# Syntactic Profiles in Secondary School Writing Using PaQu and SPOD

Hoeksema, Jack ; de Glopper, Kees; van Noord, Gertjan

*Published in:*
CLARIN

*DOI:*
10.1515/9783110767377-027

Jack Hoeksema, Kees de Glopper, and Gertjan van Noord

# Syntactic Profiles in Secondary School Writing Using PaQu and SPOD

**Abstract:** SPOD is part of the PaQu website created as a CLARIN project. It allows one to generate a syntactic profile of a corpus based on the output of the automatic parser Alpino. It runs a long sequence of queries and provides quantitative information about constituents, sentence types, coordination, length of constituents, and so on. In this chapter, we employ SPOD and the rest of PaQu to analyse a part of the *Schrijfmeterscorpus* of secondary school essays. We use a small subsection of the SPOD output for this purpose, in particular those syntactic properties that correlate most reliably with academically oriented texts. We show that SPOD is able to distinguish, on the basis of these variables, among grades and school types.

**Keywords:** automatic parsing, writing, query, secondary education

## 1 Introduction

Online corpora usually do not provide much in the way of syntactic information. Sometimes they allow searches for parts of speech or simple regular expressions, less often they come fully parsed. Even less common is a website that comes with a parser and a query interface. PaQu is such a website, developed as part of the Dutch CLARIN infrastructure, and has turned out to be useful for studying syntactic patterns in corpora (see Bloem 2020; Bouma 2017; Odijk 2015, 2020; Odijk et al. 2017; van der Wouden et al. 2015; van Noord et al. 2020). The website is in Dutch, and can only be used for analysing Dutch corpora. Users with an account can upload their corpus, have it parsed by the Alpino parser (Bouma, van Noord, and Malouf 2001; van Noord 2006) and query it to find out for example how many indirect questions it contains. There is a basic interface window allowing users

**Acknowledgements:** The development of SPOD has been funded by the Dutch national CLARIN project Common Lab Research Infrastructure for the Arts and Humanities, CLARIAH.

**Jack Hoeksema,** University of Groningen, Groningen, the Netherlands, e-mail: j.hoeksema@rug.nl
**Kees de Glopper,** University of Groningen, Groningen, the Netherlands,
e-mail: c.m.de.glopper@rug.nl
**Gertjan van Noord,** University of Groningen, Groningen, the Netherlands,
e-mail: g.j.m.van.noord@rug.nl


https://doi.org/10.1515/9783110767377-027

to search for combinations between words (for example all adjectives modifying a particular noun, or all nouns modified by a particular adjective). There is also a window in which power users can write Xpath 2.0 queries to search for syntactic patterns. Xpath is a query language for XML.

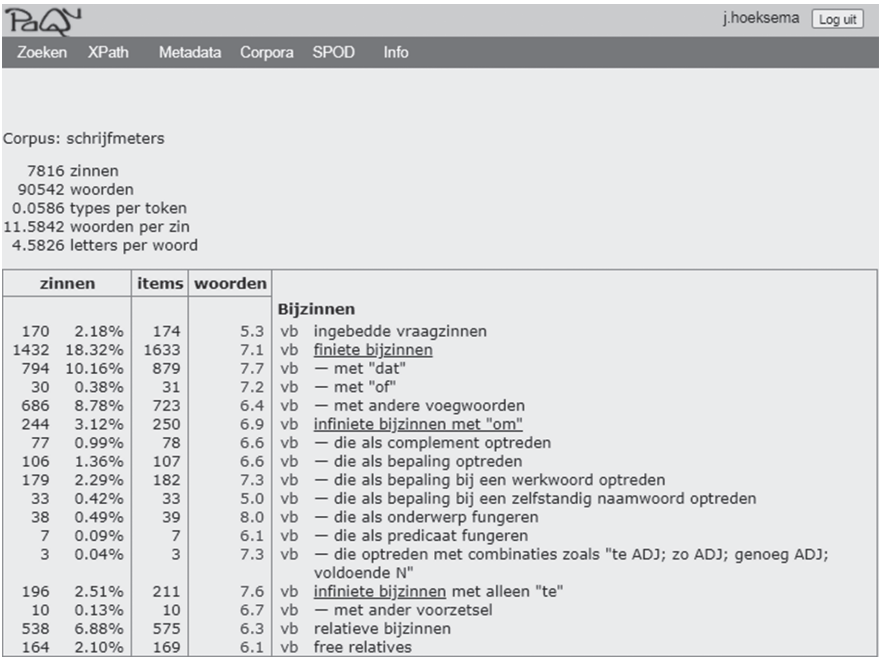
A new feature of PaQu is SPOD, the Syntactic Profiler of Dutch, which uses a battery of built-in XPath queries to provide an overview of syntactic (and some lexical) properties of the data.[1] The queries make heavy use of dedicated macro's and require knowledge of the underlying Alpino parser. Such queries are difficult to make for non-expert users, even if they are familiar with corpus linguistics, and providing this ready-made query set will help make the PaQu tools more accessible for them. By clicking on the query link, it is possible to open an XPath tab (part of PaQu) to make the query sensitive to corpus metadata. The latter are corpus-specific, and may vary according to the specs and purpose of the corpus. Among the data provided by SPOD are the following:

- basic information concerning the corpus: number of sentences, word (tokens), type/token ratio, mean sentence length, and mean word length;
- part of speech listings: numbers of nouns, verbs, adjectives and so on, including their subcategories, such as number of neuter and common gender nouns, plurals, inflected and noninflected adjectives;
- frequency of four types of main clauses: declarative, wh-questions, yes/no questions, and imperatives;
- frequency and average length of types of subordinate clauses;
- frequency of various subtypes of comparatives;
- frequency of coordinations, subdivided by conjunction word, number of conjuncts, and category of conjuncts;
- frequency and mean length of four phrasal subtypes: NP, PP, AP and AdvP
- frequency of subtypes of PP: attributive, predicative, adverbial, complement;
- frequency of verb clusters of various types;
- information about particle verbs (placement in or outside verb cluster)
- levels of finite clausal embedding;
- topicalization and extraction data;
- parser success (words skipped by the parser, sentences with a partial parsing).

Potential applications for SPOD are manifold. One can extract information about the corpora made available on PaQu, such as the corpus of spoken Dutch, Lassy Small, Basilex, and Wablieft. This can then be used for comparison with a user-provided corpus, uploaded at the PaQu site. A potential application is stylistic

**1** SPOD is available via https://www.let.rug.nl/alfa/paqu/spod.

research. There is a fair amount of n-gram based analysis of texts in computational humanities, but PaQu makes syntactic comparisons possible, at the level of individual differences among writers, but also at the level of text types, by comparing, for example, newspaper texts and academic papers, or unprepared spoken language with written genres. See van Noord et al. (2020) for more information on the set-up and main features of SPOD and PaQu. That paper also contains information about the accuracy of the Alpino parser. As with all automatic parsers, accuracy varies with text types, and sometimes manual inspection of the parsed sentences will be necessary to verify results. SPOD normally returns numbers, but it has a built-in option which lists all sentences that were selected from the corpus by a query.




| zinnen | | items | woorden | | |
|---|---|---|---|---|---|
| | | | | | **Bijzinnen** |
| 170 | 2.18% | 174 | 5.3 | vb | ingebedde vraagzinnen |
| 1432 | 18.32% | 1633 | 7.1 | vb | finiete bijzinnen |
| 794 | 10.16% | 879 | 7.7 | vb | — met "dat" |
| 30 | 0.38% | 31 | 7.2 | vb | — met "of" |
| 686 | 8.78% | 723 | 6.4 | vb | — met andere voegwoorden |
| 244 | 3.12% | 250 | 6.9 | vb | infiniete bijzinnen met "om" |
| 77 | 0.99% | 78 | 6.6 | vb | — die als complement optreden |
| 106 | 1.36% | 107 | 6.6 | vb | — die als bepaling optreden |
| 179 | 2.29% | 182 | 7.3 | vb | — die als bepaling bij een werkwoord optreden |
| 33 | 0.42% | 33 | 5.0 | vb | — die als bepaling bij een zelfstandig naamwoord optreden |
| 38 | 0.49% | 39 | 8.0 | vb | — die als onderwerp fungeren |
| 7 | 0.09% | 7 | 6.1 | vb | — die als predicaat fungeren |
| 3 | 0.04% | 3 | 7.3 | vb | — die optreden met combinaties zoals "te ADJ; zo ADJ; genoeg ADJ; voldoende N" |
| 196 | 2.51% | 211 | 7.6 | vb | infiniete bijzinnen met alleen "te" |
| 10 | 0.13% | 10 | 6.7 | vb | — met ander voorzetsel |
| 538 | 6.88% | 575 | 6.3 | vb | relatieve bijzinnen |
| 164 | 2.10% | 169 | 6.1 | vb | free relatives |

**Figure 1:** Screenshot of SPOD showing frequency and average length of types of clauses.

The screenshot in Figure 1 illustrates the output for a small part of SPOD. The full output for all variables is too large to show here. As you can see, SPOD, like the rest of PaQu, is in Dutch, and only analyses Dutch texts.

By clicking on one of the elements marked in blue, it is possible to obtain further information: clicking on the number conjures up a graph, showing frequency per unit of length (compare Figure 2), and clicking on *vb*, takes you from SPOD to the XPath window in PaQu where the query is ready to run.

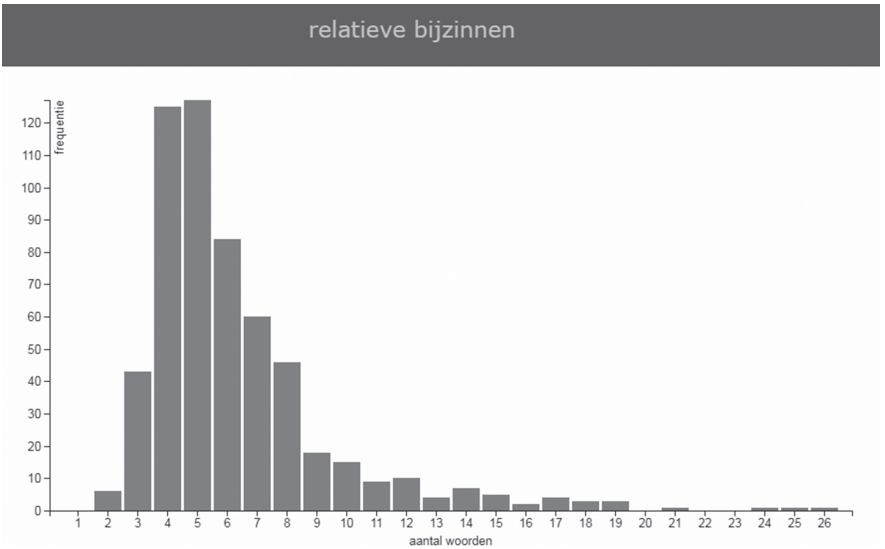



**Figure 2:** Screenshot of SPOD output: frequency (Y axis) by length (X axis) for relative clauses.

In this chapter, we use the SPOD/PaQu tools to analyse student essays from various school types (from the first three years of secondary education), and compare them along a number of syntactic dimensions that we know from previous research (cf. Hoeksema, de Glopper, and van Noord 2021) to be particularly sensitive to developmental change, in particular insofar as it involves development toward more highly academic writing styles. Syntactic properties that do not change over time, such as V2 word order in main clauses, are unlikely to vary among school types and are not included in this study. Instead, we focus on features that become more important over time and are associated with academic registers, and use the PaQu tools to see if and to what extent our main hypothesis is supported, viz. that such features will not just be a monotonically increasing function of age, but also of school type, in which higher scores are associated with more academically oriented school types.

The chapter is structured as follows: in Section 2, we sketch the Dutch system of secondary education and the various types of schools it consists of, in Section 3 we introduce our corpus, in Section 4 we discuss the variables we selected for this study and in Section 5 we present our main findings. Section 6 discusses these findings. Section 7 contains our conclusions.

## 2 School types in Dutch secondary education

Unlike primary education, which is uniform for all children attending regular education, Dutch secondary education is divided into pre-vocational secondary education (VMBO, duration four years), senior general secondary education (HAVO, duration five years) and pre-university education (VWO, duration six years).[2] Dutch children are given a secondary school level advice in the last year of primary school, typically at the age of 12.

The gymnasium is a VWO-type school which prepares students for study at the university, and offers them, along with the sciences, humanities and modern languages, classes in Greek and Latin. Atheneum is likewise a preparation for university level study, but without the classical languages.

HAVO students are not directly admitted to universities, but may go on to higher level vocational schools as well as applied universities (called HBO in Dutch, an acronym for higher vocational education). The curriculum consists of modern languages, humanities and sciences.

VMBO TL is a school type which prepares students for midlevel vocational schools (MBO), whereas VMBO BK is a more practically oriented version of the same. Students typically go on to vocational schools for hairdressers, auto mechanics, plumbers, nurses, caterers, as well as various types of office jobs.

## 3 The corpus

We make use of a 90,000 word corpus of essays, a part of the Schrijfmeterscorpus (cf. de Glopper and Prenger 2013; Pander Maat et al. 2019). This corpus was collected in the academic year 2012–13 by the former Expertise center for Language, Education and Communication (ETOC) at the University of Groningen. The essays in our corpus are based on the same writing assignment for all school types (a letter describing characteristics of the Netherlands for a Swedish girl that will soon join the class) in order to make them fully comparable. A select number of syntactic variables in SPOD will be tracked. Each query associated with one of these variables can be made sensitive to metadata such as school type, or school year (the corpus only covers the first three years of secondary school), by clicking on the *vb* button in the associated line of SPOD, and continuing in the

**2** For an overview of the Dutch education system, see https://eacea.ec.europa.eu/national-policies/eurydice/content/netherlands.

Xpath window of PaQu. Table 1 provides an overview of the size of the corpus (in number of sentences) per school type and year.

**Table 1:** Schrijfmeterscorpus: number of sentences per school type and year.

| | gymnasium | atheneum | HAVO | VMBO TL |
|---|---|---|---|---|
| year 1 | 562 | 230 | 1044 | 443 |
| year 2 | 605 | 391 | 855 | 688 |
| year 3 | 762 | 529 | 817 | 776 |

Henceforth, we combine the gymnasium and atheneum data into the category VWO. The essays were scored on a number of issues (involving structural properties of the text, such as cohesion, clarity of exposition, and so on) by a panel of experts (three raters per essay, randomly selected from a pool of eight raters). By and large, these scores show differentiation by age and school type. Scores were on a scale from 50 (minimum) to 150 (maximum). The (Cronbach alpha) reliability of the scores was 0.86.

Table 2 contains the average scores and standard deviation for the three school types in our corpus.

**Table 2:** Schrijfmeterscorpus: scores per school type.

| schooltype | average score | S.D. |
|---|---|---|
| VMBO TL | 98 | 13.3 |
| HAVO | 102 | 11.5 |
| VWO | 112 | 14.9 |

From this, we conclude that the overall ranking of essay quality mirrors the ranking of secondary school types in terms of academic rigor. In the remainder of this chapter, we want to see if this ranking is also reflected by differences at the level of sentence structure that are independent of textual qualities such as textual coherence, explicitness of argumentation, and clarity. In a number of cases to be discussed below, we add comparisons to some additional corpora that were available to us, and were parsed and queried by the same PaQu tools. This was done when it was necessary to make a point about the nature of the syntactic variables that were used in this study. They are presented in the next section.

## 4 Syntactic variables

SPOD allows us to look at a plenitude of syntactic features, not all of which are expected to be of interest for a comparison of school types. Recall that our working hypothesis is that the variables that show continuous development over time from primary school to university level writing will also distinguish texts by secondary school students of the same age, but different school types.

Some of the features identified by Biber and Gray (2010, 2016); Staples et al. (2016) as characteristic of academic writing were studied in Hoeksema, de Glopper, and van Noord (2021), and found to be relevant for analysing the developmental trajectory from early elementary school writing to academic writing. They can be seen as reflecting steady increases in phrasal complexity. The idea that academic texts differ from colloquial speech and writing in sentential complexity as well, in particular in having more subordinate clauses, has been challenged by D. Biber and his associates. They argue, instead, that academic registers abound in complex phrases, in particular elaborate noun phrases, and not in layers upon layers of clausal embedding. In short, they reject earlier accounts of academic writing as being more elaborate than other types of writing, and propose that compactness, or density, is a more apt characterization. However, this finding does not necessarily generalize to the academic registers of languages other than English. In particular, Hoeksema, de Glopper, and van Noord (2021) lists increasing levels of finite embeddings as a developmental trait for Dutch, monotonically rising all the way from elementary school writing to university level and professional academic texts. Given our focus on Dutch, we decided to include sentential complexity among the variables that may characterize differences across school types.

A striking feature about academic registers is their highly nominal character (Heylighen and Dewaele 2002). The nouns-to-verbs ratio is much higher than for fiction, or spoken language. The nominal character of academic texts is further reflected by higher frequencies for ad-nominal modifiers such as attributive adjectives, PPs and relative clauses.

In this chapter we consider the following variables: noun/verb ratio, nominal modifiers, and levels of sentential embedding. One of the features most strongly correlated with academic writing in Biber and Gray (2010, 2016), viz. nouns serving as premodifiers to nouns, is not included here since Dutch does not use nouns in this way. Just to illustrate this point, consider the linguistic term *noun phrase*. Dutch renders it as either an adjective plus noun combination (*nominale woordgroep* 'nominal phrase', or as a compound, written and treated as a single word, for example *substantiefgroep*). One of the developmental variables in Hoeksema, de Glopper, and van Noord (2021), coordination type, is not included in our study either. We intend to study aspects of coordination elsewhere.

# 5 Main findings

## 5.1 Noun/verb ratio

In Table 3, we tabulate nouns and verbs for the three school types in our corpus. For the sake of comparison, we also include the pertinent data from the university essay corpus used in Hoeksema, de Glopper, and van Noord (2021), a corpus consisting of four literary novels by Renate Dorrestein, and the corpus of spoken Dutch (CGN – cf. Oostdijk 2002). Note that the score for VWO, the school type preparing for university level higher education, has a lower N/V score than the university corpus, but it should be noted here that we only have data for the first three years of secondary school, and may expect a rising score for the upper level of secondary school, which takes another 3 years.

**Table 3:** Nouns, verbs, noun/verb ratio.

| subcorpus | N | V | N/V |
|---|---|---|---|
| VWO | 7848 | 6206 | 1.26 |
| HAVO | 6646 | 5294 | 1.26 |
| VMBO TL | 4722 | 4115 | 1.15 |
| University | 52852 | 39434 | 1.34 |
| Dorrestein | 50331 | 57180 | 0.88 |
| CGN (spoken Dutch) | 126199 | 170538 | 0.74 |

An ANOVA with noun/verb ratio as the dependent variable and school type and school year as independent variables yielded no significant results for school year ($F(2, 42) = .006$, $p = .946$), but school type was significant ($F(2, 419) = 6.33$, $p = .002$) and there was an interaction effect of school type and school year ($F(4, 419) = 421$, $p = .002$). The differences between HAVO and VMBO and between VWO and VMBO were significant ($p < .05$).

Academic registers have often been referred to as "nouny", cf. for example the findings in Heylighen and Dewaele (2002). Words that typically co-occur with nouns, such as articles and prepositions were found to correlate highly with academic success in Pennebaker et al. (2014). While the latter study is based on English academic prose, we may interpret Table 3 as providing some evidence that the same is true for Dutch. The data from the Dorrestein novels suggest that a high noun/verb ratio is not typical of Dutch literary writing. However, since the study of literary style is not our main concern here, we will not explore this matter in more detail. In the following subsections, we look for differences among the school types in noun modifiers.

## 5.2 Nominal modifiers

### 5.2.1 Attributive adjectives

In this subsection, we consider attributive versus predicative use among adjectives. Attributive adjectives modify nouns, predicative adjectives are predicates in copular, resultative, and depictive constructions. These various uses are illustrated for English below:

1. Predicative
   - This towel is dry. [copular]
   - I need to rub myself dry. [resultative]
   - The towels were given to us dry, not wet. [depictive]
2. Attributive
   - Hand me some dry towels, please.

In Dutch, attributive adjectives are inflected (they either end in a schwa or have no ending, see Haeseryn et al. (1997) for some discussion and Stowe et al. (2014) on Belgian-Dutch variation). In Hoeksema, de Glopper, and van Noord (2021) we presented data that show a continuous increase of attributive cases among all occurrences of adjectives from early elementary school to academic level and professional writing of attributive adjectives. We expect to find the same trend both across school years (1, 2, or 3) and school types in our corpus.

In Table 4, we present the PaQu counts for attributive adjectives, adjectives in general and the percentage of attributive adjectives in the Schrijfmeters corpus. The numbers 1, 2, and 3 stand for 1st, 2nd, and 3rd year classes, respectively.

**Table 4:** Attributive uses among adjectives in three school types.

| School type | year | all adjectives | attributive | pct. attr |
|---|---|---|---|---|
| VMBO TL | 1 | 496 | 145 | 29.2 |
| | 2 | 626 | 157 | 25.1 |
| | 3 | 861 | 309 | 35.9 |
| HAVO | 1 | 1067 | 377 | 35.3 |
| | 2 | 838 | 289 | 34.5 |
| | 3 | 782 | 282 | 36.1 |
| VWO | 1 | 840 | 307 | 36.5 |
| | 2 | 1025 | 378 | 36.9 |
| | 3 | 1409 | 569 | 40.4 |

An ANOVA with the percentage of attributive cases among adjectives as dependent variable and school type and school year as independent variables yielded no significant effect of school type ($F(2, 418) = 2.42$, $p = .090$). School year was significant overall ($F(2, 418) = 3.22$, $p = 0.041$), but the differences between separate years were not. Interaction of school type and school year was not significant ($F(4, 418) = 1.60$, $p = 0.173$).

### 5.2.2 Attributive and other PPs

Prepositional phrases come in a variety of uses (Pullum and Huddleston 2002; Haeseryn et al. 1997), both in English and in Dutch. They can be predicates (for example, *to be at peace*), adverbials (*we come in peace*), complements to verbs and adjectives (*to hope for peace*, *eager for peace*) and attributive (*country at peace*). Both in Dutch and English, attributive PPs are mostly postnominal (though English to a greater extent than Dutch also has prenominal PPs in compound-like combinations such as *under-the-counter sales, out of pocket expenses*. By and large, the trends among prepositional phrases are similar to those noted for adjectives: a rise in attributive cases (see Table 5).

**Table 5:** Percentage of attributive uses among PPs in three school types.

| School type | Year | PP | attr | pct. attr |
|---|---|---|---|---|
| VMBO | 1 | 411 | 84 | 20.44 |
| | 2 | 679 | 117 | 17.23 |
| | 3 | 687 | 161 | 23.44 |
| HAVO | 1 | 1033 | 215 | 20.81 |
| | 2 | 748 | 174 | 23.26 |
| | 3 | 798 | 197 | 24.69 |
| VWO | 1 | 741 | 177 | 23.89 |
| | 2 | 1055 | 250 | 23.70 |
| | 3 | 1340 | 337 | 25.15 |

Attributive PPs are among the main factors adding complexity to English noun phrases (cf. Berlage 2014). Rising trends per school year are to be expected, given similar results in Hoeksema, de Glopper, and van Noord (2021). The rising trend per school type from VMBO TL to VWO is a new finding, but in line with our hypothesis that developmental patterns on the road from elementary education to university level writing are reflected in school type diversity as well. However, our

findings of increased levels of attributive uses among PPs, though in accordance with Biber and Gray (2010, 2016); Staples et al. (2016) for written varieties of academic English, were not robust enough to be statistically significant.

An ANOVA test with the percentage of PPs that are attributive as the dependent variable and school year and school type as independent variables showed no significant effects. School type is not significant ($F(2, 419) = 2.48$, $p = .085$), nor is school year ($F(2, 419) = 2.22$, $p = .11$). The interaction of schooltype and schoolyear was not significant ($F(4, 419) = 1.35$, $p = .250$). We believe the smallish size of the corpus might be to blame for these non-results.

### 5.2.3 Relative clauses

In the case of relative clauses, we will not compare attributive with non-attributive cases (free relatives) the way we did in the case of prepositional cases (cf. the preceding subsection), because free relatives are comparatively rare anyway (free and headed relatives differ by a factor of 10 in corpora such as Lassy Small) and in our Schrijfmeters corpus they are mostly part of wh-clefts, which brings with it a host of complications (headed relatives have no comparable role in wh-clefts). Instead, we normalize raw counts by calculating occurrences per 10,000 sentences.

In Table 6, we see a notable increase of relative clauses in VWO essays, no increase in VMBO TL essays, and a weak overall growth in HAVO essays. Somewhat surprising is the relatively high score for VMBO TL in year 1. This might be a statistical fluke, in light of the fact that we have only a small sample for year 1 of VMBO TL (compare Table 1 above). The raw numbers of relative clauses suggest that relative clauses are more common with increasing grades and school levels, but corrected for the number of sentences provided by each student, an ANOVA did not find a significant effect of either school year ($F(2, 419) = .48$, $p = .622$) or school type ($F(2, 419) = 1.856$, $p = .158$), nor did it find a significant interaction effect ($F(4, 419) = 1.962$, $p = .099$). The fact that we are unable to trace this growing importance through school types and grades may be due to the smallish size of the corpus already mentioned in the previous paragraph, in combination with the limited frequency of relative clauses.

**Table 6:** Relative clauses: absolute and relative frequencies.

| School type | Year | Rel cl | per 10K sentences |
|---|---|---|---|
| VMBO TL | 1 | 30 | 677 |
| | 2 | 47 | 683 |
| | 3 | 49 | 631 |

**Table 6** (continued)

| School type | Year | Rel cl | per 10K sentences |
|---|---|---|---|
| HAVO | 1 | 62 | 594 |
| | 2 | 48 | 561 |
| | 3 | 58 | 710 |
| VWO | 1 | 47 | 593 |
| | 2 | 87 | 873 |
| | 3 | 135 | 1046 |

## 5.3 Finite embedding

A form of structurally complexity that is often associated with written registers is clausal embedding (measured in clauses per sentence, or per T-unit, cf. Hunt 1970). In this subsection we look at finite embeddings only, such as provided by finite complement clauses, relative clauses and adverbial clauses, and compare complex sentences, involving at least one finite clause embedding, with simple sentences. Other conceivable measures, such as number of nodes per syntactic tree (see Sampson 2013), or maximal length of paths from the root of the tree to its leaves, tend to be highly theory-specific, and hence less likely to be of use, especially when results for different parsers are to be compared. SPOD does not include them. However, finite embeddings can be counted in a theory-neutral way. Table 7 contains data from Hoeksema, de Glopper, and van Noord (2021), showing continuous growth of finite embedding from elementary to higher education (note that these data are from different corpora than the ones considered in this chapter).

**Table 7:** Complex finite clauses in texts by elementary school children (BasiScript), secondary school students (Hofstad corpus), university students and linguists.

| Corpus | *FinEmb* = 0 | *FinEmb* > 0 | *Pct. Finemb* > 0 |
|---|---|---|---|
| BasiScript | 614815 | 128187 | 17.3 |
| Hofstad | 17877 | 10727 | 37.5 |
| UnivStud | 7735 | 5136 | 40.1 |
| Linguists | 3522 | 2966 | 45.7 |

The (maximal) level of finite embedding (referred to in Table 7 as *Fin Emb*) is a variable running from 0 (no embedding whatever) to 6 or 7 in very complex cases. The Schrijfmeterscorpus does not go beyond level 3. This means that the most

complex sentences according to this measure have a finite clause inside another finite clause that is part of yet another finite clause which is part of the main clause. So the measure does not look at the number of clauses in a sentences, but at their hierarchical structure. The following example from the corpus will illustrate this; each square left bracket indicates a further level of embedding:

(1) Dat is een superleuk feest [waarbij er wordt gevierd [dat
That is a superfun feast whereby there gets celebrated that
Sinterklaas (een man uit Spanje) in ons land is [die
St. Nicholas (a man from Spain) in our country is who
onsterfelijk is.]]]
immortal is
"That is a superfun feast which celebrates that Santa Claus (a man from Spain) is in our country who is immortal"

Finite subordination plays a role in various linguistic phenomena, such as long-distance extraction (Ross 1967; Bouma 2017; Schippers and Hoeksema 2021), NEG-raising (Horn 1989; Collins and Postal 2014), long-distance licensing of negative polarity items (Hoeksema 2017) and sequence of tense (Boogaart 1999; Hollebrandse 2000). Consequently, it has been considered one of the core properties of language. While we cannot study these related phenomena in any detail here, we can take a closer look at their common denominator, the presence of finite subordination. Table 8 presents our main findings. Note that we only look at (at least) one level of embedding versus no level of embedding. An ANOVA revealed significant effects of school type ($F(2, 419) = 4.84$, $p = .008$), school year ($F(2, 419) = 17.07$, $p = .000$), and interaction of school type and school year ($F(4, 419) = 3.71$, $p = .006$). For school type there was a significant difference ($p < .05$) between VWO and HAVO.

**Table 8:** Finite embedding per school type and grade.

| School type | Year | *FinEmb* > 0 | *FinEmb* = 0 | *Pct. Finemb* > 0 |
|---|---|---|---|---|
| VMBO | 1 | 121 | 502 | 19.4 |
| | 2 | 234 | 648 | 26.5 |
| | 3 | 210 | 743 | 22.0 |
| HAVO | 1 | 234 | 1099 | 17.6 |
| | 2 | 183 | 764 | 19.3 |
| | 3 | 243 | 720 | 25.2 |
| VWO | 1 | 182 | 796 | 18.6 |
| | 2 | 305 | 879 | 25.8 |
| | 3 | 446 | 1041 | 30.0 |

# 6 Discussion

Our findings bear out the correctness of our hypothesis that variables which show continuous change from elementary school to academic level writing will also differentiate between levels of high school. The degree to which pupils master the demands of academic and professional writing is without doubt important in their academic career, including choice of secondary school level and type of tertiary education. It would therefore be odd if those features which most strongly characterize academic prose were to be randomly scattered across the secondary school essays, rather than clustering around those levels (gymnasium and atheneum) which prepare for university education.

We found that the noun/verb ratio is a reflection of both school type and school year. Higher years and higher school types correspond to a higher noun/verb ratio. Nominal modifiers become relatively more important in higher grades, as we managed to show for attributive adjectives (though not for school types). An increase in attributive prepositional phrase and relative clause usage was also predicted, but could not be established, perhaps owing to the limitations (in size) of the corpus. In Hoeksema, de Glopper, and van Noord (2021) growing amounts of relative clauses were found from elementary school essays all the way to professional academic writing.

Sentential complexity, measured in terms of the percentage of all sentences that involved at least one level of finite embedding, also correlated with higher years and school levels. It is claimed in studies by Biber and his associates that such complexity is not typical of academic prose. The data in Biber and Gray (2010) show that spoken English has more subordinate complement clauses and more adverbial clauses than academic English, and only relative clauses were more prominent in academic than in spoken English. In line with this is a finding of Myhill (2008), a study of writing quality in secondary education, where it was discovered that better writers in that age bracket use significantly less clausal embedding. However, a different conclusion was drawn in Hoeksema, de Glopper, and van Noord (2021) and van Rijt, van den Broek, and Maeyer (2021) for Dutch. While many of the features typical of academic English carry over to Dutch, sentential complexity may well be a factor distinguishing academic English from Dutch, and perhaps, we speculate, from the continental European languages more generally. It should be noted here as well that academic writing styles are not set in stone but may change rapidly, much like any other type of language register, as shown for English by some striking graphs in Biber and Gray (2010). Mean sentence length has declined over time in a variety of English text types, such as fiction and nonfiction (see in particular Rudnicka 2018).

# 7 Conclusions

PaQu and its new component SPOD make it possible to look at a broad range of syntactic phenomena in automatically parsed corpora in a user-friendly way. Corpora can be uploaded and parsed, in order to be queried by SPOD. In this chapter, we probed the possibilities of this application for analysing syntactic variation in the Schrijfmeterscorpus, a collection of essays from different levels and grades of Dutch secondary education. It was shown for a number of syntactic properties associated with academic writing that the writing of students varies in predicted ways across levels and grades, in particular noun/verb ratio, number of nominal modifiers and the percentage of complex sentences.

The use of noun/verb ratios is not standard in studies of writing proficiency, but might be worthwhile considering for future research. There are studies of noun/verb ratios in the typological literature (for example Polinsky and Magyar 2020), but these are focused on types, not tokens. Languages like Dutch have far more nouns in their lexicon than verbs, but token frequency is more balanced, and sensitive to developmental as well as register variation.

# Bibliography

Berlage, Eva. 2014. *Noun phrase complexity in English*. Cambridge: Cambridge University Press.

Biber, Douglas & Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9: 2–20.

Biber, Douglas & Bethany Gray. 2016. Phrasal versus clausal discourse styles: A synchronic grammatical description of academic writing contrasted with other registers. In Douglas Biber & Bethany Gray (eds.), *Grammatical complexity in academic English: Linguistic change in writing*, 67–124. Cambridge: Cambridge University Press.

Bloem, Jelke. 2020. Een corpus waar alle constructies in gevonden zouden moeten kunnen worden? Corpusonderzoek met behulp van automatisch gegenereerde syntactische annotatie. *Nederlandse Taalkunde* 25: 39–71.

Boogaart, Ronny. 1999. *Aspect and temporal ordering. A contrastive analysis of Dutch and English*. Utrecht: Netherlands Graduate School of Linguistics.

Bouma, Gosse. 2017. Finding long-distance dependencies in the Lassy corpus. In Hilke Reckman, Lisa Lai-Shen Cheng, Maarten Hijzelendoorn & Rint Sybesma (eds.), *Crossroads semantics: Computation, experiment and grammar*, 39–56. Amsterdam: Benjamins.

Bouma, Gosse, Gertjan van Noord & Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra & Jakub Zavrel (eds.), *Computational linguistics in the Netherlands 2000*, 45–59. Amsterdam: Rodopi.

Collins, Chris & Paul M. Postal. 2014. *Classical neg raising*. Cambridge, MA: MIT Press.

Glopper, Kees de & Joanneke Prenger. 2013. Schrijfmeters maken. *Zevenentwintigste conferentie onderwijs Nederlands*. Gent: Academia Press.

Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij & Maarten van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Groningen/Deurne: Martinus Nijhoff and Wolters Plantyn.

Heylighen, Francis & Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of science* 7: 293–340.

Hoeksema, Jack. 2017. Neg-raising and long-distance licensing of negative polarity items. In Debra Ziegeler & Zhiming Bao (eds.), *Negation and contact: With special focus on Singapore English*, 33–61. Amsterdam: John Benjamins.

Hoeksema, Jack, Kees de Glopper & Gertjan van Noord. 2021. The development of syntactic structure in written Dutch. Submitted.

Hollebrandse, Bart. 2000. The acquisition of sequence of tense. Ph.D. diss., University of Massachusetts, Amherst.

Horn, Laurence. 1989. *A natural history of negation*. Chicago: University of Chicago Press.

Hunt, Kellogg. 1970. *Syntactic maturity in school children and adults*. Monographs of the Society of Research in Child Development. Chicago: University of Chicago Press.

Myhill, Debra. 2008. Towards a linguistic model of sentence development in writing. *Language and Education* 22 (5): 271–288.

Noord, Gertjan van. 2006. At Last Parsing Is Now Operational. *Taln 2006 Verbum ex machina, Actes de la 13e Conference sur le Traitement Automatique des Langues Naturelles*, 20–42. Leuven.

Noord, Gertjan van, Jack Hoeksema, Peter Kleiweg & Gosse Bouma. 2020. Spod: Syntactic profiler of Dutch. *Computational Linguistics in the Netherlands Journal* 10: 129–145.

Odijk, Jan. 2015. Linguistic research with PaQu. *Computational Linguistics in the Netherlands Journal* 5: 3–14.

Odijk, Jan. 2020. De verleidingen en gevaren van GrETEL. *Nederlandse Taalkunde* 25 (1): 7–37.

Odijk, Jan, Gertjan van Noord, Peter Kleiweg & Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk & Arjan van Hessen (eds.), *Clarin in the Low Countries*, 281–297. London: Ubiquity Press.

Oostdijk, Nelleke. 2002. The design of the spoken Dutch corpus. *New frontiers of corpus research*, 105–112. Amsterdam: Rodopi.

Pander Maat, Henk, Kay Raaijmakers, Dennis Vermeulen & Kees de Glopper. 2019. Tekst-kenmerken en tekstkwaliteit van leerlingteksten. *Tijdschrift voor Taalbeheersing* 41: 331–361.

Pennebaker, James W., Cindy K. Chung, Joey Frazee, Gary M. Lavergne & David I. Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PLoS ONE* 9, no. 12.

Polinsky, Maria & Lilla Magyar. 2020. Headedness and the lexicon: The case of verb-to-noun ratios. *Langages* 5: 1–25.

Pullum, Geoffrey K. & Rodney Huddleston. 2002. Prepositions and preposition phrases. In Rodney Huddleston & Geoffrey K. Pullum (eds.), *The Cambridge grammar of the English language*, 597–662. Cambridge: Cambridge University Press.

Rijt, Jimmy van, Brenda van den Broek & Sven De Maeyer. 2021. Syntactic predictors for text quality in Dutch upper-secondary school students' L1 argumentative writing. *Reading and Writing* 34 (2): 449–465.

Ross, John Robert. 1967. Constraints on variables in syntax. Ph.D. diss., MIT, Cambridge, MA.

Rudnicka, Karolina. 2018. Variation of sentence length across time and genre. In Richard J. Whitt (ed.), *Diachronic corpora, genre, and language change*, 220–240. Amsterdam: John Benjamins.

Sampson, Geoffrey. 2013. The structure of children's writing. In Geoffrey Sampson & Anna Babarczy (eds.), *Grammar without grammaticality: Growth and limits of grammatical precision*, 155–171. Berlin: Walter De Gruyter.

Schippers, Ankelien & Jack Hoeksema. 2021. Langeafstandsverplaatsing in het Nederlands, Engels en Duits: de sandwich ontleed. *Nederlandse Taalkunde* 26: 41–78.

Staples, Shelley, Jesse Egbert, Douglas Biber & Bethany Gray. 2016. Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication* 33: 149–183.

Stowe, Laurie, Robert Hartsuiker, Magdalena Devos & Jack Hoeksema. 2014. Measuring variation in perception of acceptability: a magnitude estimation investigation of Netherlands and Belgian Dutch. In Jack Hoeksema & Dicky Gilbers (eds.), *Black book: a Festschrift in honor of Frans Zwarts*, 311–329. Groningen: University of Groningen.

Wouden, Ton van der, Gosse Bouma, Marjo van Koppen, Frank Landsbergen, Jan Odijk & Matje van de Camp. 2015. Enriching a descriptive grammar with treebank queries. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk & Adam Przepiórkowski (eds.), *Proceedings of the fourteenth international workshop on treebanks and linguistic theories (tlt14)*, 13–25. Warszawa: Polish Academy of Sciences.