LEARNING FROM NATIVES' ERRORS: AN ANALYSIS OF STUDENTS' ERRORS IN WRITTEN LANGUAGE

Annelies Deveneyns, Jose Tummers

Leuven University College (BELGIUM)
annelies.deveneyns@khleuven.be, jose.tummers@khleuven.be

Abstract

There is a growing concern in Flanders about the deterioration of native (written) language proficiency amongst youngsters. In this paper, we will outline a study of the errors in texts written in Dutch by bachelor students. We will pin-point the most acute and most frequent errors in order to develop adapted language material to bridge the gap between the actual and the desired level of proficiency. An error coding scheme was designed that, in line with learner corpus research, combines linguistic information (spelling; lexicon; syntax; textual structure) and error information (erroneous use; omission; redundancy). The most widespread and recurrent errors belong to the categories textual grammar (especially referential coherence), syntax, punctuation and lexical use. Those errors typically cause interpretative problems which interrupt the reading process. The results are the starting point of a usage-based remediation process of the students' written language proficiency by creating a growing awareness of correct formal language use.

Keywords: applied linguistics, learner corpus, language proficiency, writing research

1 INTRODUCTION

There is a growing concern that the **language skills** of youngsters are deteriorating. In Flanders, this concern is translated into an increasing attention towards native and foreign language skills in the policy plans of the Department of Education [1,2]. In higher education, auditing committees and work placement reports increasingly make mention of students being unable to write an error-free text in their mother tongue. As a result, higher education institutions are exploring various pathways to improve the language skills of their students, such as language screening [3,4] and summer classes [5].

The present paper focuses on one particular, but very noticeable aspect of written language proficiency, namely the **errors** occurring in texts. More specifically, we will present a taxonomy of the errors made by first year Flemish university college students in argumentative texts written in their mother tongue Dutch. It will be shown that students' texts contain a high number of various errors. Moreover, the most persistent errors appear in a majority of the texts. Looking at the nature of the errors, it will be shown that most recurrent and persistent errors affect the argumentative and referential structure of the text, disturbing its intelligibility.

The remainder of this paper is structured as follows. First, a brief outline is given of the broader context of language error research within the field of language learning (section 2). Next, the data gathering will be described (section 3). In the subsequent section, we will address the issue of defining and identifying language errors (section 4) before presenting the results (section 5) and discussing the implications for language education (section 6).

2 LANGUAGE ERRORS AND WRITING RESEARCH

There is a **paradox** between the growing realization that there is – or might be – a problem with the native (written) language proficiency of students on the one hand, and the amount of scientific research of (written) language skills of adult natives [6,7] on the other. This is partly a corollary of the communicative approach in language teaching, favoring the development of communicative skills over the formal aspects of language proficiency. The last decades, there is a tendency in written language research to focus on writing processes to the detriment of formal textual properties [8]. Most entry tests measuring written language proficiency mainly use multiple choice questions and fill in exercises. As a consequence, there is need for objective and authentic data about students' formal language proficiency in their mother tongue.

Recently, there is a **renewed attention** for the formal and structural properties of texts in the fields of readability research [9], automatic text evaluation [10,11] and automatic text correction [12]. In foreign language learning, learner corpora are compiled and analyzed to link foreign language proficiency to language properties [13,14]. Those research domains share the interest for the text as a product, focusing on various formal aspects of language proficiency, including errors.

3 DATA GATHERING

This research into students' language errors is part of a project to gather data for descriptive research supporting the development and implementation of a language policy at Leuven University College. The project is confined to written language.[1]

Leuven University College is a Flemish Higher Education Institution organizing 12 bachelor programs. The programs represent 180 ECTS spread over 3 years. At the end, students have reached EQF level 6. Because of the wide range of study programs, ranging from business studies over social work, teacher education and nursery to IT, the population of about 6,500 students reflects the diversity in secondary education[2] better than traditional universities, which mainly attract students from general secondary education.

A sample of 346 texts written by first year students has been collect for the research project. The students were asked during the second semester of year 1 to write a 500 word argumentative text on social network sites. In order to simulate the students' text writing settings as much as possible, participants were given a computer and were allowed to use any information they deemed useful. For the same reason, we opted for an indirect language test in which the participants were not informed about the linguistic goal of the project, thus avoiding students to consider the writing task as a language test and therefore paying extra attention to language and style in comparison to the texts they generally produce when writing papers and essays.[3] The topic, social network sites, belongs to the participants' daily life, minimizing the chance that they would be confronted with an unfamiliar subject.[4]

The respondents were selected by means of cluster sampling [15]: in the first year of every bachelor program, one class group was selected at random to participate in the project. Students had one hour to write an argumentative text about social network sites. In order to minimize the number of absentees and to motivate the students, the hour was integrated in the students' schedule and there was a 1/3 chance to win a movie ticket. This is the assignment as it was presented to the students:

> The government is interested in the impact of social network sites, such as Facebook, MySpace, Netlog …, on social life.
>
> Imagine you are a newspaper journalist who has to write a critical article on the use and the impact of social network sites. Formulate your opinion in a coherent and well-structured text of about 500 words. Convince your audience of your point of view, which can be positive or negative.
>
> You have 60 minutes to formulate your opinion. You can use all resources deemed useful.

Of the original 382 texts gathered, 36 were removed due to an unreadable document format (n=3), an empty document (n=1) and plagiarism (n=32)[5], amounting to a corpus of 346 texts. The average text length is 549.42 words.

The present article is based on a random sample of 90 texts which has been extracted from the dataset. This sample corpus has been manually analyzed to identify and code language errors.

---

[1] This project is funded with grant 020_PWO_TAAL_09 of the Leuven University research fund (PWO).

[2] In Flanders, secondary eduction is attended by youngsters from 12 to 18 years old. It is a 6 year program training graduates to reach level 4 of the EQF. Pupils can choose between general secondary education, technical secondary education, professional secondary education and artistic secondary education.

[3] Since only 2 out of the initial 382 participants explicitly asked whether the assignment was not a language assignment, we consider the distraction a success.

[4] This goal was reached since no significant interdependency was found between the respondents' reported familiarity with the subject and the scores obtained on the four assessment criteria.

[5] We used the software package *Turnitin* to identify plagiarism [16]. Texts were entered in the system which compared them to a world-wide database of texts and to the internet. All documents showing an overlap with existing sources exceeding 12% were removed from the corpus. Although the selection of this threshold is partly random, the overlap identified in the remaining documents involved single words, sequences of words and formulaic expressions.

4 ERROR ANALYSIS

The goal of the error analysis is twofold: the identification as well as the coding of errors occurring in the sample corpus. The results of this analysis will be used to develop teaching materials adapted to the actual language level of university college students and to awake their consciousness about the importance of writing well-structured and error-free texts.

In this section, we will first outline the research questions (section 4.1). Next, we will describe the procedure to identify and code language errors (section 4.2).

4.1 Research questions

The research analysis addresses the following research questions:

1 What are the most frequently occurring language errors?

2 What are the most typical language errors?

The first research question concerns the frequency of occurrence of language errors within the sample corpus. Hence, this research question will be made operational as the **corpus frequency** (CF) of an error. The CF is the number of occurrences of an error in the sample corpus. The second research question deals with distribution of language errors over the texts in the sample corpus, viz. their distribution over the students. Hence, it will be measured by means of the **document frequency** (DF) of an error. This is a density measure computing the number of texts in which an error occurs.

4.2 Error identification and annotation

Although seemingly self-evident, the identification of language errors is by no means a sinecure [17]. The mere definition of an error is often vague and opaque. In foreign language research, the domain most interested in error analysis, an error is often defined by referring to the mother tongue and/or the intuitions of native speakers. Those approaches are not viable for our purpose, since we need a replicable criterion without reference to the mother tongue.

We have chosen for a stepwise approach. In the first stage, we adopted James' broad definition of an error as "an unsuccessful bit of language" [18]. This definition implicitly refers to the breaking of language rules, which raises the issue of what **language norm** to use. Roughly speaking, two language norms can be distinguished: the formal codified norm and the (non-codified) norm of actual language use. The latter is a weakened version of the former, because the former is not strictly applied in real language use. Although the latter can be considered the actual norm of real-life communication, we have opted to use the former, as codified in the *Van Dale Groot Woordenboek der Nederlandse Taal* [19] for spelling and lexicon and the *Algemene Nederlandse Spraakkunst* [20] for grammar and textual structure. This choice for the codified norm is motivated by the need of an objective and replicable norm. The norm of actual language is thus unusable because of its variation amongst situations and language users. Moreover, the texts are written by first year bachelor students who all have finished secondary education and who, according to the learning outcomes of secondary education, ought to be able to write a well-structured and error-free text.

In the second stage, an **error coding scheme** was designed that, in line with learner corpus research, combines linguistic information

- spelling
- punctuation
- use of capital letters
- lexicon
- grammar: syntax, for the errors made at sentence level
- grammar: textual structure, for the errors made at the level of sentence combination
- document structure (material structure of text, including title, paragraphs, etc.)

with error information

- erroneous use

- omission
- redundancy

We will now present the coding scheme. For each linguistic category, we will enumerate and illustrate the different error types. This coding scheme is the result of an incremental adaption during the labor intensive error coding process.

For spelling errors, no subcategories have been created. Table 1 summarizes and illustrates the coding scheme for punctuation errors:

Table 1: Coding scheme punctuation errors

Error type	Code	Example
Missing punctuation	<P_M>	*Mensen die die foto niet moeten zien Ø kunnen dat ook niet* ('people who do not have to see that picture Ø simply cannot')
Redundant punctuation	<P_R>	*Dus gaf ik het, het voordeel van de twijfel* ('that is why I gave it, the benefit of the doubt')
Erroneous punctuation	<P_E>	Use of a comma instead of a full stop at end of a sentence

For errors in the use of capital letters, two categories are distinguished:

Table 2: Coding scheme errors against use capital letters

Error type	Code	Example
Missing capital letter	<C_M>	*Bijna iedereen kent wel een site zoals Facebook, Netlog … veel mensen zijn er ook gebruiker van*. ('Almost everybody knows a site like Facebook, Netlog, … a lot of people are also a user of it.')
Capital letter: error	<C_E>	*Kortom: Sociale netwerksites hebben een grote impact* ('In brief: Social network sites have a big impact')

For lexical errors, eleven categories have been distinguished:

Table 3: Coding scheme lexical errors

Error type	Code	Example
Non-existing words	<L_U>	*ergens vertroeven* instead of *toeven* ('to stay somewhere')
Erroneous use	<L_E>	*een vertrouwelijke site* instead of *betrouwbare* ('a confidential site' instead of 'a reliable site')
Erroneous combination	<L_EC>	*hij is groter als haar* instead of *dan* ('he is bigger than her')
Redundant word	<L_R>	*Persoonlijk vind ik…* ('Personally I think...')
Flemish dialect	<L_F>	*deftig* instead of *fatsoenlijk, net* ('respectable')
Blend	<L_C>	*iets afprinten* instead of *printen* or *afdrukken* ('to print something')
Pleonasm	<L_P>	*alle vrienden die je kent* ('all the friends you know')
Computer & MSN	<L_MSN>	*iemand taggen* ('to tag someone')
Loanwords	<L_B>	*socializen* instead of *optrekken met* ('to socialize')
Pars pro toto	<L_PPT>	*de chat* instead of *de chatroom* ('the chat' instead of 'the chatroom')
Totum pro parte	<L_TTP>	*facebook* instead of *facebookpagina* ('Facebook instead of a 'Facebook page')

Grammatical errors are divided in errors occurring at sentence level, syntactic errors, and those occurring above sentence level, textual errors.

Table 4: Coding scheme grammatical errors

	Error type	Code	Example
Syntactic errors	Word order	<G_WO>	*weten wanneer de posts mogen geplaatst worden* instead of *geplaatst mogen worden* ('to know when the posts can made be public') *iedereen toevoegen als vriend die ze maar tegenkomen* instead of *iedereen die ze maar tegenkomen, toevoegen als vriend* ('to add everyone as friend they meet')
	Valence (direct object, prepositional object, complement clause, …)	<G_V>	*ik kwam Ø tegen en zei …* ('I met Ø and said …') *ik sprak hem aan met die zaak* instead of *over* ('I contacted him with that issue' instead of 'about')

	Auxiliary verbs	<G_AV>	*ik heb hem tegengekomen* instead of *ben* ('I was met him' instead of 'have')
	Tenses	<G_T>	*hij kijkt en riep* instead of *hij keek en riep* or *hij kijkt en roept* ('he looks and yelled')
	Gender	<G_G>	*die moment* instead of *dat* ('that moment')
	Inflection	<G_I>	*een mooie meisje* instead of *mooi* ('a pretty girl') *ons televisie* instead of *onze* ('our television')
	Redundancy	<G_R>	*Terwijl dat de echte personen …* ('while that the real persons …')
	Wrong part of speech	<G_POS>	*hij praat met hun instead* of *hen* ('he speaks to their' instead of 'them')
	Missing words	<G_MW>	*Ik surf op Ø internet instead* of *het* ('I'm surfing on Ø internet' instead of 'the internet')
	Erroneous contraction	<G_EC>	*dat het van een gebruiker uit een sociale netwerksite moet blijven en dus alleen je vrienden toegang geven tot je profiel* ('that it has to remain of a user of a social network site and thus only your friends permit acces to your profile')
Textual errors	Anaphora	<G_T_A>	*Er is een probleem. Dit valt niet te ontkennen.* instead of *Dat* ('There is a problem. This is beyond denial.')
	Erroneous referential links	<G_T_L>	*Een 13-jarige … Zij …* ('A thirteen year old … They …)
	Introduction new referent	<G_T_R>	*die sites* ('those sites', while this is the first mention of (social network sites in the text)
	Coordinating conjunction	<G_T_CC>	(argument 1). *En* (argument 2). ('(argument 1). And (argument 2)')
	Subordinating conjunction	<G_T_SC>	*Hij deed dat doordat ze boos zou worden* instead of *opdat* ('he did it because she would become angry' instead of 'so that')
	Pronominal adverbs	<G_T_PA>	*hij maakte fouten waarbij hij zijn werk verloor.* instead of *waardoor* ('he made mistakes at which he lost his work' instead of 'which caused him to lose his work')
	Conjunctional adverbs	<G_T_AD>	*Die mensen komen niet meer buiten en daarentegen hebben ze geen vrienden meer.* ('Those people no longer leave their house and on the other hand they don't have friends anymore')
	Relative pronouns	<G_T_RP>	*het meisje die daar loopt* instead of *dat* ('the girl which is walking there' instead of 'who')

Finally, table 5 presents an overview of the error types identified at the level of the material **document structure**.

Table 5: Coding scheme errors document structure

Error type	Code	Example
Missing title	<D_T>	Document without a title
Use of paragraphs	<D_P>	Document without paragraphs or document where every sentence is a separate paragraph
Macro-structure	<D_MS>	Documents without a clearly articulated structure "introduction > body > conclusion"
Style: colloquial language	<D_S_C>	Euhm ik bedoel dus … ('Well in fact I mean that …')
Style: use of block letters	<D_S_B>	*Ik geloof dat het wel tof is om als fan de activiteiten van je idool te kunnen volgen!* ('I do think that it is nice to be able to follow the activities of your idol!') *Maar ze zouden standaard op ZEER privé moeten staan*. ('But their default setting should be VERY private.')
Style: stream of consciousness	<D_S_S>	*Wij hebben de opdracht gekregen om een kritisch artikel te schrijven over sociale netwerksites.* ('We got the assignment to write a critical article on social network sites.')
Style: quotes	<D_S_Q>	*Zoals Einstein zei ";" voor elke reactie* … ('As Einstein said: for each reaction …')
Style: (useless) repetitions	<D_S_R>	*Een ander minpunt is dat* …('Another disadvantage is that …' in a text where every disadvantage is introduced by this exact sequence)
Style: too formal	<D_S_F>	*Dit is ook mede het gevolg van* … ('This is partway a consequence of …')

5 RESULTS

The presentation of the results will be broken up in two parts: we will start with a global overview of the errors made; then, we will proceed with a detailed analysis of the different error types occurring in the linguistic categories we have distinguished. This section will be concluded by presenting a usage-based error typology based on the combined CF and DF scores of each error type.

5.1 Global analysis

In the sample corpus of 90 texts, 6,592 errors have been identified. This amounts to 13.97 errors per 100 words of text (SD = 3.98). Although we have used the strict codified language norm, this still remains a very high error rate for students who have all obtained a secondary education degree. In this section, we will describe the distribution of the values of CF and DF over the texts in de sample corpus and over the different error types. The CF is a standardized score identifying the occurrences of an error per 100 words. The CF of an error is computed based on exclusively the documents in which the error effectively occurs, since the documents in which an error does not occur significantly decrease the CF score (95% $CI_{difference}$ = [0.02;0.29]; t = 2.2019, df = 107.048, p = 0.02982). The DF identifies the proportion of texts in which the error occurs.

We will first look at the distribution of the CF and DF scores over the texts, viz. the students, in the sample corpus. The left-hand plot in figure 1 shows that the number of errors per 100 words ranges between 5.49 and 29.59 (outliers included) with 13.53 as median. Since the average sentence length is 17 words, those data indicate that for almost half of the bachelor students in the sample corpus, it is taxing to write an error-free sentence.

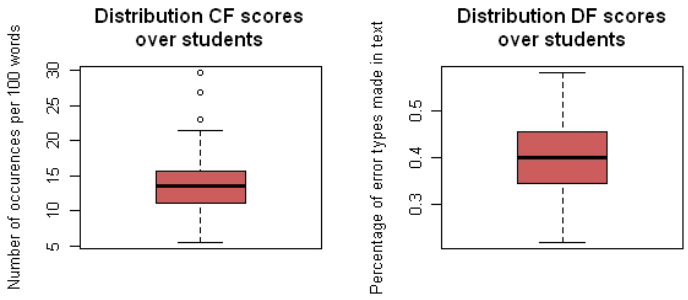

Figure 1: Distribution CF and DF over texts in the sample corpus

When we look at the distribution of the DF values in the right-hand plot, we see that the average text contains a great diversity of errors. Half of the students make more than 40% of the errors listed in the taxonomy presented in section 4.2. Finally, the high correlation score between the CF and DF scores (r = 0.8378) indicates that the most frequently occurring errors are those recurring in most texts. In other words, there appears to be a subset of highly frequent and recurrent errors which will be identified in section 5.2.

Figure 2 shows the distribution of the CF and DF values over the error types.

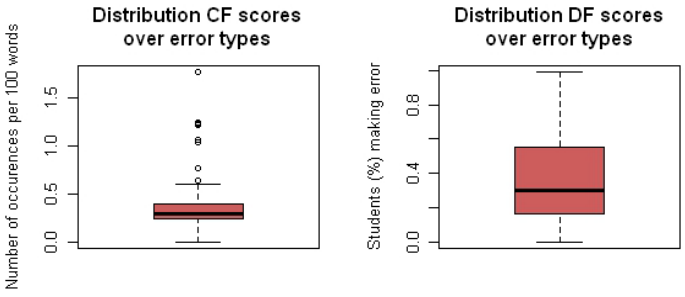

Figure 2: Distribution CF and DF over error types in the sample corpus

The distribution of the CF values in the left-hand boxplot shows that the overwhelming majority of errors has a relative low frequency of occurrence. After exclusion of the outliers, all errors occur less than 0.5 times per 100 words. Meanwhile, the outliers suggest the existence of a small number of highly recurrent errors. The right-hand plot visualizes the distribution of the DF values over the error types. Although about three quarters of the errors occur in less than half of the texts, the boxplot

clearly shows that a majority of errors occur in at least one out of three texts. Putting both plots together, we can conclude that the relatively high number of errors per 100 words of text results from a small number of highly frequent errors recurring in a majority of the texts on the one hand and a great number of infrequent errors recurring in a lot of texts on the other hand.

5.2 Analysis per linguistic category

We will start the detailed analysis of the error types listed in the taxonomy by presenting the global CF and DF scores for the **linguistic categories**:

Table 6: CF and DF per linguistic component

Error type	DF	CF
Spelling	0.90	0.63
Punctuation	0.98	2.21
Use of capital letters	0.24	0.25
Lexicon	1.00	2.85
Grammar: syntax	1.00	2.97
Grammar: textual structure	1.00	3.98
Document structure	1.00	0.95

These results clearly show that the most problematic domains are punctuation, lexical use and grammar. In spite of the text being written on a computer with a spelling checker (MS Word 2007), a majority of the students still makes spelling errors. Their document frequency, however, is relatively low, meaning that most students make 2 to 3 spelling errors in a 500-word text.[6]

In the remainder of this section, we will focus on the aforementioned error types with a high score for both CF and DF, viz. the error types that are recurrently committed by a majority of students. Table 7 contains the CF and DF scores for punctuation. Those figures show that more than 9 out of 10 students repeatedly struggle with missing and erroneous punctuation.

Table 7: CF and DF punctuation error types

Error type	DF	CF
Missing punctuation	0.93	1.21
Erroneous punctuation	0.93	1.06
Redundant punctuation	0.21	0.36

In table 6, we have seen that every student in the sample corpus makes on average 2.85 lexical errors per 100 words. When we look at table 8, it appears that a limited number of lexical error types have both a high frequency of occurrence and a high recurrence amongst texts, namely erroneous word use, redundant use of words and *pars pro toto*. It must be noted that the high density of the final category in the sample corpus might be a corollary of the topic of the writing assignment since most occurrences involve *mijn Facebook* ('my Facebook' instead of 'my Facebook page').

Table 8: CF and DF lexical error types

Error type	DF	CF
Non-existing words	0.47	0.34
Erroneous use	0.98	1.76
Erroneous combination	0.48	0.25
Redundant word	0.56	0.47
Flemish dialect	0.16	0.27
Blend	0.05	0.26
Pleonasm	0.18	0.24
Computer & MSN	0.33	0.37
Loanwords	0.30	0.33
Pars pro toto	0.54	0.38
Totum pro parte	0.04	0.17

[6] This finding that spelling is not the most pinching problem might wonder laymen but concurs with language teachers' intuitions.

The discussion of the grammatical errors will be divided in two parts: those on sentence level (syntax) and those above sentence level (textual structure). Table 9 presents an overview of the syntactic errors.

Table 9: CF and DF syntactic error types

Error type	DF	CF
Word order	0.92	0.42
Valence	0.17	0.18
Auxiliary verbs	0.31	0.20
Tenses	0.43	0.37
Gender	0.03	0.35
Inflection	0.90	0.46
Redundancy	0.82	0.60
Wrong part of speech	0.37	0.34
Missing words	0.94	0.76
Erroneous contraction	0.52	0.27

In these data, it appears that a majority of the students faces problems with word order, inflection, use of redundant elements and unfinished or incomplete sentences (viz. sentences missing words).

Let us finally look at the textual errors made by the students in our sample.

Table 10: CF and DF textual error types

Error type	DF	CF
Anaphora	0.96	1.24
Erroneous referential links	0.98	1.04
Introduction new referent	0.98	1.22
Coordinating conjunction	0.73	0.47
Subordinating conjunction	0.26	0.23
Pronominal adverbs	0.19	0.22
Conjunctional adverbs	0.28	0.29
Relative pronouns	0.28	0.24

Looking at the first three figures (above the double line) which all relate to the referential structure of the text, viz. how new ideas are introduced and imbedded in the existing conceptual structure of the text, we have to conclude that an overwhelming majority of the students face severe problems when having to build a referentially coherent and transparent text. For the reader, this means that his/her reading process will be repeatedly interrupted and that it is hard to find out which referent the writer is speaking of. Those referential errors disturb a smooth and transparent understanding of the message conveyed by the text. The second part of the table (below the double line) shows the results for the argumentative structure of the text, i.e. how the ideas conveyed by sentences are related. At first sight, it might seem remarkable that the category causing most problems is the coordinative conjunctions, which are the least elaborate set of conjunctive elements and hence expected to cause little problems. A closer look at the data, however, reveals that the apparently low occurrence and recurrence of the other error types for conjunctive elements is due to their very low overall frequency in the corpus.

We will conclude this analysis by concretizing the high correlation coefficient between CF and DF (see section 5.1). This boils down to answering the following questions: (i) what are the low frequency errors made by few students, and (ii) what are the highly frequent errors recurring in the texts of a majority of students? Table 11 combines the CF and DF score of every error code, creating a two-dimensional usage-based **error taxonomy**. Both scores are split up in four ordinal categories: for the DF (in essence a proportion), its range is cut into four parts ($DF < 0.25$; $0.25 \leq DF < 0.50$; $0.50 \leq DF < 0.75$; $DF \geq 0.75$); for the CF, which has a skewed distribution, as can be deduced from the left-hand boxplot in figure 1, we split up the data in four parts based on the deciles (CF < decile 5; decile 5 ≤ CF < decile 8; decile 8 ≤ CF < decile 9; CF ≥ decile 9). The goal of this decile-based breakdown is to identify the error types with a very high CF score. Due to limited space, the error types are identified by their codes, as mentioned in tables 1 to 5.

Table 11: Combination of CF and DF scores of error types

	DF < 0.25	**0.25 ≤ DF < 0.50**	**0.50 ≤ DF < 0.75**	**DF ≥ 0.75**
CF < decile 5	<C_M> <L_PPT> <G_V> <G_T_A> <G_T_PA> <D_S_B> <D_S_Q> <D_T> <D_P> <D_MS>	<G_AV> <G_T_SC>		
Decile 5 ≤ CF < Decile 8	<C_E> <L_F> <L_C> <L_P> <G_T_RP> <P_R>	<L_EC> <L_B> <L_U> <L_MSN> <G_T> <G_G> <G_POS> <D_S_C> <D_S_F>	<L_R> <L_TTP> <G_T_CC> <G_EC> <D_S_R>	
Decile 8 ≤ CF < Decile 9				<S> <G_WO> <G_I> <G_R> <G_MW> <D_S_S>
CF ≥ Decile 9				<P_M> <P_E> <L_E> <G_T_A> <G_T_L> <G_T_R>

The error tags in table 11 mainly occupy the diagonal from to the top left (viz. low frequent error types made by a minority of students) to the bottom right (viz. highly frequent error types made by a majority of students), confirming the high positive correlation coefficient reported in section 5.1. We clearly see that the majority of error types is situated in the top left-hand quadrant of the error space. In the bottom right-hand quadrant, we find a limited number of highly frequent and recurrent error types, belonging to syntax, textual grammar, lexical use and punctuation.

6 DISCUSSION

The data presented in this paper corroborate the general social concern that it is hard for youngsters, viz. first year bachelor students, to write well-structured and error-free texts. The results can be used to develop a usage-based language policy and teaching materials dealing with the actual problems faced by students. Firstly, the texts contain a large number of errors which instantiate a great variety of error types. This finding argues for language classes reviewing the principles of spelling and grammar. Those language classes should focus on the formal aspects of written language and textual structure as well as the meta-language necessary to explain and understand the problems at hand. Next to those knowledge-based aspects, students have to be made aware of the mistakes they make and the possible impact on the understanding of the text. Those classes are ideally organized for all students in the first year of higher education. Secondly, there is a positive correlation between the recurrence of an error and its document frequency. In view of remediation, this means that within the large variety of errors, a relatively small number of errors is committed by most of the students and hence needs extra attention during remediation.

Let us now turn to the types of errors made. The highly frequent and recurrent errors instantiate so-called 'markers', to be more precise referential, syntactic, lexical and punctuation markers. Those elements play a central role in the conceptualization and construction of a transparent and well-structured message. Their role is to facilitate the understanding of the message conveyed by the text by articulating the relations between the different referents and actions. As a result, those errors disturb the textual understanding [9], not only affecting the textual evaluation [21], but also the persuasiveness of the message [22] and even the author's image [23]. Spelling errors, which have a low degree of occurrence in many texts have a smaller impact on text evaluation, since they hardly impede the reading process.

At a general level, the results of the error analysis show the need to focus on the formal aspects of language. This is especially the case when it concerns proficiency in the mother tongue, since society expects its members, the more so when they have obtained a higher education degree, to be able to write transparent and error-free texts in their mother tongue. This claim is by no means a nostalgic plea to return to language drills reducing language to its formal manifestation. We do not argue for an abolition of the communicative approach, which made a very important contribution to the improvement of language skills from a pragmatic point of view: students have become proficient communicators able to adapt to various situations. However, this approach needs to be complemented with a formal component that has a twofold goal. First of all, students' attention has to be drawn to the formal and structural properties of their (written) language production. They have to learn to write error-free and well-structured texts. Second, this knowledge-based goal can only be

achieved when it is linked to an attitudinal goal. Students have to be made aware of the importance of well-structured and error-free texts, not only in education but also in their future career. Those goals can only be reached successfully when the complete teaching staff confronts students with the errors they write in texts and with the effect those have on the readability of the text and the way those errors are perceived by the reader.

REFERENCES

[1] Vandenbroucke, F. (2007). *Gelijke kansen op de hele onderwijsladder Een tienkamp. Beleidsbrief onderwijs en vorming 2007-2008*. Url: www.ond.vlaanderen.be/beleid/brief/2007-2008.pdf.

[2] Smet, P. (2011). *Taalnota: Samen taalgrenzen verleggen*. Url: www.ond.vlaanderen.be/nieuws/2011/doc/Talennota_2011.pdf.

[3] Deygers, B. & Kanobana, S. (2010). Taaltoetsen: waarom, wat en hoe? In E. Peeters & T. Van Houtven (eds.) *Taalbeleid in het hoger onderwijs: de hype voorbij?* Antwerpen: Acco. 23-36.

[4] De Wachter, L. & J. Heeren (2011). *Taalvaardig aan de start. Een behoefteanalyse rond taalproblemen en remediëring van eerstejaarsstudenten aan de KU Leuven.* Interfacultair Instituut voor Levende Talen (ILT)/Katholieke Universiteit Leuven.

[5] Sterckx, L. & D. Vanhoren (2011). *ToTaal Klaar? Startklaar! Een instapcursus 'Nederlands in het hoger onderwijs': tool in studiebegeleiding en taalondersteuning*. Leuven: KHLeuven.

[6] Berckmoes, D. & Rombouts, H. (2009). *Intern rapport verkennend onderzoek naar knelpunten taalvaardigheid in het hoger onderwijs*. Antwerpen: Universiteit Antwerpen/Linguapolis.

[7] Peeters, E. & Van Houtven, T. (eds.) (2010). *Taalbeleid in het hoger onderwijs: de hype voorbij?* Antwerpen: Acco.

[8] Hyland, K. (2002). *Teaching and Research Writing*. London: Longman.

[9] Pittler, E. & A. Nenkova. (2008). "Revisiting Readability: A Unified Framework for Predicting Text Quality". *Proceedings of the Empirical Methods in Natural Language Processing*.

[10] Crosley, S.A. & D.S. McNamara. 2011. "Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing". *International Journal of Continuing Engineering Education and Life-Long Learning* 21(2/3). 170-191.

[11] Yannakoudakis, H. T. Briscoe & B. Medlock (2011) "A New Dataset and Method for Automatically Grading ESOL Texts". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. 180-189.

[12] Albert, C., M. Garnier, A. Rykner & P. Saint-Dizier (2009). "Analyzing a corpus of documents written in English by native speakers of French: Classifying and annotating lexical and grammatical errors". Paper presented at the Corpus Linguistics Conference, Liverpool, July 2009.

[13] Granger, S. (2003). "The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research". *TESOL Quarterly* 37(3). 538-546.

[14] Pravec, N. (2002). "Survey of learner corpora". *ICAME Journal* 26. 81-114.

[15] McDaniel, C. & Gates, R. (2007). *Marketing Research*. Hoboken, N.J.: Wiley.

[16] iParadigms. (2005). *TurnItIn Instructor Guide*, Url: http://www.turnitin.com/static/training_support/tii_instructor_guide.pdf.

[17] Tono, Y. (2003). "Learner corpora: design, development and applications". Paper presented at the Corpus Linguistics 2003 Conference (CL 2003), Lancaster.

[18] James, C. 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. London, New York: Longman.

[19] Boon, T. de & D. Geeraerts. (2008). *Van Dale Groot Woordenboek van de Nederlandse Taal*. (14th edition). Utrecht: Van Dale Lexicografie.

[20] Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (1997) *Algemene Nederlandse Spraakkunst*. Groningen/Deurne, Martinus Nijhoff uitgevers/Wolters Plantyn.

[21] Kloet, L., J. Renkema & C. van Wijk (2003). "Waarom foutloos schrijven? Het effect van taalfouten op tekstwaardering, imago en overtuigingskracht". In L. van Waes, P. Cuvelier, G. Jacobs & I. de Ridder (eds.) *Studies in Taalbeheersing.* Assen: Koninklijke Van Gorcum. 270-279.

[22] McCroskey, J. C. & R. S. Mehrley (1969). "The effects of disorganization and nonfluency on attitude change and source credibility." *Speech Monographs* 36. 13-21.

[23] Burgoon, M. & G. Miller (1985). "An expectancy interpretation of language and persuasion." In H. Giles & R. Clair (eds.) *The social and psychological contexts of language*. London: Erlbaum. 199–229.