch-speaking community as well.

Bibliographical information

This text was first published in the book:

Tanneke Schoonheim/Johan Van Hoorde (eds.) (2019): Language variation. A factor of increasing complexity and a challenge for language policy within Europe. Contributions to the EFNIL Conference 2018 in Amsterdam. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences. [250 pages.]

The electronic PDF version of the text is accessible through the EFNIL website at:

http://www.efnil.org