UPPSALA UNIVERSITY

Project ref. no. *LE3-4239*

Project title *SCARRIE Scandinavian Proof-reading Tools*

Deliverable number *DEL 2.1.3.2*

Deliverable title ***An Error Database of Swedish***

Number of pages *54*

WP/Task responsible *Anna Sågvall Hein, Department of Linguistics, Uppsala University, Box 513, S-751 20 Uppsala, Sweden*
Email: anna@ling.uu.se

Author(s) *Olga Wedbjer Rambell, Bengt Dahlqvist, Erik Tjong Kim Sang, Nils Hein*

EC Project Officer *Pierre-Paul Sondag*

Keywords *Error Corpora Database, error database, error corpus, proof-reading, Swedish*

Abstract Two Swedish newspapers have supplied non-proof-read and proof-read versions of articles, from which language errors have been collected. Erroneous and corrected versions of text fragments are stored together with information of their origin. An error type code has been assigned to each error according to the error typology (DEL 2.1). Error frequencies are presented in the report.

The database has an SQL interface. Information can also be retrieved by a webbrowser through a www-based search interface developed for the database.

Executive Summary

The Error Corpora Database (ECD) has been constructed for the purpose of storing and making searchable instances of language errors and their corrections together with information of their origin and error types. It is to be used in the development of a proof-reading tool for Danish, Norwegian, and Swedish in the SCARRIE project. In specific, it will serve as a source of information about error types and their frequencies. The ECD seems to provide a very good starting point for the work with the grammar checker. It will also be a useful source of information for the word checking module.

The ECD consists of two components, a relational database for data storage and a www-interface for searching. The interface consists of a number of forms written in HTML. From the webpages, information is retrieved by sending SQL queries to the database and presenting the received results.

The Swedish language errors have been provided by two newspapers, Svenska Dagbladet (SvD) and Upsala Nya Tidning (UNT). Non-proof-read and proof-read texts were supplied from SvD in electronic form. The two versions were processed automatically to find the sentences which had been altered by the proof-reader. The SvD corpus contains exactly 2,100 errors. During the winter/spring of 1997, UNT supplied the department with the proof-readers' paper copies on which they had marked the corrections to be made. The corpus covers the sections normally proof-read at the newspaper. From this supply, 25 days' production was analysed and 6,801 errors were found.

Each language error is stored as a text fragment in the database along with the corrected version of the fragment. Information about newspaper, publishing date, text section (e.g. domestic news, editorial) and text type (e.g. headline, plain text) is also given. An error type code is attached to each entry according to an error typology.

The interface is meant for the casual user, but also for persons who are familiar with the error typology. The user is guided through the error codes, but is also given the opportunity to perform free text searches. It is possible to search in only one of the newpaper corpora, as well as in only one of the text types. For the revision of the database, the possibility of making changes by using the www-based search interface was added. The relational database with its interface has been an efficient tool for working with the corpora. The www-based search interface has functioned very well.

The Error Corpora Database has become larger, containing more error entries, than was initially planned. The size of the database, nearly 9,000 error entries, was found to be necessary for reaching reliable conclusions about error frequencies. The results are in that respect dependent on the size and the quality of the corpora.

Spelling errors are the most common cause for proof-readers' corrections. In total, over 40% of the errors in the ECD are spelling errors. Grammar problems, punctuation problems, and graphical problems are equally common, approximately 16% each. Graphical problems is the least common error group, less than 10%.

In the UNT corpus, hyphenation errors account for 40% of the spelling errors, and ordinary spelling errors are more common than word formation errors. The SvD corpus has no end of line hyphenation errors because the material was delivered as raw text. In the SvD corpus, word formation errors account for nearly half of the spelling errors and are thus more common than ordinary spelling errors.

On the whole, the distributions of grammar problems in the two newspapers are quite similar. Four grammar problems out of ten are located in the noun phrase. Agreement errors are the most common error type, especially number agreement problems are frequent accounting for nearly half of the agreement errors. The second most common noun phrase problem concerns species, embracing problems with article usage and usage of the definite and indefinite forms, followed by case problems.

One grammar problem out of four concerns the verb, either in a combination of verbs (i.e. in a verb phrase in a limited sense) or as a verb valency error. Most verb valency errors occur in infinitive phrases, over 90% of these errors concern missing infinitive mark *att*, in particular after the verb *komma*.

Problems with prepositions are distributed over four categories depending on what governs the prepositions: a preceeding noun, adjective or verb, or other factors. The prepositional phrase category gathers the latter ones, being the third largest category. Together, preposition related problems account for 15–20% of the total number of grammar problems.

The least common grammar problems occur in adjective phrases and adverb phrases on the top level of the clauses. Wrong word category problems on clause or sentence levels are also quite uncommon, and so are pronoun case errors.

By far the most frequent punctuation problem concerns the comma, over 70% in both newspapers. The most frequent graphical problem, in both newspapers, is related to spaces.

The most frequent style, meaning, and reference problem is preferred spelling, i.e. choice between correct word forms. Choice between correct abbreviations is also a common problem; points are often removed from abbreviations. How to present numbers is also a quite common problem. Changing words and expressions has been done to a larger degree in the UNT corpus than in the SvD corpus.

Uppsala university
Department of Linguistics
SCARRIE
21 January 1998

An Error Database of Swedish

Olga Wedbjer Rambell
Bengt Dahlqvist
Erik Tjong Kim Sang
Nils Hein

SCARRIE

DEL 2.1.3.2

FINAL VERSION 1.0

Contents

Tables

Figures

Acknowledgements

Without the contributions from Svenska Dagbladet and Upsala Nya tidning, the Error Corpora Database would not exist in its present form. Thanks to these newspapers, their reporters and proof-readers the error collection was made possible.

Students' in language engineering at the Department of Linguistics have also contributed to the work with the Error Corpora Database: Kristina Bäckström and Stina Nylander have classified language errors, Ulrika Hedström, Per Sandhammar, Anna Sterner and Natalia Zinovjeva have created the input files and helped with the inspection of the database. Law student Sandra Hein has worked with the classification of errors, and Gunilla Fredriksson, personnel and financial officer at the Department of Linguistics, and Anna Koch, student in linguistics, have helped with the inspection of the database.

1 Introduction

The Error Corpora Database (ECD) has been constructed for the purpose of storing and making searchable instances of language errors and their corrections together with information of their origin and error types. It is to be used in the development of a proof-reading tool for Danish, Norwegian, and Swedish in the SCARRIE project. In specific, it will serve as a source of information about error types and their frequencies. The ECD is a relational database with an SQL interface, and it can be reached on the Internet[1].

The Swedish language errors have been provided by two newspapers, Svenska Dagbladet and Upsala Nya Tidning. Each language error is stored as a text fragment in the database along with the corrected version of the fragment. Information about newspaper, publishing date, text section (e.g. domestic news, editorial) and text type (e.g. headline, plain text) is also given. An error type code is attached to each entry according to an error typology.

The error typology is a part of work package 2 of the SCARRIE project which is funded by the Language Engineering Sector in the Telematics Application Programme of the European Union.[2] The SCARRIE consortium consists of a co-ordinating partner, four project partners, and nine sub-contractors. Center for Sprogteknologi in Copenhagen will develop the Danish part of the SCARRIE pilot application, Humanistik Datasenter in Bergen will develop the Norwegian part and the Department of Linguistics at Uppsala university will develop the Swedish part. One of the subcontractors, Stichting Cognitieve Technologie, has already developed a proof-reading tool for Dutch that will be used in the SCARRIE project.

Newspapers and publishing houses in Sweden (Svenska Dagbladet, Upsala Nya Tidning), Norway (Bergen Trykk AS), and Denmark (Berlingske Tidende, Munksgaard International Publishers) have contributed to the project by defining user demands on an automated proof-reading tool. They are also the main suppliers of text material for the dictionaries and the error corpora. In the final phase, these users will act as test beds for the SCARRIE proof-reading software.

After the project, the co-ordinating partner of the project, WordFinder Software, will package the SCARRIE results into its own interface, and market it as a product. The ultimate goal for WordFinder Software is to develop a proof-reading tool for everyone using a word processor when they write in Swedish, Danish or Norwegian.

In the next chapter, a technical description of the Error Corpora Database is presented. The interface used over the Internet is described in chapter 3. The collection of language errors is accounted for in chapter 4, and in the following chapter, statistical information is presented about the different error types. The last chapter contains closing remarks about the error corpora, the work done, and remaining work.

---

[1] http://stp.ling.uu.se/cgi-bin/w3-msql/ECD/index.html
User name: user
Password: error

[2] More information about the SCARRIE project can be found on the Internet:
http://www2.echo.lu/langeng/en/le3/scarrie/scarrie.html
http://www.scarrie.com

2 Technical Specification

As a first step, the design of the database was outlined from a discussion about what information would be needed and retrievable in order to fulfil the purpose of the error database as a source of information about error types and their frequencies.

2.1 Design of the Error Corpora Database

The information to be stored in the database is, beside the erroneous and corrected text fragments and their error type codes, information about their origin: newspaper or publishing house, section of the newspaper, text type, and publishing date. To locate the error to its specific page in the newpaper would be difficult, especially when a corpus is supplied in electronic format. Since the database is to be used by all partners of the SCARRIE project, information about responsible organisation or partner regarding the entries could be useful. An internal sequence number would also be needed, and perhaps also a document number to connect the error to the newspapers' archives. This information would be gathered in Table1 (see table 2.1 and tabel 2.2).

Table 2.1 Design of Table1

SEQ NO	numerical value, 1,2...n [internal usage only]
ORG ID	alphanumerical code, 4 chars, organisation
USER ID	alphanumerical code, 4 chars, paper or publishing house
PUBL DATE	alphanumerical code, 8 chars, date of publication
TEXT SECTION	alphanumerical code, 14 chars, text category (e.g. culture, politics)
TEXT TYPE	alphanumerical code, 11 chars, text type (e.g. headline, plain text)
ERROR TYPE CODE	alphanumerical code, 8 chars (fixed pos for cat and sub cat)
ERROR TOKEN	full sentence or sentence fragment (max 400 chars)
CORRECTED TOKEN	full sentence or sentence fragment (max 400 chars)
COMMENT	free text (max 400 chars)

Table 2.2 An example of a Table1 entry

ORG ID	UU
USER ID	SVD
PUBL DATE	19950214
TEXT SECTION	KULTUR
TEXT TYPE	TEXT
ERROR TYPE CODE	GPNPNN01
ERROR TOKEN	Även om Alfred Nobel bara bodde här i två är det så oerhört mycket som påminner om hans innehållsrika liv från födseln 1833 till bortgången 1896, säger Tina Svanberg-Lundgren.
CORRECTED TOKEN	Även om Alfred Nobel bara bodde här i två år är det så oerhört mycket som påminner om hans innehållsrika liv från födseln 1833 till bortgången 1896, säger Tina Svanberg-Lundgren.
COMMENT	none

A table would also be needed for storage and retrieval of the error type codes and explanations of them. For this purpose, Table2 was designed (see table 2.3).

Table 2.3 Design of Table2

ERROR TYPE CODE	alphanumerical code (max 8 chars)
ERROR DESCRIPTION	free text (max 120 chars)

2.2 Implementation of the Error Corpora Database

The ECD consists of two components, a relational database for data storage and a www-interface for searching. The database is implemented in mSQL 2.0.1, a relational database management system available from Hughes Technologies[3]. The relational database contains two separate tables, related by the error type code. The actual table structures used in the ECD is showed in table 2.4 and 2.5. One field has been added, a utility value which will be used when analysing errors and sorting out those errors that can be handled by the parser (work package 6, task 6.2.3).

Table 2.4 Implementation of Table1

Field name	**Size**	**Data type**	**Description**
seqno[4]	4	INT	Internal entry no
orgid	4	CHAR	Organisation, e.g. UU
userid	4	CHAR	Paper or publishing house, e.g. SvD
lang[4]	2	CHAR	Language
publdate	8	CHAR	Publication Date
docno[4]	8	CHAR	For future use
section	14	CHAR	Section in the paper, e.g. culture, politics
texttype	11	CHAR	Type of text, e.g. headline, plain text
utilvalue[4]	4	INT	For future use: Evaluation of error recognition capability
errtypecode	8	CHAR	Error type code
errtoken	400	CHAR	Error token
corrtoken	400	CHAR	Corrected token
comment	400	CHAR	Free text

Table 2.5 Implementation of Table2

Field name	**Size**	**Data type**	**Description**
errtypecode	8	CHAR	Error type code
errdescr	120	CHAR	Error description

The ECD www-interface is written in HTML, with certain parts coded in JavaScript and mSQL's own scripting language W3-mSQL. By means of the scripting language, it is possible to include SQL queries to the database in the HTML code of the interface and thus enable on-line communication between the webpages and the database.

[3] The mSQL system has an official home page at the URL http://www.Hughes.com.au

[4] Field not displayed in the ECD www-interface.

The www-interface of the ECD system consists of 17 separate HTML-files that are invoked in turn depending on the user's actions and choices. The communication with the database is primarily done via embedded SQL queries in the HTML-code with key values generated on-line by the system. The result is then used to build an HTML page displaying the selected entries from the database query. Since the W3-mSQL module handles the communication to and from the mSQL database, the need for a multitude of standard CGI scripts for every web page with dynamic content is removed.

Further, the module W3-auth in W3-mSQL was utilised for password protection of the ECD web-pages.

2.3 Input Data File Format

To manually insert data, case for case, into the ECD database, would be both time-consuming and inefficient. Therefore, data to be inserted into the Table1 in the ECD should be delivered in batch, i.e. as a text file with all the error data given in a standardised form. By an automatic procedure, all cases in the batch file were then inserted into the ECD database. Table 2.6 shows in what format information should be given, a template that will be referred to as the input data file format. The data for Table1 should be delivered as a text document in ASCII, where each error entry consists of four or five fields, each starting with a numeral in column one, identifying the field. An example of an error instance written in the input data file format is given in table 2.7. The data for Table2 was inserted by direct use of mSQL.

Table 2.6 Input data file format for Table1

Field no	Contents
1	Error token (can be several lines)
2	Corrected token (can be several lines)
3	Error type code
4	org/user id/publ date/section/text type
5	Comment (optional)
and a blank line as a separator.	

Table 2.7 Example of an error instance in the input data file format for Table1

```
1 Även om Alfred Nobel bara bodde här i två är det så oerhört mycket som påminner
1 om hans innehållsrika liv från födseln 1833 till bortgången 1896, säger Tina
1 Svanberg-Lundgren.
2 Även om Alfred Nobel bara bodde här i två år är det så oerhört mycket som påminner
2 om hans innehållsrika liv från födseln 1833 till bortgången 1896, säger Tina
2 Svanberg-Lundgren.
3 GPNPNN01
4 UU/SvD/19950214/KULTUR/TEXT
5 none
```

3 Search Interface

For easy retrieval of information, a www-based search interface has been developed. The interface is meant for the casual user, but also for persons who are familiar with the error typology. The user is guided through the error codes, but is also given the opportunity to perform free text searches. It is possible to search in only one of the newpaper corpora, as well as in only one of the text types. For the revision of the database, the possibility of making changes by using the www-based search interface was added.

The interface consists of a number of forms written in HTML. From the webpages, information is retrieved by sending SQL queries to the database and presenting the received results.

The Error Corpora Database is found at:

http://stp.ling.uu.se/cgi-bin/w3-msql/ECD/index.html

User name: user

Password: error

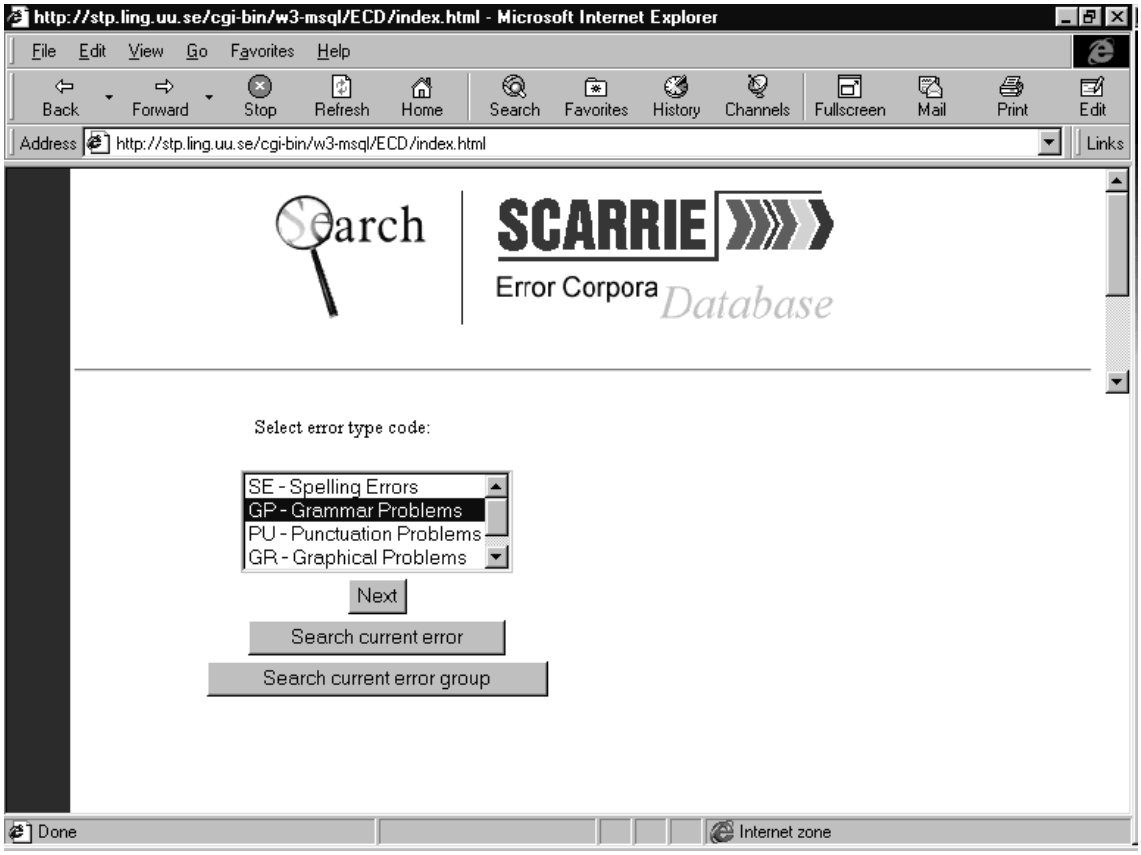

Figure 3.1 Start page at the www-based search interface

The user is met by the start page shown in figure 3.1. The error typology consists of five error group, and the user marks one of them. The button ”Next” takes the user one step

down in the error typology hierarchy to a similar page. When the user marks one of the error codes, the description of the error is shown at the right, see figure 3.2.

Figure 3.2 The error type codes are explained as the user marks them

Both buttons ”Search current error” and ”Search current error group” lead to the search page, as does the Search-picture at the top of the page. The search page is shown in figure 3.3 after `GPNP` has been choosen and the ”Search current error” button has been pushed.

Beside the error type code, the user can chose User ID, i.e. to search in all corpora, or only in the SvD corpus or in the UNT corpus. The user is also able to specify the text type: caption (`BILDTEXT`), by-line (`BYLINE`), introduction (`INGRESS`), headline (`RUBRIK`), plain text (`TEXT`), subheading (`UNDERRUBRIK`), or all text types (`ALL`). There is also a possibility to search for words and parts of words in the erroneous sentences. Note that the search engine is case sensitive.

The search in table1 in the database is started by the ”Search” button. The result is presented as a table, showing 250 errors at the time (see figure 3.4). At the bottom of the table, the user will find a button ”Show next 250 hits”.

At the top of the result page, there is a button ”Show current error descriptions ”. When the button is pressed, a new window will appear in which (parts of) the typology is presented. For instance, if the search was made for the error code `GPNP`, the codes `GP`, `GPNP`, and all codes beginning with `GPNP` are fetched from table2 of the database.

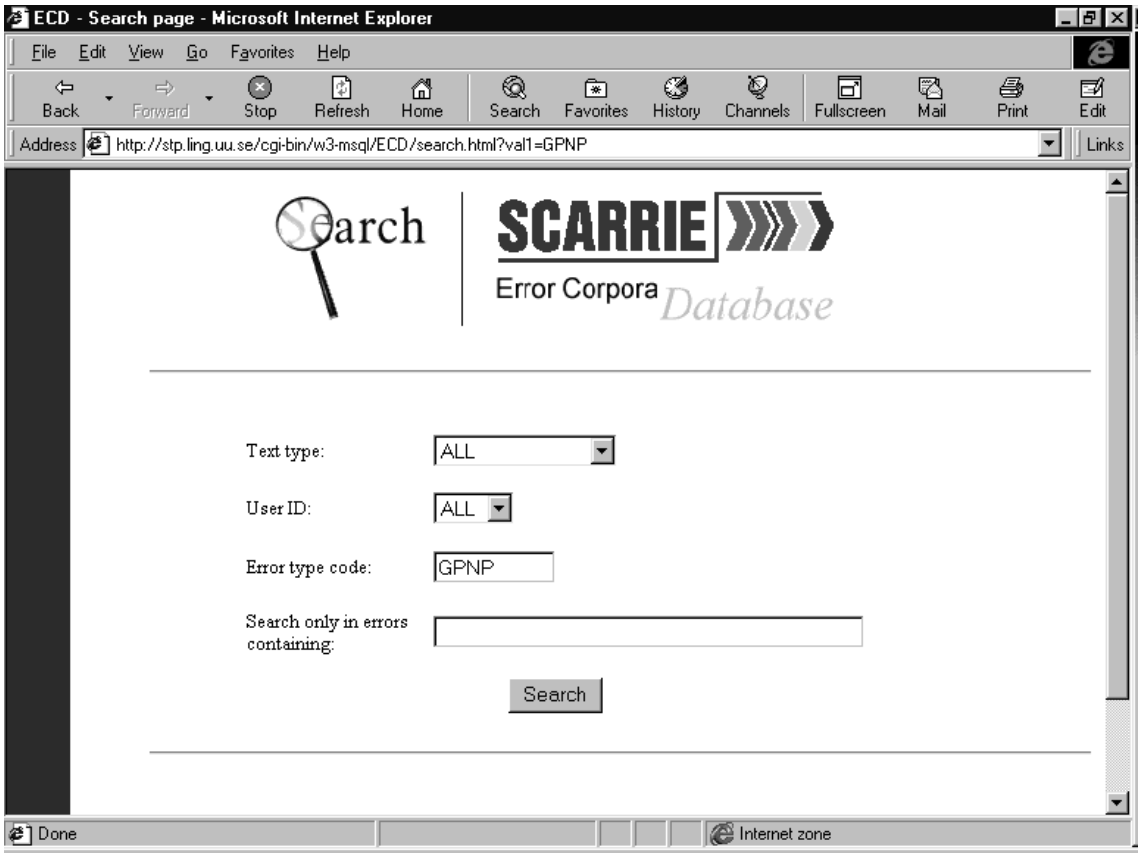

Figure 3.3 Search page

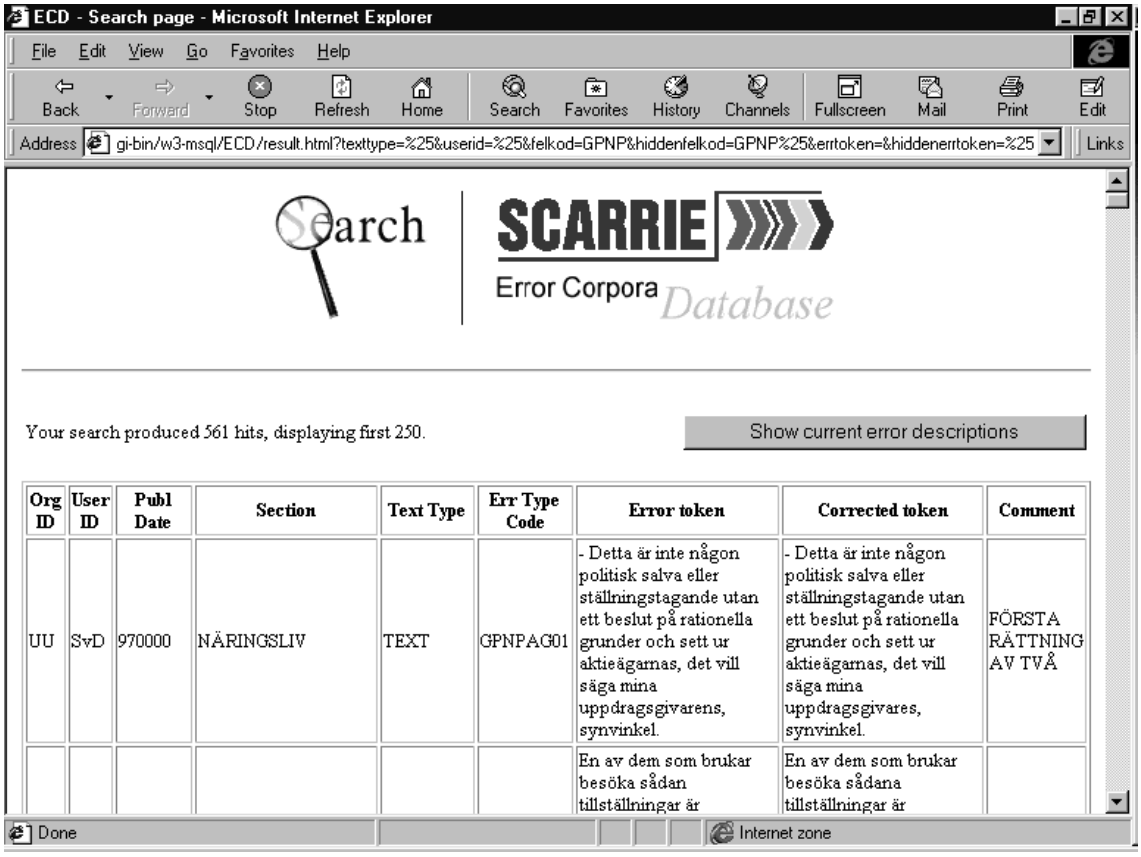

Figure 3.4 Result page

4 Error Collection

Non-proof-read and proof-read texts were supplied from Svenska Dagbladet (SvD) and Upsala Nya Tidning (UNT) for the Error Corpora Database. All proof-reading has been done at the newspapers by professional proof-readers. No jugdement has been made at the Department of Linguistics whether a correction is appropriate or not – all alterations in the material have been inserted into the database.

The material from SvD was delivered in electronic form. The two versions were processed automatically to find the sentences which had been altered by the proof-reader. UNT supplied the department with the proof-readers' paper copies on which they have marked the corrections to be made. The erroneous and the correct versions of the text fragments were inserted into the database, along with information about their origin. All errors were manually classified and assigned error type codes in accordance with a preliminary error typology.

If the incorrect sentence has been corrected in more than one aspect, it is first established whether the corrections depend on each other or not. If they do, they will be treated as one error, otherwise as separate errors. The erroneous text fragment contains only one error in each entry, which means that the other errors have been corrected. Notes are given in the comment fields when the original sentence contained more than one error.

If a reporter repeats the same error throughout an article, each error occurrence is registrated as a separate error. In the UNT corpus, it could be the case that an error has been corrected by the proof-reader but was not afterwards corrected by the editor, why the same error occurs in the next proof-reading turn. The second occurrence of exactly the same error ought not to be registrated, but so has been done.

The ECD is currently being revised in accordance with the Error Typology for Automatic Proof-reading Purposes (Deliverable 2.1 of the SCARRIE project; Wedbjer Rambell 1998)[5]. For some error instances, their sources need to be consulted; these instances are marked in the comment field (with KOLLA) until they have been properly checked. So far, nearly 7,000 of the 9,000 errors have been inspected. End of line hyphenation errors and problems with commas remain to be inspected.

4.1 The SvD Error Corpus

The reporters' versions of 734 articles were proof-read by a professional proof-reader at Svenska Dagbladet. The articles represent seven different sections in the newspaper: editorials, culture, foreign affairs, domestic affairs, local news, economy, and sports (see table 4.1). The paper is published seven days a week. Two weeks' production were covered for each section except for the sport pages which include more than two weeks. The articles were all written during the first eight months of 1997.

[5] The Error Typology for Automatic Proof-reading Purposes can be found on the Internet: http://stp.ling.uu.se/~olgaw/errortyp/index.html

Table 4.1 Supply of texts from Svenska Dagbladet

Section	Name in ECD	Time period	No of articles in 2 versions	No of errors
Editorial	LEDARE	970216–970301	93	205
Culture	KULTUR	970608–970621	66	424
Foreign affairs	UTRIKES	970518–970531	78	192
Domestic affairs	INRIKES	970112–970125	198	427
Local news	STOCKHOLM	970316–970329	121	216
Economy	NÄRINGSLIV	970727–970809	70	326
Sports	SPORT	970720–970809	108	310
SUM			734	2100

The non-proof-read and the proof-read articles were delivered in electronic form. The texts were presented in two formats. The most frequent format was a binary format for Apple Mac files. However, some files were encoded in an Apple Mac plain text format. There was no way of determining in what format a text was delivered without inspecting the text itself. Paper copies of the articles were not supplied by SvD.

In order to automatically compare the two versions of the articles, they had to be converted and aligned. First, the files were converted to the SGML format TEI Lite, a process which also divided the texts into sentences. The sentence boundaries were recognised by a point, a question mark or an exclamation mark immediately followed by a space. The end of paragraph token resulted in a sentence boundary tag as well.

After the format conversion, an alignment algorithm (Gale & Church 1993) were applied to the texts. This algorithm generated parallel texts: texts in which corresponding sentences have been linked with each other. A parallel corpus of proof-read and non-proof-read text in Swedish (Del 2.1.3.3 in the SCARRIE project) was automatically extracted from the collection of texts.

The next step was an automatic comparison of the aligned sentences. For every pair of sentences in which the proof-reader had made a correction, the comparison process (character by character) found that the aligned sentences differed from each other. All these sentence pairs were extracted and saved in a separate file. Most alignment structures contained a sentence together with an exact copy of itself, since most sentences in the original files were correct.

The sentence pairs with different members were manually checked to see if they contained alignment and conversion errors. The most problematic processing step was the first conversion step from Mac binary format to TEI Lite. No guarantee can be given that that the TEI files are 100% correct. In 50 texts conversion errors were found, and 33 texts contained alignment errors. A frequent conversion problem was the insertion of a space in the middle of a word. Alignment errors occurred when a point had been removed by the proof-reader, resulting in an extra sentence boundary in the first text version. The errors were corrected manually. However, four texts contained errors that could not be corrected: these errors were duplicate occurrences of text parts and missing text parts.

The final error corpus contained 1965 aligned structures in which errors were detected (2.47 aligned structures per text). An aligned structure may contain more than one error. The corpus was converted to the input data file format (see section 2.3). All dates were specified as 970000 because the exact publication dates of the articles were not provided. Since the articles were delivered as raw text, all error instances were given the text type TEXT (plain text). If the sentence were longer than 400 characters, it had to be shortened to conform to the ECD specification.

The errors in the database input file were manually analysed and assigned an error type code, and sometimes a comment. Eventually, the file was inserted into the ECD. The SvD corpus contains exactly 2,100 errors.

4.2 The UNT Error Corpus

During the winter/spring of 1997, Upsala Nya Tidning continuously delivered paper copies of proof-read articles. At UNT, articles are printed with the page lay-out, the proof-readers make the corrections on the paper, and the editors enter the corrections. An article may be proof-read several times, and exists in numerous versions, from raw text to printed article. Because of the difficulties in gathering articles in electronic form at the same stage in the production process, UNT supplied the proof-readers' paper copies.

The corpus covers the sections normally proof-read at the newspaper. From this supply, 25 days' production has been analysed and 6,801 errors were found, classified, typed in input files and inserted into the ECD (see table 4.2). There is no record of number of articles that are included in the corpus, nor of their distribution over the sections in the paper.

Upsala Nya Tidning is published six days a week, i.e. it does not have a Sunday edition. Due to typing errors, some errors have been assigned Sundays as publishing dates (2 errors 970413 and 4 errors 970427, see table 4.2), and two errors have dates that do not exist. In addition, there are 120 errors that have not been assigned a publishing date. The reason herefore is that some paper copies lack information about publishing dates.

Table 4.2 Publishing dates and numbers of errors in the UNT corpus

Publ date	No of errors
970214	9
970215	26
970217	47
970218	173
970219	237
970220	158
970221	22
970303	39
970304	210
970305	5
970306	244
970307	26
970405	26
970408	393
970409	31
970410	329
970411	52
970412	131
970413	2
970414	219
970415	317
970416	410
970417	255
970418	30
970419	410
970421	341
970422	204
970423	414
970424	276
970427	4
970428	226
970429	306
970430	307
970502	384
970503	153
970505	264
971419	1
97050	1
?	120

Sections in the newspaper that occur in the ECD are presented in table 4.3 with their error frequencies. The names in the ECD differ from the SvD corpus; they are in accordance with those used in the separate newspapers. For 44 of the errors, the section is unknown.

Table 4.3 Sections in UNT and error frequencies

Section	Name in ECD	No of errors
First page	ETTAN	471
Editorial	LEDARE	771
Signed articles	SIGNERAT	20
Debate	DEBATT	628
Letters to the editor	LÄSARNAS_FORUM	169
Local news	UPPLAND	997
	UPPSALA	966
	UPPSALA_I_DAG	34
Economy	EKONOMI	33
Domestic affairs	SVERIGE	183
Foreign affairs	VÄRLDEN	174
Culture	KULTUR	443
Entertainment	NÖJE	269
	NÖJESSVEPET	13
Research	FORSKNING	14
Light articles	LIKT_OCH_OLIKT	139
Today	FÖR_DAGEN	290
Special feature	TEMA	23
IT	IT	57
Magazine	MAGASINET	9
Sports	SPORT	492
Family pages	FAMILJ	219
	FAMILJENYTT	288
TV & radio programs	TV_OCH_RADIO	21
Chess	SCHACK	35
Unknown	?	44

4.3 Problems and limitations

The work with the Error Corpora Database has not been without problems. Here, the most important problems and limitations are outlined.

The character set used in the Error Corpora Database is limited. Special characters and signs used at the newspapers are not always included in the character set, e.g. dash. This problem has not been approached systematically, different solutions to the problem appear in the database. As it is now, a dash is respresented by a hyphen (with or without comments) or by two hyphens. A hyphen may thus represent a hyphen or a dash. To know which, one would have to check with the source.

Lack of knowledge of how the database input files were to be transferred to the Error Corpora Database, at the time they were produced, has resulted in errors. For instance, a new line cannot be represented by a new line token in the input file, since the new line token is converted into a space token when the input file is inserted into the database. As a result, the new line token lacks a representation in the database. It ought to have been represented by another character, such as the vertical line |.

Even though characters are in the database character set, they can not be represented in an appropriate manner by a webbrowser. The tabulation is such a character. Also, double space tokens are represented as a single space. Thus this kind of information in the database is lost when presented on the Internet.

The text fragment of the language error is sometimes too small. The context might be sufficient for identification of the error type, but too small for the further use of the database as a source of information on error recognition and correction. This is especially true for the UNT corpus, but also for the SvD corpus when an error extends over the sentence boundaries.

For the SvD corpus, errors that have been made during the conversion and the alignment processes may have been overlooked in the manual inspection. Conversion errors may influence the character used, so that the original character is replaced by another character. Printed copies of the articles are not available to consult in problematic cases. In the alignment process a sentence boundary was recognised by the combination of a point, a question mark or an exclamation mark and a following space. Thus a point after an abbreviation (not ending the sentence) was incorrectly interpreted as a sentence boundary. If the proof-reader removed the point, the result would not only be the removed point but also a removed new line token. This problem was known during the work with the SvD files and taken into consideration during the classification of the errors, but errors caused by the alignment process may still be present in the database.

For the UNT corpus, the proof-readers' correction marks may have been misinterpreted. This is very difficult to discover, but may lead to either wrong error type code or errors in the corrected version of the text fragment. Since the classification and the typing were done by different persons, there might have been misunderstandings of the code notation made by the person who classified. In addition, ordinary typing errors may also be present in the database.

The problems mentioned here may be overcome. As a first step, probable alignment errors and typing errors will be examined. Too small text fragments will be expanded, especially if the error type code indicates a grammatical error.

5 Error Frequencies

The error typology is presented in more detail in the report Error Typology for Automatic Proof-reading Purposes (Deliverable 2.1; Wedbjer Rambell 1998). The typology consists of four levels: group, category, subcategory, and specification. Each level is assigned a two character code which is concatenated into the resulting 8 character code. For example, the spelling errors group is assigned the code SE. The category of word formation errors within this group are assigned the code SEWF. This category is in turn divided into subcategories, so that split words are assigned the error type code SEWFSW. There are different kinds of split words, which are reflected in the specifications. The specifications are given a 2 figure code, e.g. 01 if the the two words that are erroneously separated may correctly appear on their own. The resulting error type code for this particular error is SEWFSW01.

The most important findings are presented in this chapter, a full list of the error type code frequencies is found in Appendix A. The numbers do not always add up: Some errors have by mistake been assigned non-existing error type codes, other errors have not yet been subdivided on the specification level. Because of the differences in the material, the two newspapers are presented both together and separately. A study of the error distribution over the text types and sections in the newspapers have not yet been made.

5.1 The five error groups

The error typology consists of five major error groups: spelling errors, grammar prolems, punctuation problems, graphical problems, and style, meaning, and reference problems. The spelling errors group contain capital letter errors, word formation errors, end of line hyphenation errors, and ordinary spelling errors. The grammar problems group involve grammatical errors that may be recognised and corrected within the individual sentence. The style, meaning, and reference group contain problems reaching over sentence boundaries, and problems involving choice between alternative expressions.

Graphical problems concern graphical and typographical representation, such as the length of the dash. Problems with quotation marks and parentheses are also viewed primarily as graphical problems. Punctuation problems are about the usage of punctuation marks, such as full stops, question marks, exclamation marks, commas, dashes, colons, and semicolons. Missing capital letter at the beginning of a sentence is also dealt with in the punctuation problems group.

The material provided by Svenska Dagbladet consisted of the reporters' versions delivered as raw text. Therefore, several error types are lacking, e.g. end of line hyphenation errors and certain graphical problems. As shown in table 5.1, the relative numbers of errors in the spelling error group and in the graphical problems group are smaller for the SvD corpus than for the UNT corpus.

Spelling errors are the most common cause for proof-readers' corrections. In total, over 40% of the errors in the ECD are spelling errors. Grammar problems, punctuation problems, and graphical problems are equally common, approximately 16% each. Graphical problems is the least common error groups, less than 10%.

Table 5.1 Distribution of errors in the five error groups

Error group	Code	UNT		SvD		TOT	
Spelling Errors	SE	3 086	45,4%	723	34,5%	3 809	42,8%
Grammar Problems	GP	984	14,5%	390	18,6%	1 374	15,4%
Punctuation Problems	PU	1 009	14,8%	468	22,3%	1 477	16,6%
Graphical Problems	GR	670	9,9%	120	5,7%	790	8,9%
Style, Meaning and Reference	SP	1 049	15,4%	397	18,9%	1 446	16,3%
SUM		6 798	100%	2 098	100%	8 896	100%

5.2 Spelling Errors

Spelling errors is the most common error group in the ECD. The total number of spelling errors differ between table 5.1 above and table 5.2 below because there are error instances in the database which have been assigned a non-existing error type code beginning with SE, but not continuing with CP, WF, HY, or OS. These error entries are to be found and corrected.

As stated earlier, the SvD corpus lacks end of line hyphenation errors due to the stage in the production chain in which the articles were proof-read, i.e. before the articles were put into the page lay-out. In the UNT corpus, hyphenation errors account for 40% of the spelling errors (see table 5.2). No further subcategorisation has been made for the end of line hyphenations errors, e.g. if the morpheme boundary rule, the one consonant rule, or both rules have been violated.

Table 5.2 Spelling error categories

Spelling error category	Code	UNT		SvD		TOT	
Capital Letter Errors	SECP	409	13,3%	128	17,7%	537	14,1%
Word Formation Errors	SEWF	630	20,4%	346	47,9%	976	25,6%
End of Line Hyphenation	SEHY	1 263	40,9%	0	0,0%	1 263	33,2%
(Other) Spelling Errors	SEOS	783	25,4%	248	34,3%	1 031	27,1%
SUM		3 085	100%	722	100%	3 807	100%

Word formation errors consist of erroneous binding morpheme or erroneous binding hyphen, but also problems with split words and concatenated words. In the SvD corpus, word formation errors are more common than ordinary spelling errors (48% compared to 34%). The opposite is true for the UNT corpus, the ordinary spelling errors are more common than word formation errors (25% versus 20%). This is also true for the material as a whole. Capital letter errors, i.e. whether the first letter in a word should be in the lower or in the upper case, show the smallest number of occurrences in both newspapers.

5.2.1 Word Formation Errors

Word formation errors may be of different types, which is shown in table 5.3. About 50% of the word formation errors in the UNT corpus concern split words and concatenated words, i.e. one word has erroneously been written as two words and vice versa. These error types are not as frequent in the SvD corpus as in the UNT corpus: about 30% of the word formations errors. Instead, incorrect hyphens which are to be removed are much more common accounting for nearly half of the word formation errors in the SvD corpus. For both newspapers, problems with the binding morpheme *s* are not very common, neither are missing hyphens, misplaced space, coordination with common word part, or erroneously formed abbreviations.

Table 5.3 Word formation errors subcategories

Word formation subcategory	Code	UNT		SvD		TOT	
Bindings -s- incorrect	SEWFSI	19	3,0%	8	2,3%	27	2,8%
Binding -s- missing	SEWFSM	48	7,6%	34	9,8%	82	8,4%
Hyphen missing	SEWFHM	43	6,8%	12	3,5%	55	5,6%
Incorrect hyphen	SEWFHI	87	13,8%	162	46,8%	249	25,5%
Split words: several words => one word	SEWFSW	182	28,9%	65	18,8%	247	25,3%
Concatenated words: one word => several words	SEWFCW	144	22,9%	40	11,6%	184	18,9%
Misplaced space	SEWFMS	3	0,5%	0	0,0%	3	0,3%
Coordination with common word part	SEWFCO	38	6,0%	13	3,8%	51	5,2%
Abbreviations	SEWFAB	57	9,0%	10	2,9%	67	6,9%
Other word formation errors	SEWFOP	9	1,4%	2	0,6%	11	1,1%
SUM		630	100%	346	100%	976	100%

The split words and concatenated words subcategories are further divided on the specification level (see Appendix A). Most of the problems with split words involve two lexical words, i.e. two words that may appear on their own should be written together as one word. This problem can not be detected by a word checker, since the two words probably would be in the dictionary. For the second most frequent split words error, one of the words needs correction, and do therefore not appear in a dictionary. Combinations of split words errors and capital problems are uncommon.

Half of the concatenated words errors in both newspapers involve two lexical words that ought to be separated with a space. Other fairly common errors concern two words one of which needs correction, and two words that have been concatenated with a hyphen which should be replaced by a space.

5.3 Grammar Problems

Four problems out of ten are located in the noun phrase (see table 5.4). One problem out of four concerns the verb, either in a combination of verbs (i.e. in a verb phrase in the limited sense) or as a verb valency error. Problems with prepositions are located in four categories depending on what govern the prepositions: a preceeding noun, adjective or verb, or other factors. The prepositional phrase category gathers the latter ones, being the third largest category. Together, preposition related problems account for 15–20% of the total number of grammar problems. The least common problems occur in adjective phrases and adverb phrases on the top level of the clauses. Wrong word category problems on clause or sentence levels are also quite uncommon, and so are pronoun case errors.

Table 5.4 Distribution of errors in the grammar problems categories

Grammar problem category	Code	UNT		SvD		TOT	
Noun Phrase	GPNP	414	42,1%	147	37,7%	561	40,8%
Adjective Phrase	GPAP	5	0,5%	3	0,8%	8	0,6%
Adverb Phrase	GPAB	5	0,5%	1	0,3%	6	0,4%
Prepositional Phrase	GPPP	114	11,6%	38	9,7%	152	11,1%
Conjunctions and Conjunctive Adverbs	GPCN	50	5,1%	21	5,4%	71	5,2%
Verb Phrase in the Limited Sense	GPVF	79	8,0%	33	8,5%	112	8,2%
Verb Valency	GPVV	151	15,3%	88	22,6%	239	17,4%
Pronoun Case	GPPC	11	1,1%	10	2,6%	21	1,5%
Agreement	GPAG	42	4,3%	15	3,8%	57	4,1%
Referential Problems	GPRP	26	2,6%	10	2,6%	36	2,6%
Word Order	GPWO	48	4,9%	8	2,1%	56	4,1%
Wrong Word Category	GPWC	13	1,3%	1	0,3%	14	1,0%
Other Grammar Problems	GPOG	26	2,6%	15	3,8%	41	3,0%
SUM		984	100%	390	100%	1374	100%

The distribution of errors in the two newspapers are quite similar; there is no particular deviance to be remarked upon. The five largest categories are presented (after size) in more detail in the following sections.

5.3.1 Noun Phrase

Problems located in the noun phrase belong to the noun phrase category, which is the largest grammar problem category. As shown in table 5.5, within the noun phrase agreement errors are the most common error type, being more frequent in the SvD corpus (33%) than in the UNT corpus (26%). The second largest subcategory is the species subcategory (22% in the corpora), embracing problems with article usage and usage of

the definite and indefinite forms, followed by the case subcategory (14% in the corpora).

Table 5.5 Noun phrase subcategories

Noun phrase subcategory	Code	UNT		SvD		TOT	
Agreement	GPNPAG	107	26,0%	48	33,1%	155	27,8%
Gender	GPNPGE	25	6,1%	11	7,6%	36	6,5%
Number	GPNPNB	4	1,0%	3	2,1%	7	1,3%
Species	GPNPSS	97	23,5%	27	18,6%	124	22,3%
Case	GPNPCA	56	13,6%	23	15,9%	79	14,2%
Adjective phrase	GPNPAP	19	4,6%	5	3,4%	24	4,3%
Participles	GPNPPE	10	2,4%	1	0,7%	11	2,0%
Numerals	GPNPNL	2	0,5%	4	2,8%	6	1,1%
Nouns	GPNPNN	24	5,8%	7	4,8%	31	5,6%
Pronouns	GPNPPN	23	5,6%	3	2,1%	26	4,7%
Choice of preposition after a noun	GPNPCP	19	4,6%	4	2,8%	23	4,1%
Preposition missing after a noun	GPNPMP	7	1,7%	3	2,1%	10	1,8%
Other noun valency problems	GPNPNV	1	0,2%	1	0,7%	2	0,4%
Coordination	GPNPCO	11	2,7%	1	0,7%	12	2,2%
Word order	GPNPWO	2	0,5%	1	0,7%	3	0,5%
Other problems	GPNPOP	5	1,2%	3	2,1%	8	1,4%
SUM		412	100%	145	100%	557	100%

All other subcategories have a relative number of errors less than 8% each. Together, the three subcategories dealing with noun valency problems account for less than 7% of the noun phrase problems. Word order problems in noun phrases are very rare.

Agreement

The agreement errors are distributed over the specifications as shown in table 5.6. Number agreement errors is the most frequent error type, accounting for nearly half of the total number of agreement errors. The number agreement errors occur in most cases between a premodier (an article, a pronoun, an adjective etc on the left-hand side of the head noun) and the noun. However, number agreement errors are more frequent in the UNT corpus than in the SvD corpus. Gender agreement errors are more common than species agreement errors in the UNT corpus, while the two error types are equally common in the SvD corpus.

Table 5.6 Agreement errors in noun phrases

Agreement specification	Code	UNT		SvD		TOT	
number agreement in:							
premodifier - noun	GPNPAG01	44	44,0%	16	34,0%	60	40,8%
coordinated head nouns	GPNPAG05	1	1,0%	0	0,0%	1	0,7%
coordinated nouns in the genitive	GPNPAG06	0	0,0%	0	0,0%	0	0,0%
noun - postmodifier	GPNPAG07	2	2,0%	1	2,1%	3	2,0%
noun phrases in apposition	GPNPAG08	2	2,0%	1	2,1%	3	2,0%
species agreement in:							
premodifier - noun	GPNPAG03	19	19,0%	13	27,7%	32	21,8%
premodifier - adjective functioning as a noun	GPNPAG10	0	0,0%	1	2,1%	1	0,7%
premodifier - coordinated nouns	GPNPAG11	0	0,0%	0	0,0%	0	0,0%
gender agreement in:							
premodifier - noun	GPNPAG02	31	31,0%	14	29,8%	45	30,6%
noun - relative pronoun	GPNPAG04	0	0,0%	0	0,0%	0	0,0%
premodifier - adjective/participle functioning as a noun	GPNPAG09	0	0,0%	1	2,1%	1	0,7%
noun - postmodifier	GPNPAG13	1	1,0%	0	0,0%	1	0,7%
agreement in coordinated nouns	GPNPAG12	7	7,0%	1	2,1%	8	5,4%
SUM		100	100%	47	100%	147	100%

Species

Species agreement rules may be overruled by other language rules. Problems in this respect are gathered in the species subcategory. In table 5.7, the distribution of errors in the specifications of the species subcategory show that the most frequent error (approx. 26%) in both newspapers is problematic indefinite noun phrases in the singular without article. The noun phrase is corrected either by inserting the definite article or by inserting the definite article and adding the definite inflection to the head noun.

The second most common error type in the UNT corpus is problems in definite noun phrase with adjective premodifier but with no definite article (20%) in which the definite article is inserted or the definite inflection is removed. In the SvD corpus, this error type is as common as erroneous definite inflection after genitive attribute (18%), which is a quite uncommon error type in the UNT corpus (4%).

The distribution among the other specifications differs between the two newspapers. Specifications accounting for 8–10% of the total number of errors each are: missing definite inflections in noun phrases in prepositional phrases, other cases of missing definite inflections, and indefinite articles which are to be removed. On the whole, definite inflections account for many of the species problems. Articles are inserted more often than they are removed.

Table 5.7 Error specifications in the species subcategory

Species specification	Code	UNT		SvD		TOT	
def art missing or erroneous def infl in def NP with adj premod	GPNPSS01	19	19,6%	5	18,5%	24	19,4%
erroneous def infl after gen attr	GPNPSS02	4	4,1%	5	18,5%	9	7,3%
indef art missing or def art (and def infl) missing in indef NP in the sing without article	GPNPSS03	26	26,8%	7	25,9%	33	26,6%
definite inflection missing in noun phrase in PP	GPNPSS09	11	11,3%	1	3,7%	12	9,7%
other cases of missing definite inflection	GPNPSS12	5	5,2%	1	3,7%	6	4,8%
erroneous def infl before a necessary relative clause	GPNPSS04	3	3,1%	0	0,0%	3	2,4%
erroneous def infl after certain pronouns and adjectives	GPNPSS05	4	4,1%	0	0,0%	4	3,2%
erroneous definite inflection in titles	GPNPSS10	3	3,1%	1	3,7%	4	3,2%
other cases of erroneous definite inflection	GPNPSS11	7	7,2%	3	11,1%	10	8,1%
demonstrative pronoun / definite article should be removed	GPNPSS06	1	1,0%	3	11,1%	4	3,2%
the indefinite article should be removed	GPNPSS07	10	10,3%	1	3,7%	11	8,9%
double articles	GPNPSS08	4	4,1%	0	0,0%	4	3,2%
SUM		100	100%	47	100%	147	100%

Table 5.8 Error specifications in the case subcategory

Case specification	Code	UNT		SvD		TOT	
common noun should be in the genitive case	GPNPCA01	18	32,1%	4	17,4%	22	27,8%
proper noun should be in the genitive case	GPNPCA02	12	21,4%	12	52,2%	24	30,4%
the genitive case => the basic case	GPNPCA03	23	41,1%	5	21,7%	28	35,4%
error in forming the genitive case in word group	GPNPCA04	0	0,0%	0	0,0%	0	0,0%
pronoun should be possessive pronoun	GPNPCA05	1	1,8%	0	0,0%	1	1,3%
adjective used as a noun should be in the genitive case	GPNPCA06	1	1,8%	2	8,7%	3	3,8%
other problems with case	GPNPCA07	1	1,8%	0	0,0%	1	1,3%
SUM		56	100%	23	100%	79	100%

The distribution of errors in the UNT corpus and the SvD corpus differs for the case subcategory as well (see table 5.8). But if the two first specifications are added, nouns being in the basic case while they should be in the genitive case are the most frequent problem in both UNT and SvD (54% and 70%, respectively). In the UNT corpus, over 40% of the case errors involve changing the genitive case to the basic case, while this error type is only half as frequent in the SvD corpus (22%).

Problems with possessive pronouns are very rare, along with errors in forming the genitive case in word groups. There are a few occurrences with adjectives used as nouns that should be in the genitive case but are not.

5.3.2 Verb Valency

Verb valency is the second largest grammar problem category: 15% in the UNT corpus, 23% in the SvD corpus, 17% in the corpora. Table 5.9 shows that the largest verb valency subcategory in both newspapers is the infinitive phrase (18% and 35% in UNT and SvD respectively, 25% in the corpora). Over 90% of the problems with infinitive phrases in the verb valency category concern missing infinitive mark *att*. One third of these instances occur after the verb *komma*. Problems with infinitive phrases are also classified in the other verb related category. Those problems, concerning the infinitives, are rare.

Table 5.9 Verb valency subcategories

Verb valency subcategory	Code	UNT		SvD		TOT	
Intransitivity	GPVVIN	6	4,0%	2	2,3%	8	3,3%
Transitivity	GPVVTR	5	3,3%	5	5,7%	10	4,2%
Copula	GPVVCO	1	0,7%	0	0,0%	1	0,4%
Reflexivity	GPVVRE	4	2,6%	3	3,4%	7	2,9%
Passive constructions	GPVVPC	14	9,3%	3	3,4%	17	7,1%
Object with infinitive	GPVVOI	1	0,7%	2	2,3%	3	1,3%
Prepositional phrase	GPVVPP	1	0,7%	0	0,0%	1	0,4%
Infinitive phrase	GPVVIP	28	18,5%	31	35,2%	59	24,7%
Clause	GPVVCL	8	5,3%	5	5,7%	13	5,4%
Position holding "det"	GPVVID	9	6,0%	1	1,1%	10	4,2%
VF missing	GPVVVM	20	13,2%	8	9,1%	28	11,7%
NP missing	GPVVNM	22	14,6%	12	13,6%	34	14,2%
Choice of preposition/adverb after verbs	GPVVCP	22	14,6%	4	4,5%	26	10,9%
Preposition/adverb missing after verbs	GPVVMP	9	6,0%	8	9,1%	17	7,1%
Repetition of preposition/adverb	GPVVRP	1	0,7%	4	4,5%	5	2,1%
SUM		151	100%	88	100%	239	100%

Problems with missing noun phrases as subjects of clauses are the second most common subcategory in both UNT and SvD. Missing or erroneous subjects, objects, and verb phrases in the limited sense account together for more than a third of the verb valency errors. Problems with verb governed prepositions and adverbs account for a fifth of the verb valency errors.

5.3.3 Prepositional Phrase

The prepositional phrase category are the third largest category in the grammar problems group accounting for a tenth of the grammar problems. Together with noun valency, adjective valency, and verb valency problems concerning prepositions, the percentage preposition related problems is higher, approximately 17% of the corpora.

There are two subcategories in the prepositional phrase category: prepositions and complements with 91–92% and 8–9% of the errors respectively in both newspaper materials.

Prepositions

The prepositions subcategory contain ten specifications (see table 5.10). The distribution among them are not the same in the two newspapers. In the UNT corpus, the most common problem is choice of preposition (42%), and the second most common problem is missing preposition (27%). Together, these problems account for over two thirds of the errors in the prepositions subcategory.

Table 5.10 Error specifications in the prepositions subcategory

Prepositions specification	Code	UNT		SvD		TOT	
preposition to be removed - should not be a prepositional phrase	GPPPPR01	5	4,8%	4	11,4%	9	6,5%
one preposition too many	GPPPPR02	8	7,7%	5	14,3%	13	9,4%
preposition missing	GPPPPR03	28	26,9%	9	25,7%	37	26,6%
wrong preposition; choice of preposition	GPPPPR04	44	42,3%	6	17,1%	50	36,0%
wrong word category	GPPPPR05	12	11,5%	8	22,9%	20	14,4%
wrong preposition in coordinated PPs	GPPPPR06	1	1,0%	0	0,0%	1	0,7%
doubled preposition	GPPPPR07	2	1,9%	2	5,7%	4	2,9%
preposition missing in coordination of phrases - phrases of different types	GPPPPR08	3	2,9%	1	2,9%	4	2,9%
preposition missing in coordination of phrases - phrases of the same type	GPPPPR09	0	0,0%	0	0,0%	0	0,0%
comma => preposition	GPPPPR10	1	1,0%	0	0,0%	1	0,7%
SUM		104	100%	35	100%	139	100%

The errors are more spread among the specifications in the SvD corpus compared to the UNT corpus. In the SvD corpus, the most frequent problem is missing preposition (26%) followed by wrong word category (23%). (Choice between *innan* and *före* is conceived of as a wrong word category error.) The third most frequent error in the UNT corpus is wrong word category (12%), and in the SvD corpus it is choice of preposition (17%). In total, these three specifications mentioned account for 77% of the errors.

5.2.4 Verb Phrase in the Limited Sense

The category of the verb phrase in the limited sense is the forth largest category, accounting for 8–9% of the grammar problems in both newspapers. The distribution within the category differs between the two papers (see table 5.11). In the SvD corpus, the occurrences are more spread over the different error types. However, the three largest groups are the same as for the UNT corpus.

Table 5.11 Subcategories in the verb phrase in the limited sense

Verb phrase subcategory	Code	UNT		SvD		TOT	
Main verb in the finite form	GPVFMF	27	34,2%	7	21,9%	34	30,6%
Temporal auxiliary verb in the finite form + Main verb in the supine	GPVFTS	11	13,9%	9	28,1%	20	18,0%
Existential auxiliary verb in the finite form + Main verb in the perfect participle	GPVFEP	1	1,3%	1	3,1%	2	1,8%
Auxiliary verb in the finite form + Main verb in the infinitive	GPVFAI	31	39,2%	6	18,8%	37	33,3%
Combination of auxiliary verb + Main verb	GPVFAM	2	2,5%	5	15,6%	7	6,3%
Coordination of verbs	GPVFCO	3	3,8%	0	0,0%	3	2,7%
Infinitive in infinitive phrase	GPVFIP	2	2,5%	4	12,5%	6	5,4%
Other problems	GPVFOP	2	2,5%	0	0,0%	2	1,8%
SUM		79	100%	32	100%	111	100%

In the UNT corpus, errors in the sequence auxiliary verb + inifinitive are most common problem (39%), closely followed by errors concerning main verbs in the finite form (34%). Temporal auxiliary verb + the supine is also a problematic verb combination (14%). The other error types have only a few occurrrences each.

The most frequent error made concerning main verbs in the finite form is an infinitive that should be changed into the present or the past verb form (see Appendix A). For the combination of a temporal auxiliary verb followed by the main verb in the supine, a missing auxiliary verb is the most common error. (Omitted temporal auxiliary verbs in subordinate clauses are style problems, and as such handled in the style, meaning, and reference group.) For auxiliary verbs followed by infinitives, the combination of two verbs in the present tense is the most frequent specification. Two thirds of the problems in this subcategory concern the infinitive standing in an incorrect verb form.

5.3.5 Conjunctions and Conjunctive Adverbs

Conjunctions and conjunctive adverbs is the fifth largest grammar problem category with 5–6% of the grammar problems. Erroneously formed complex conjunctions is the most common subcategory in the corpora (see table 5.12), of which two thirds are discontinuous. The second most frequent error type is missing conjunctions or conjunctive adverbs, some of which are replacing commas.

Table 5.12 Error specifications in the conjunctions and conjunctive adverbs subcategory

Conjunctions and conjunctive adverbs specification	Code	UNT		SvD		TOT	
Conjunction or conjunctive adverb missing	GPCNCM	14	28,6%	4	19,0%	18	25,7%
Complex conjunction	GPCNCC	12	24,5%	9	42,9%	21	30,0%
Doubled conjunctions	GPCNDW	8	16,3%	2	9,5%	10	14,3%

Erroneous conjunction	GPCNEC	8	16,3%	3	14,3%	11	15,7%
Wrong word category	GPCNWC	7	14,3%	3	14,3%	10	14,3%
SUM		49	100%	21	100%	70	100%

5.4 Punctuation Problems

Punctuation problems concern the use of punctuation marks: full stop (or point), comma, dash within the sentence, colon, and semicolon. The punctuation problems are as common as grammar problems and style, meaning, and reference problems. By far the most frequent punctuation problem concerns the comma, over 70% in both newspapers (see table 5.13). The second largest subcategory involves end of sentence punctuation (18%). The comma subcategory has not been revised, why more detailed information about the distribution of comma errors is not given.

Table 5.13 Punctuation problems categories

Punctuation problems category	Code	UNT		SvD		TOT	
End of Sentence Punctuation	PUES	183	18,2%	77	16,5%	260	17,6%
Capital Letter	PUCP	34	3,4%	23	4,9%	57	3,9%
Comma	PUCO	716	71,0%	351	75,0%	1 067	72,3%
Dash within the Sentence	PUDW	9	0,9%	7	1,5%	16	1,1%
Colon	PUCN	20	2,0%	7	1,5%	27	1,8%
Semicolon	PUSN	2	0,2%	3	0,6%	5	0,3%
Other Punctuation Problems	PUOP	44	4,4%	0	0,0%	44	3,0%
SUM		1 008	100%	468	100%	1 476	100%

5.4.1 End of sentence punctuation

A third of the end of sentence punctuation problems, in both newspapers, are missing punctuation marks (see Appendix A). Problems in choosing the proper punctuation mark are more common in the UNT corpus (22% in UNT versus 10% in SvD), while problems with combining punctuation marks with quotations marks or parentheses are more frequent in the SvD corpus (38% in SvD versus 14% in UNT). Problems with punctuation marks in the middle of sentences are also quite common (19% in UNT and 12% in SvD).

5.5 Graphical Problems

Graphical problems is the smallest error group gathering less than 10% of the total material. The error instances are not equally distributed in the two newspapers (see table 5.14). In the UNT corpus, the second largest category involves dash within the sentence. This category is not represented at all in the SvD corpus because the SvD corpus was

collected at an earlier stage in the production process before the page lay-outs were made. Because of the same reason, there is no typographical error in the SvD corpus while it is the third most common category in the UNT corpus. The second most frequent problem in the SvD corpus concerns new lines or new paragraphs. This may be a result of the alignment and comparison process of the non-proof-read and proof-read text versions, and not a result of the proof-reader's corrections.

Table 5.14 Graphical problems categories

Graphical problems category	Code	UNT		SvD		TOT	
Space	GRSC	411	61,3%	59	49,2%	470	59,5%
New Line / Paragraph	GRNL	38	5,7%	42	35,0%	80	10,1%
Dash before Direct Speech	GRDS	17	2,5%	3	2,5%	20	2,5%
Dash within the Sentence	GRDW	91	13,6%	0	0,0%	91	11,5%
Quotation marks	GRQM	57	8,5%	10	8,3%	67	8,5%
Parentheses	GRPA	5	0,7%	2	1,7%	7	0,9%
Typographical Errors	GRTY	49	7,3%	0	0,0%	49	6,2%
Other graphical errors	GROP	2	0,3%	4	3,3%	6	0,8%
SUM		670	100%	120	100%	790	100%

5.5.1 Space

The most frequent graphical problem category is space in both newspapers. Problems with space involve either too little space or too much space, of which too little space is more frequent (including missing space). The problems with space in the graphical problems group do not include space between words, but only at punctuation marks and graphical signs. Errors involving space between words are dealt with in the spelling errors group.

5.6 Style, Meaning and Reference

The style, meaning and reference group contains problems in choosing between alternatives, and problems reaching over sentence boundaries. The most frequent category is preferred spelling (28–29% of the style, meaning, and reference problems in both newspapers), i.e. choice between correct word forms (see table 5.15). Choice between correct abbreviations is also a common category (14% in the UNT corpus and 22% in the SvD corpus). Many of these problems involve removing points from abbreviations.

How to present numbers is also a common problem (15% in the UNT corpus and 24% in the SvD corpus). Changing words and expressions has been done to a larger degree in the UNT corpus than in the SvD corpus (30% and 14%, respectively). This could partly be due to the proof-readers' experience and professionalism, since changing the reporters' words and expressions is a more delicate matter than correcting spelling errors.

Table 5.15 Style, meaning, and reference categories

Style, meaning, and reference category	Code	UNT		SvD		TOT	
Preferred Spelling	SPPS	297	28,6%	112	28,3%	409	28,5%
Abbreviation	SPAB	144	13,8%	87	22,0%	231	16,1%
Number Style	SPNS	153	14,7%	97	24,5%	250	17,4%
Correct Word Category but Wrong Word	SPWN	170	16,3%	28	7,1%	198	13,8%
Choice of Words and Expressions	SPCW	142	13,7%	27	6,8%	169	11,8%
Choice of Signs	SPCS	40	3,8%	8	2,0%	48	3,3%
Choice of Sentence Boundaries	SPCB	27	2,6%	11	2,8%	38	2,6%
Choice of Syntactic Construction	SPSC	15	1,4%	6	1,5%	21	1,5%
Consistency	SPCN	0	0,0%	2	0,5%	2	0,1%
Redundancy	SPRD	32	3,1%	14	3,5%	46	3,2%
Referential Problems	SPRP	20	1,9%	4	1,0%	24	1,7%
SUM		1 040	100%	396	100%	1 436	100%

Words may also be removed when the information already is given to the reader (redundancy). Sentence boundaries are seldom changed, which means that the proof-reader does not alter the text structure in any significant way.

6 Closing Remarks

The relational database with its interface has been an efficient tool for working with the corpora. The www-based search interface has functioned very well. It can though be made even better by future extensions of its functionality. For instance, the ability to search in more fields, and to perform more complex searches may be added.

A more thorough analysis of the error collection process may had prevented some of the problems that have been discussed in this report. Some minor polishing needs to be done in order to make the database more reliable as a source of information. Probable alignment errors are going to be checked, and so are the typing error in the error codes. The work with the inspection of the database will be finished in the near future.

The Error Corpora Database has become larger, containing more error entries, than was initially planned. The size of the database, nearly 9,000 error entries, was found to be necessary for reaching reliable conclusions about error frequencies. The results are in that respect dependent on the size and the quality of the corpora. The ECD seems to provide a very good starting point for the work with the grammar checker (work package 6). It will also be a useful source of information for the word checking module. The efforts made in building the error database will, no doubt, pay off.

Literature

Gale, William A. & Church, Kenneth W (1993): A program for aligning sentences in bilingual corpora. In: *Computational Linguistics*, 19(1), 1993.

Wedbjer Rambell, Olga (1998): *Error Typology for Automatic Proof-reading Purposes*. SCARRIE, Deliverable 2.1, version 1.1.

Appendix A

Error Frequencies

Code	Error Type Code Description	UNT	SvD	TOT
SE	**Spelling Errors**	**3086**	**723**	**3809**
SECP	Capital Letter Errors	**409**	**128**	**537**
SECPPN	Proper nouns	335	76	**411**
SECPPN01	lower case letter => upper case letter	156	23	**179**
SECPPN02	upper case letter => lower case letter	178	53	**231**
SECPCC	Compounds with proper nouns	42	44	**86**
SECPCC01	lower case letter => upper case letter	29	31	**60**
SECPCC02	upper case letter => lower case letter	13	13	**26**
SECPDC	Derivations of proper nouns	9	2	**11**
SECPDC01	lower case letter => upper case letter	4	0	**4**
SECPDC02	upper case letter => lower case letter	5	2	**7**
SECPPT	Personal titles	11	2	**13**
SECPPT01	lower case letter => upper case letter	2	0	**2**
SECPPT02	upper case letter => lower case letter	9	2	**11**
SECPFT	Foreign names	11	4	**15**
SECPFT01	lower case letter => upper case letter	5	1	**6**
SECPFT02	upper case letter => lower case letter	6	3	**9**
SEWF	Word Formation Errors	**630**	**346**	**976**
SEWFSI	Bindings -s- incorrect	19	8	**27**
SEWFSI00	Bindings -s- incorrect	16	8	**24**
SEWFSM	Binding -s- missing	48	34	**82**
SEWFSM00	Binding -s- missing	18	34	**52**
SEWFHM	Hyphen missing	43	12	**55**
SEWFHM01	without an incorrect binding -s- and capital letter problem	29	12	**41**
SEWFHM02	with an incorrect binding -s-	0	0	**0**
SEWFHM03	with a capital letter problem	13	0	**13**
SEWFHM04	hyphen to be moved	1	0	**1**
SEWFHI	Incorrect hyphen	87	162	**249**
SEWFHI01	without a capital letter problem	82	147	**229**
SEWFHI02	with a capital letter problem	3	15	**18**
SEWFHI03	consonant to be removed	2	0	**2**
SEWFSW	Split words: several words => one word	182	65	**247**
SEWFSW01	2 lexical words	70	26	**96**

Code	Error Type Code Description	UNT	SvD	TOT
SEWFSW02	3 lexical words	12	3	**15**
SEWFSW03	compound with hyphen - proper noun	18	0	**18**
SEWFSW04	2 words - at least one word is non-lexical	48	16	**64**
SEWFSW05	2 lexical words + a hyphen after the first word or before the second word	9	1	**10**
SEWFSW06	compound with hyphen - with figures, abbreviations etc	6	12	**18**
SEWFSW07	compound with hyphen - common word	2	3	**5**
SEWFSW08	2 lexical words + a hyphen between them	4	2	**6**
SEWFSW09	2 words + capital letter problem	2	1	**3**
SEWFSW10	2 words + hyphen to be removed in the last word	2	1	**3**
SEWFSW11	compound with hyphen + capital letter problem	2	0	**2**
SEWFSW12	2 words - one needs correction	7	0	**7**
SEWFCW	Concatenated words: one word => several words	144	40	**184**
SEWFCW01	2 words - both correct	76	21	**97**
SEWFCW02	2 words - one word needs correction	3	3	**6**
SEWFCW03	erroneous compound with a hyphen - proper noun	6	1	**7**
SEWFCW04	2 words - with figures, letters etc	24	2	**26**
SEWFCW05	words with a common word part	10	1	**11**
SEWFCW06	erroneous compound with a hyphen - common word	13	9	**22**
SEWFCW07	2 words - foreign words	0	0	**0**
SEWFCW08	other	2	1	**3**
SEWFCW09	3 words - all correct	5	2	**7**
SEWFCW10	erroneous compound with a hyphen - comma to be inserted	4	0	**4**
SEWFMS	Misplaced space	3	0	**3**
SEWFMS00	Misplaced space	1	0	**1**
SEWFCO	Coordination with common word part	38	13	**51**
SEWFCO01	hyphen missing	14	2	**16**
SEWFCO02	incorrect hyphen; no common part exists	10	3	**13**
SEWFCO03	misplaced hyphen	0	2	**2**
SEWFCO04	space to be removed	4	6	**10**
SEWFCO05	space to be moved	10	0	**10**
SEWFAB	Abbreviations	57	10	**67**
SEWFAB00	Abbreviations	57	10	**67**
SEWFOP	Other word formation errors	9	2	**11**
SEWFOP00	Other word formation errors	9	2	**11**
SEHY	End of Line Hyphenation	**1263**	**0**	**1263**
SEHYHY	End of Line Hyphenation	1263	0	**1263**
SEHYHY00	End of Line Hyphenation	1263	0	**1263**
SEOS	(Other) Spelling Errors	**783**	**248**	**1031**
SEOSPN	Proper nouns	221	38	**259**
SEOSPN00	Proper nouns	221	38	**259**

Code	Error Type Code Description	UNT	SvD	TOT
SEOSFW	Foreign words	24	0	**24**
SEOSFW00	Foreign words	24	0	**24**
SEOSNB	Number expressions	4	2	**6**
SEOSNB00	Number expressions	4	2	**6**
SEOSOW	Other words	534	208	**742**
SEOSOW00	Other words	533	208	**741**
GP	Grammar Problems	**984**	**390**	**1374**
GPNP	Noun Phrase	**414**	**147**	**561**
GPNPAG	Agreement	107	48	**155**
GPNPAG01	number agreement in premodifier - noun	44	16	**60**
GPNPAG02	gender agreement in premodifier - noun	31	14	**45**
GPNPAG03	species agreement in premodifier - noun	19	13	**32**
GPNPAG04	gender agreement in noun - relative pronoun	0	0	**0**
GPNPAG05	number agreement in coordinated head nouns	1	0	**1**
GPNPAG06	number agreement in coordinated nouns in the genitive	0	0	**0**
GPNPAG07	number agreement in noun - postmodifier	2	1	**3**
GPNPAG08	number agreement in noun phrases in apposition	2	1	**3**
GPNPAG09	gender agreement in premodifier - adjective/participle functioning as a noun	0	1	**1**
GPNPAG10	species agreement in premodifier - adjective functioning as a noun	0	1	**1**
GPNPAG11	species agreement in premodifier - coordinated nouns	0	0	**0**
GPNPAG12	agreement in coordinated nouns	7	1	**8**
GPNPAG13	gender agreement in noun - postmodifier	1	0	**1**
GPNPGE	Gender	25	11	**36**
GPNPGE01	grammatical gender versus semantic gender	15	6	**21**
GPNPGE02	wrong gender of the indefinite article in the genitive premodifier	10	5	**15**
GPNPNB	Number	4	3	**7**
GPNPNB01	the plural => the singular	2	3	**5**
GPNPNB02	number problems between premodifier - noun: uncountable / countable	2	0	**2**
GPNPNB03	semantic number different from grammatical number	0	0	**0**
GPNPNB04	the singular => the plural	0	0	**0**
GPNPSS	Species	97	27	**124**
GPNPSS01	definite article missing or erroneous definite inflection in definite noun phrase with adjective premodifier	19	5	**24**
GPNPSS02	erroneous definite inflection after genitive attribute	4	5	**9**
GPNPSS03	indefinite article missing or definite article (and definite inflection) missing in indef. NP in the sing. without article	26	7	**33**
GPNPSS04	erroneous definite inflection before a necessary relative clause	3	0	**3**
GPNPSS05	erroneous definite inflection after certain pronouns and	4	0	**4**

Code	Error Type Code Description	UNT	SvD	TOT
	adjectives			
GPNPSS06	demonstrative pronoun / definite article should be removed	1	3	**4**
GPNPSS07	the indefinite article should be removed	10	1	**11**
GPNPSS08	double articles	4	0	**4**
GPNPSS09	definite inflection missing in noun phrase in PP	11	1	**12**
GPNPSS10	erroneous definite inflection in titles	3	1	**4**
GPNPSS11	other cases of erroneous definite inflection	7	3	**10**
GPNPSS12	other cases of missing definite inflection	5	1	**6**
GPNPCA	Case	56	23	**79**
GPNPCA01	common noun should be in the genitive case	18	4	**22**
GPNPCA02	proper noun should be in the genitive case	12	12	**24**
GPNPCA03	the genitive case => the basic case	23	5	**28**
GPNPCA04	error in forming the genitive case in word group	0	0	**0**
GPNPCA05	pronoun should be possessive pronoun	1	0	**1**
GPNPCA06	adjective used as a noun should be in the genitive case	1	2	**3**
GPNPCA07	other problems with case	1	0	**1**
GPNPAP	Adjective phrase	19	5	**24**
GPNPAP01	wrong word category of the premodifier	9	5	**14**
GPNPAP02	wrong type of adverb in premodifier	0	0	**0**
GPNPAP03	adjective used as a noun	0	0	**0**
GPNPAP04	wrong word category of the head adjective	6	0	**6**
GPNPAP05	other problems	3	0	**3**
GPNPPE	Participles	10	1	**11**
GPNPPE01	wrong verb form in premodifier	9	1	**10**
GPNPPE02	wrong verb form in postmodifier	1	0	**1**
GPNPNL	Numerals	2	4	**6**
GPNPNL01	approximate numbers	0	1	**1**
GPNPNL02	numeral missing in certain expressions	1	2	**3**
GPNPNL03	wrong word category	1	1	**2**
GPNPNN	Nouns	24	7	**31**
GPNPNN01	head noun missing	5	1	**6**
GPNPNN02	wrong word category	16	4	**20**
GPNPNN03	doubled noun	2	2	**4**
GPNPPN	Pronouns	23	3	**26**
GPNPPN01	relative pronoun missing	8	0	**8**
GPNPPN02	doubled pronoun	4	1	**5**
GPNPPN03	wrong type of pronoun	5	0	**5**
GPNPPN04	wrong word category	6	1	**7**
GPNPPN05	other problems	0	1	**1**
GPNPCP	Choice of preposition after a noun	19	4	**23**
GPNPCP01	noun + preposition + NP	10	1	**11**

Code	Error Type Code Description	UNT	SvD	TOT
GPNPCP02	noun + preposition + infinitive phrase	1	0	**1**
GPNPCP03	noun + preposition + subordinate clause	2	1	**3**
GPNPCP04	noun + PP + preposition + NP	0	0	**0**
GPNPCP05	noun + infinitive phrase; no preposition	1	0	**1**
GPNPCP06	noun + "att"-clause; no preposition	2	0	**2**
GPNPCP07	doubled preposition	2	1	**3**
GPNPCP08	noun + noun; no preposition	1	0	**1**
GPNPCP09	one preposition too many	0	1	**1**
GPNPCP10	noun + preposition + "att"-clause	0	0	**0**
GPNPMP	Preposition missing after a noun	7	3	**10**
GPNPMP01	noun + preposition + "att"-clause	0	0	**0**
GPNPMP02	noun + preposition	0	1	**1**
GPNPMP03	noun + preposition + NP	3	1	**4**
GPNPMP04	other missing prepositions	2	0	**2**
GPNPMP05	noun + preposition + clause	1	1	**2**
GPNPMP06	noun + preposition + infinitive phrase	1	0	**1**
GPNPNV	Other noun valency problems	1	1	**2**
GPNPNV01	noun + preposition + "att"-clause; "att" missing	0	1	**1**
GPNPNV02	wrong word category	1	0	**1**
GPNPNV03	question of repetition of preposition	0	0	**0**
GPNPCO	Coordination	11	1	**12**
GPNPCO01	conjunction missing	3	0	**3**
GPNPCO02	asymmetric coordination	1	1	**2**
GPNPCO03	comma replaced by coordinating conjunction	1	0	**1**
GPNPCO04	other coordination problems	6	0	**6**
GPNPWO	Word order	2	1	**3**
GPNPWO01	noun & adjective	1	0	**1**
GPNPWO02	noun & participle	1	1	**2**
GPNPOP	Other problems	5	3	**8**
GPNPOP00	Other problems	5	3	**8**
GPAP	Adjective Phrase	**5**	**3**	**8**
GPAPWC	Wrong word category	0	1	**1**
GPAPWC01	adjective => adverb	0	1	**1**
GPAPCP	Choice of preposition after an adjective	0	0	**0**
GPAPCP01	adjective + "att"-clause; no preposition	0	0	**0**
GPAPCP02	adjective + preposition + "att"-clause	0	0	**0**
GPAPCP03	adjective + infinitive phrase; no preposition	0	0	**0**
GPAPCM	Comparing "än"	2	0	**2**
GPAPCM00	Comparing "än"	2	0	**2**
GPAB	Adverb Phrase	**5**	**1**	**6**
GPABWM	Word missing	1	0	**1**

Code	Error Type Code Description	UNT	SvD	TOT
GPABWM00	Word missing	1	0	**1**
GPABDW	Doubled word	1	0	**1**
GPABDW00	Doubled word	1	0	**1**
GPABWO	Word order	1	0	**1**
GPABWO00	Word order	1	0	**1**
GPABOP	Other problems	2	1	**3**
GPABOP00	Other problems	2	1	**3**
GPPP	Prepositional Phrase	**114**	**38**	**152**
GPPPPR	Prepositions	104	35	**139**
GPPPPR01	preposition to be removed - should not be a prepositional phrase	5	4	**9**
GPPPPR02	one preposition too many	8	5	**13**
GPPPPR03	preposition missing	28	9	**37**
GPPPPR04	wrong preposition; choice of preposition	44	6	**50**
GPPPPR05	wrong word category	12	8	**20**
GPPPPR06	wrong preposition in coordinated PPs	1	0	**1**
GPPPPR07	doubled preposition	2	2	**4**
GPPPPR08	preposition missing in coordination of phrases - phrases of different types	3	1	**4**
GPPPPR09	preposition missing in coordination of phrases - phrases of the same type	0	0	**0**
GPPPPR10	comma => preposition	1	0	**1**
GPPPCO	Complements	10	3	**13**
GPPPCO01	erroneous construction after a certain preposition	6	3	**9**
GPPPCO02	consistency in complements	1	0	**1**
GPPPCO03	complement missing	2	0	**2**
GPPPCO04	med phrase	1	0	**1**
GPCN	Conjunctions and Conjunctive Adverbs	**50**	**21**	**71**
GPCNCM	Conjunction or conjunctive adverb missing	14	4	**18**
GPCNCM01	subordinating conjunction or conjunctive adverb missing	6	3	**9**
GPCNCM02	comma => subordinating conjunction	1	0	**1**
GPCNCM03	comma => coordinating conjunction	5	1	**6**
GPCNCM04	coordinating conjunction missing	2	0	**2**
GPCNCC	Complex conjunction	12	9	**21**
GPCNCC01	continuous	3	4	**7**
GPCNCC02	discontinuous	9	5	**14**
GPCNDW	Doubled conjunctions	8	2	**10**
GPCNDW01	coordinating conjunction	1	2	**3**
GPCNDW02	subordinating conjunction	7	0	**7**
GPCNEC	Erroneous conjunction	8	3	**11**
GPCNEC01	subordinating conjunction - no clause	7	2	**9**
GPCNEC02	subordinating conjunction - two conjunctions	1	1	**2**

Code	Error Type Code Description	UNT	SvD	TOT
GPCNWC	Wrong word category	7	3	**10**
GPCNWC01	pronoun	1	0	**1**
GPCNWC02	other	2	3	**5**
GPCNWC03	adverb	1	0	**1**
GPCNWC04	preposition	1	0	**1**
GPCNWC05	verb	1	0	**1**
GPCNWC06	adjective	1	0	**1**
GPVF	Verb Phrase in the Limited Sense	**79**	**33**	**112**
GPVFMF	Main verb in the finite form	27	7	**34**
GPVFMF01	presens + presens => presens	2	2	**4**
GPVFMF02	presens + preteritum => preteritum	0	1	**1**
GPVFMF03	preteritum + preteritum => preteritum	2	1	**3**
GPVFMF04	infinitiv => presens	11	3	**14**
GPVFMF05	supinum => imperativ	1	0	**1**
GPVFMF06	perfektparticip => preteritum	1	0	**1**
GPVFMF07	supinum => preteritum	0	0	**0**
GPVFMF08	infinitiv => preteritum	5	0	**5**
GPVFMF09	presensparticip => preteritum	1	0	**1**
GPVFTS	Temporal auxiliary verb in the finite form + Main verb in the supine	11	9	**20**
GPVFTS01	har/"hade" + presens => "har"/"hade" + supinum	2	1	**3**
GPVFTS02	har/"hade" + infinitiv => "har"/"hade" + supinum	0	1	**1**
GPVFTS03	wrong auxiliary verb	0	0	**0**
GPVFTS04	doubled auxiliary verb	0	2	**2**
GPVFTS05	missing auxiliary verb	4	4	**8**
GPVFTS06	wrong word category of the auxiliary verb	2	0	**2**
GPVFTS07	auxiliary verb omitted + perfect participle	0	1	**1**
GPVFTS08	har/"hade" + perfektparticip => "har"/"hade" + supinum	1	0	**1**
GPVFTS09	ha + supinum => "har"/"hade" + supinum	2	0	**2**
GPVFEP	Existential auxiliary verb in the finite form + Main verb in the perfect participle	1	1	**2**
GPVFEP01	är/"var" + presens =>"är"/"var" + perfektparticip	1	0	**1**
GPVFEP02	auxiliary verb missing	0	1	**1**
GPVFAI	Auxiliary verb in the finite form + Main verb in the infinitive	31	6	**37**
GPVFAI01	infinitiv + infinitiv => presens/preteritum + infinitiv	1	1	**2**
GPVFAI02	presens + preteritum => presens + infinitiv	3	0	**3**
GPVFAI03	presens + presens => presens + infinitiv	12	1	**13**
GPVFAI04	preteritum + preteritum => preteritum + infinitiv	0	1	**1**
GPVFAI05	supinum + infinitiv => presens/preteritum + infinitiv	0	0	**0**
GPVFAI06	missing infinitive	1	0	**1**
GPVFAI07	presens/preteritum + "att" + infinitiv; "att" should be removed	2	1	**3**
GPVFAI08	preteritum + supinum => preteritum + infinitiv	1	0	**1**

Code	Error Type Code Description	UNT	SvD	TOT
GPVFAI09	wrong word category of the infinitive	1	0	**1**
GPVFAI10	wrong word category of the auxiliary verb	1	1	**2**
GPVFAI11	preteritum + imperativ => preteritum + infinitiv	1	0	**1**
GPVFAI12	att + presens + infinitiv; "att" should be removed	1	0	**1**
GPVFAI13	missing auxiliary verb	2	0	**2**
GPVFAI14	doubled auxiliary verb	3	1	**4**
GPVFAI15	doubled infinitive or two infinitives after the auxiliary verb	2	0	**2**
GPVFAM	Combination of auxiliary verb + Main verb	2	5	**7**
GPVFAM01	two finite auxiliary verbs	2	2	**4**
GPVFAM02	infinitive + infinitive	0	0	**0**
GPVFAM03	supine + imperative/perfect participle	0	0	**0**
GPVFAM04	ha + perfect participle	0	1	**1**
GPVFAM05	ha doubled	0	1	**1**
GPVFAM06	wrong auxiliary verb	0	1	**1**
GPVFCO	Coordination of verbs	3	0	**3**
GPVFCO01	auxiliary verb + coordinated infinitives	1	0	**1**
GPVFCO02	bliva + coordinated perfect participles	0	0	**0**
GPVFCO03	coordinating conjunction missing	1	0	**1**
GPVFIP	Infinitive in infinitive phrase	2	4	**6**
GPVFIP01	presens => infinitiv	1	0	**1**
GPVFIP02	supinum => infinitiv	0	1	**1**
GPVFIP03	missing infinitive	0	1	**1**
GPVFIP04	wrong word category	0	0	**0**
GPVFIP05	perfektparticip => infinitiv	0	1	**1**
GPVFIP06	presens + infinitiv => infinitiv	1	0	**1**
GPVFOP	Other problems	2	0	**2**
GPVFOP00	Other problems	1	0	**1**
GPVV	Verb Valency	**151**	**88**	**239**
GPVVTR	Transitivity	5	5	**10**
GPVVTR01	transitive verb => transitive context	4	4	**8**
GPVVTR02	intransitive verb => transitive verb	1	1	**2**
GPVVCO	Copula	1	0	**1**
GPVVCO01	transitive verb => copula	1	0	**1**
GPVVRE	Reflexivity	4	3	**7**
GPVVRE01	reflexive => non-reflexive	1	0	**1**
GPVVRE02	non-reflexive => reflexive	0	3	**3**
GPVVRE03	one reflexive pronoun too many	1	0	**1**
GPVVRE04	wrong word - should be a reflexive pronoun	1	0	**1**
GPVVRE05	other problems	1	0	**1**
GPVVPC	Passive constructions	14	3	**17**
GPVVPC01	s-form: active voice => passive voice	5	0	**5**

Code	Error Type Code Description	UNT	SvD	TOT
GPVVPC02	s-form: passive voice => active voice	2	2	**4**
GPVVPC03	construction with "få"	0	0	**0**
GPVVPC04	active context => passive context	5	1	**6**
GPVVPC05	passive context => active context	1	0	**1**
GPVVOI	Object with infinitive	1	2	**3**
GPVVOI01	preteritum + NP + "att" + infinitive; "att" to be removed	0	1	**1**
GPVVOI02	other erroneous construction	1	1	**2**
GPVVPP	Prepositional phrase	1	0	**1**
GPVVPP01	PP missing	1	0	**1**
GPVVIP	Infinitive phrase	28	31	**59**
GPVVIP01	the infinitive mark "att" missing after the verb "komma"	5	14	**19**
GPVVIP02	the infinitive mark "att" missing - other cases	21	14	**35**
GPVVIP03	the infinitive mark "att" doubled	0	2	**2**
GPVVIP04	the infinitive mark "att" to be removed	1	0	**1**
GPVVIP05	wrong word	1	1	**2**
GPVVCL	Clause	8	5	**13**
GPVVCL01	att missing in "att"-clause	8	5	**13**
GPVVCL02	erroneous PP - preposition to be removed	0	0	**0**
GPVVID	Position holding "det"	9	1	**10**
GPVVID01	existential "det"	8	1	**9**
GPVVID02	det in emphatic constructions	1	0	**1**
GPVVVM	VF missing	20	8	**28**
GPVVVM01	verb inserted	6	6	**12**
GPVVVM02	wrong word category	14	2	**16**
GPVVNM	NP missing	22	12	**34**
GPVVNM01	subject in clause with inversion	12	2	**14**
GPVVNM02	subject in clause without inversion	6	7	**13**
GPVVNM03	subject in "att"-clause	3	3	**6**
GPVVNM04	subject in relative clause	1	0	**1**
GPVVCP	Choice of preposition/adverb after verbs	22	4	**26**
GPVVCP01	verb + preposition + NP	4	0	**4**
GPVVCP02	verb + preposition/adverb + "att"-clause	1	0	**1**
GPVVCP03	verb + preposition + infinitive phrase	0	0	**0**
GPVVCP04	verb + adverb + NP	1	0	**1**
GPVVCP05	verb + pronoun + "som" + NP	0	0	**0**
GPVVCP06	verb + NP; no preposition or adverb	0	1	**1**
GPVVCP07	one preposition too many	2	0	**2**
GPVVCP08	verb + noun + preposition	2	0	**2**
GPVVCP09	verb + reflexive pronoun + preposition + NP	3	0	**3**
GPVVCP10	verb + adverb + PP	2	0	**2**
GPVVCP11	verb; no preposition or adverb	4	0	**4**

Code	Error Type Code Description	UNT	SvD	TOT
GPVVCP12	verb + infinitive phrase; no preposition or adverb	1	0	**1**
GPVVCP13	verb + "att"-clause; no preposition or adverb	0	1	**1**
GPVVCP14	verb + reflexive pronoun + preposition/adverb	1	1	**2**
GPVVCP15	verb + adjective + NP; no preposition or adverb	1	0	**1**
GPVVCP16	verb + adverb + preposition + NP	0	1	**1**
GPVVMP	Preposition/adverb missing after verbs	9	8	**17**
GPVVMP01	verb + preposition/adverb + clause	1	0	**1**
GPVVMP02	verb + reflexive pronoun + noun + preposition + noun	1	0	**1**
GPVVMP03	verb + adverb + preposition + infinitive phrase ("att" may also be missing)	0	1	**1**
GPVVMP04	verb + reflexive pronoun + "som" + NP	0	1	**1**
GPVVMP05	verb + preposition/adverb + NP	2	2	**4**
GPVVMP06	verb + "som" + clause	0	1	**1**
GPVVMP07	verb + preposition + infinitive phrase ("att" may also be missing)	1	0	**1**
GPVVMP08	verb + reflexive pronoun + preposition + infinitive phrase	1	0	**1**
GPVVMP09	verb + reflexive pronoun + preposition + NP	1	1	**2**
GPVVMP10	verb + adverb + preposition + "att"-clause	0	1	**1**
GPVVMP11	verb + adverb + preposition + NP	1	0	**1**
GPVVMP12	verb + preposition + noun + infinitive phrase	0	1	**1**
GPVVMP13	verb + noun + preposition	1	0	**1**
GPVVRP	Repetition of preposition/adverb	1	4	**5**
GPVVRP01	phrases of the same type	1	3	**4**
GPVVRP02	phrases of diffferent types	0	1	**1**
GPPC	Pronoun case	**11**	**10**	**21**
GPPCSF	Subjective form correct	1	0	**1**
GPPCSF02	objective form => subjective form, followed by a relative clause	1	0	**1**
GPPCOF	Objective form correct	10	10	**20**
GPPCOF01	subjective form => objective form	0	2	**2**
GPPCOF02	subjective form => objective form, followed by a relative clause	10	8	**18**
GPAG	Agreement	**42**	**15**	**57**
GPAGNA	NP and AP - subject and complement	34	9	**43**
GPAGNA01	number in non-collective nouns	10	1	**11**
GPAGNA02	number in collective nouns	7	2	**9**
GPAGNA03	gender	9	6	**15**
GPAGNA04	gender in specific/general meaning	5	0	**5**
GPAGNA05	head noun/relative pronoun and Ap in relative clause	0	0	**0**
GPAGNA06	number in heading without copula	1	0	**1**
GPAGNA07	number in coordinated noun phrases	2	0	**2**
GPAGNO	NP and AP - object and complement	3	0	**3**
GPAGNO01	gender	2	0	**2**

Code	Error Type Code Description	UNT	SvD	TOT
GPAGNO02	species	1	0	**1**
GPAGAA	AP and AP - subject and complement	1	0	**1**
GPAGAA01	number	1	0	**1**
GPAGNE	NP and perfect participle - subject and complement	0	1	**1**
GPAGNE01	gender	0	1	**1**
GPAGNE02	number	0	0	**0**
GPAGNP	NP and NP - subject and complement	1	1	**2**
GPAGNP01	number	1	1	**2**
GPAGNS	NP and NP in "som" phrases - subject and complement	0	3	**3**
GPAGNS01	number	0	3	**3**
GPAGNS02	gender	0	0	**0**
GPAGNN	NP and NP - object and complement	0	0	**0**
GPAGNN01	number	0	0	**0**
GPRP	Referential problems	**26**	**10**	**36**
GPRPPN	Pronoun reference	9	4	**13**
GPRPPN01	anaforic reference	7	4	**11**
GPRPPN02	deictic reference	2	0	**2**
GPRPVF	Choice of VF	16	5	**21**
GPRPVF01	conditional subordinate clause	1	1	**2**
GPRPVF02	comparative subordinate clause	1	0	**1**
GPRPVF03	consistency	11	3	**14**
GPRPVF04	combination of verb form and temporal adverbial	2	1	**3**
GPWO	Word Order	**48**	**8**	**56**
GPWOIN	Inversion	5	0	**5**
GPWOIN01	inversion => not inversion	3	0	**3**
GPWOIN02	not inversion => inversion	2	0	**2**
GPWOIP	Inserted phrase	0	0	**0**
GPWOIP01	before => after the finite verb	0	0	**0**
GPWOAB	Adverb phrase	22	4	**26**
GPWOAB01	noun phrase	4	2	**6**
GPWOAB02	preposition	3	0	**3**
GPWOAB03	finite verb	11	1	**12**
GPWOAB04	infinite verb	2	0	**2**
GPWOAB05	adverb governed by a verb	2	0	**2**
GPWOAB06	prepositional phrase	0	1	**1**
GPWONP	Noun phrase	2	0	**2**
GPWONP01	reflexive pronoun	1	0	**1**
GPWONP02	infinite verb	1	0	**1**
GPWOPP	Prepositional phrase	7	0	**7**
GPWOPP01	infinitive phrase	1	0	**1**
GPWOPP02	finite verb	2	0	**2**

Code	Error Type Code Description	UNT	SvD	TOT
GPWOPP03	finite verb + adverb	1	0	**1**
GPWOPP04	prepositional phrase	1	0	**1**
GPWOPP05	finite verb + noun phrase	1	0	**1**
GPWOPP06	infinitive mark	1	0	**1**
GPWOOP	Other word order problems	12	4	**16**
GPWOOP01	både ... "och" ...	0	0	**0**
GPWOOP02	såväl ... "som" ...	2	0	**2**
GPWOOP03	other problems	7	1	**8**
GPWC	Wrong Word Category	**13**	**1**	**14**
GPWCAV	Adjective	3	0	**3**
GPWCAV01	verb	1	0	**1**
GPWCAV02	preposition	2	0	**2**
GPWCAB	Adverb	9	1	**10**
GPWCAB01	noun	1	0	**1**
GPWCAB02	verb	0	1	**1**
GPWCPN	Pronoun	1	0	**1**
GPWCPN01	other	1	0	**1**
GPOG	Other Grammar Problems	**26**	**15**	**41**
GPOGCO	Coordinations	2	0	**2**
GPOGCO00	Coordinations	2	0	**2**
GPOGWM	Word missing	2	1	**3**
GPOGWM00	Word missing	2	1	**3**
GPOGDW	Doubled words	7	1	**8**
GPOGDW00	Doubled words	6	1	**7**
GPOGHE	Heading	0	0	**0**
GPOGHE00	Heading	0	0	**0**
GPOGOP	Strange syntax and other grammatical problems	15	13	**28**
GPOGOP00	Strange syntax and other grammatical problems	14	13	**27**
PU	Punctuation Problems	**1009**	**468**	**1477**
PUES	End of Sentence Punctuation	**183**	**77**	**260**
PUESPM	Punctuation mark missing	61	27	**88**
PUESPM01	point / full stop	58	26	**84**
PUESPM02	question mark	2	1	**3**
PUESPM03	exclamation mark	1	0	**1**
PUESEC	Choice of end of sentence punctuation	40	8	**48**
PUESEC01	point => colon	9	2	**11**
PUESEC02	point => exclamation mark	2	0	**2**
PUESEC03	point => question mark	11	1	**12**
PUESEC04	colon => point	8	1	**9**
PUESEC05	exclamation mark => point	1	1	**2**

Code	Error Type Code Description	UNT	SvD	TOT
PUESEC06	question mark => point	3	0	**3**
PUESEC07	question mark => exclamation mark	1	0	**1**
PUESEC08	hyphen => point	1	0	**1**
PUESEC09	point => three dots	0	1	**1**
PUESEC10	comma => point	4	1	**5**
PUESEC09	comma => exclamation mark	0	1	**1**
PUESFS	Full stop together with quotation marks or parentheses	26	29	**55**
PUESFS01	citation within the sentence	3	8	**11**
PUESFS02	whole sentence within parentheses	8	3	**11**
PUESFS03	whole sentence is a citation	15	17	**32**
PUESFS04	parentheses within the sentence	0	1	**1**
PUESPT	One punctuation mark too many	19	3	**22**
PUESPT01	point + question mark => question mark	1	0	**1**
PUESPT02	comma + point => point	3	0	**3**
PUESPT03	point + point => point	12	2	**14**
PUESPT04	point + exclamation mark => exclamation mark	3	1	**4**
PUESNE	Not end of sentence	35	9	**44**
PUESNE01	incorrect point should be removed	24	3	**27**
PUESNE02	point => comma	5	3	**8**
PUESNE03	other	4	0	**4**
PUESNE04	question mark => comma	0	2	**2**
PUESNE05	exclamation mark => comma	0	1	**1**
PUESNE06	comma => question mark	2	0	**2**
PUESOP	Other end of sentence punctuation problems	2	0	**2**
PUESOP00	Other end of sentence punctuation problems	2	0	**2**
PUCP	Capital letter	**34**	**23**	**57**
PUCPPT	Point	15	11	**26**
PUCPPT00	Point	14	11	**25**
PUCPCN	Colon	10	7	**17**
PUCPCN01	capital letter => no capital letter	4	4	**8**
PUCPCN02	no capital letter => capital letter	6	3	**9**
PUCPQN	Quotation	3	0	**3**
PUCPQN01	no capital letter => capital letter	3	0	**3**
PUCPNO	Not beginning of sentence	6	5	**11**
PUCPNO01	capital letter => no capital letter	6	5	**11**
PUCO	Comma	**716**	**351**	**1067**
PUCOMC	Main clauses	65	160	**225**
PUCOMC01	coordination of main clauses	65	159	**224**
PUCOSC	Subordinate clause	202	69	**271**
PUCOSC01	necessary subordinate clause	91	12	**103**
PUCOSC02	not necessary subordinate clause	111	57	**168**

Code	Error Type Code Description	UNT	SvD	TOT
PUCOPH	Phrases / units	314	76	**390**
PUCOPH01	shared unit	24	6	**30**
PUCOPH02	necessary unit	55	12	**67**
PUCOPH03	inserted unit	193	51	**244**
PUCOPH04	wrongly assumed inserted unit	42	7	**49**
PUCOPA	Parts of phrases / units	53	9	**62**
PUCOPA01	erroneous comma in attributes	8	1	**9**
PUCOPA02	erroneous comma in "enumerations" ("uppräkningar")	7	2	**9**
PUCOPA03	comma missing in attributes	37	2	**39**
PUCOPA04	comma missing in "enumerations"	1	4	**5**
PUCOCC	Clarity criteria	67	22	**89**
PUCOCC00	Clarity criteria	3	5	**8**
PUCOIW	Comma instead of word	2	1	**3**
PUCOIW00	Comma instead of word	2	1	**3**
PUCOCO	Comma correct	5	8	**13**
PUCOCO01	semicolon => comma	3	3	**6**
PUCOCO02	dash => comma	0	4	**4**
PUCOCO03	other sign => comma	0	1	**1**
PUCOCO04	colon => comma	2	0	**2**
PUDW	Dash within the sentence	**9**	**7**	**16**
PUDWPH	Phrases / units	6	4	**10**
PUDWPH01	dash to be moved	2	0	**2**
PUDWPH02	dash to be inserted	4	3	**7**
PUDWPH03	dash to be removed	0	1	**1**
PUDWDC	Dash correct	2	3	**5**
PUDWDC01	colon => dash (with capital letter problem)	0	1	**1**
PUDWDC02	semicolon => dash	1	0	**1**
PUDWDC03	comma => dash	1	2	**3**
PUCN	Colon	**20**	**7**	**27**
PUCNCC	Colon correct	9	4	**13**
PUCNCC01	semicolon => colon	5	3	**8**
PUCNCC02	comma => colon	3	1	**4**
PUCNCM	Colon missing	10	3	**13**
PUCNCM01	before a citation, a quotation etc	4	1	**5**
PUCNCM02	introducing an explanation, an example etc	6	2	**8**
PUCNIC	Incorrect usage of colon	1	0	**1**
PUCNIC01	before a citation within quotation marks	0	0	**0**
PUCNIC02	other problems	1	0	**1**
PUSN	Semicolon	**2**	**3**	**5**
PUSNCS	Semicolon correct	1	1	**2**
PUSNCS01	colon => semicolon	0	0	**0**

Code	Error Type Code Description	UNT	SvD	TOT
PUSNCS02	comma => semicolon	1	1	**2**
PUSNSM	Semicolon missing	1	0	**1**
PUSNSM00	Semicolon missing	1	0	**1**
PUSNIS	Semicolon incorrect	0	2	**2**
PUSNIS00	Semicolon incorrect	0	2	**2**
PUOP	Other Punctuation Problems	**44**	**0**	**44**
PUOPEP	Erroneous punctuation in certain texttypes	36	0	**36**
PUOPEP01	point to be removed in bylines, captions, headings, etc	14	0	**14**
PUOPEP02	other problems	22	0	**22**
PUOPEM	Other erroneous punctuation marks	5	0	**5**
PUOPEM00	Other erroneous punctuation marks	5	0	**5**
GR	Graphical Problems	**670**	**120**	**790**
GRSC	Space	**411**	**59**	**470**
GRSCBA	Missing space around signs	63	3	**66**
GRSCBA01	dash within sentence ("tankstreck")	59	3	**62**
GRSCBA02	three dots	2	0	**2**
GRSCSB	Missing space before signs	27	16	**43**
GRSCSB01	dash within sentence ("tankstreck")	4	10	**14**
GRSCBA02	three dots	2	0	**2**
GRSCSB03	left parenthesis	2	6	**8**
GRSCSB04	left quotation mark	1	0	**1**
GRSCSM	Missing space after signs	131	18	**149**
GRSCSM01	comma	12	8	**20**
GRSCSM02	colon	3	3	**6**
GRSCSM03	point	35	1	**36**
GRSCSM04	dash before direct speech ("pratminus")	73	4	**77**
GRSCSM05	dash within sentence ("tankstreck")	5	1	**6**
GRSCSM06	three dots	1	0	**1**
GRSCSM07	right parenthesis	0	1	**1**
GRSCSM08	other signs	1	0	**1**
GRSCSM09	exclamation mark	1	0	**1**
GRSCSM10	right quotation mark	0	0	**0**
GRSCSL	Too little space	36	0	**36**
GRSCSL01	between words in a row	7	0	**7**
GRSCSL02	indent missing (at new paragraph)	29	0	**29**
GRSCSL03	before heading	0	0	**0**
GRSCST	Too much space	154	22	**176**
GRSCST01	after left bracket	0	4	**4**
GRSCST02	doubled space	47	1	**48**
GRSCST03	before ringht quotation mark	0	0	**0**

Code	Error Type Code Description	UNT	SvD	TOT
GRSCST04	before and/or after dash within the sentence ("tankstreck") meaning "till" and the like	10	7	**17**
GRSCST05	before and/or after hyphen in telephone numbers and the like	4	0	**4**
GRSCST06	too much indent	54	0	**54**
GRSCST07	before point	21	3	**24**
GRSCST08	before comma	14	4	**18**
GRSCST09	after left quotation mark	1	3	**4**
GRSCST10	after slash	1	0	**1**
GRSCST11	other erroneous space tokens	2	0	**2**
GRNL	New Line / Paragraph	**38**	**42**	**80**
GRNLNR	New line / paragraph to be removed	15	16	**31**
GRNLNR01	within a sentence	12	16	**28**
GRNLNR02	between colon and citation or the like	1	0	**1**
GRNLNR03	other	2	0	**2**
GRNLAB	Erroneously placed line break	23	0	**23**
GRNLAB01	abbreviation	13	0	**13**
GRNLAB02	number	9	0	**9**
GRNLAB03	other	1	0	**1**
GRNLNI	New line / paragraph to be inserted	0	13	**13**
GRNLNI00	New line / paragraph to be inserted	0	13	**13**
GRDS	Dash before Direct Speech	**17**	**3**	**20**
GRDSDM	Dash missing	2	1	**3**
GRDSDM00	Dash missing	2	0	**2**
GRDSIH	Incorrect hyphen	12	0	**12**
GRDSIH00	Incorrect hyphen	7	0	**7**
GRDSID	Incorrect dash	1	0	**1**
GRDSID00	Incorrect dash	1	0	**1**
GRDSIU	Incorrect underscore	2	2	**4**
GRDSIU01	also missing space	0	2	**2**
GRDSIU02	space not missing	2	0	**2**
GRDW	Dash within the Sentence	**91**	**0**	**91**
GRDWIH	Incorrect hyphen	89	0	**89**
GRDWIH01	meaning "till"	45	0	**45**
GRDWIH02	at inserted units	31	0	**31**
GRDWIH03	meaning "mot"	10	0	**10**
GRDWIU	Incorrect underscore	1	0	**1**
GRDWIU00	Incorrect underscore	1	0	**1**
GRDWID	Incorrect dash	1	0	**1**
GRDWID01	dash => hyphen	0	0	**0**
GRQM	Quotation marks	**57**	**10**	**67**
GRQMWQ	Quotation within a quotation	1	1	**2**

Code	Error Type Code Description	UNT	SvD	TOT
GRQMWQ01	single quotations marks correct	1	1	**2**
GRQMIS	Incorrect usage of single quotation marks	0	0	**0**
GRQMIS01	double quotation marks correct	0	0	**0**
GRQMTI	Quotation marks around titles, names etc	41	4	**45**
GRQMTI01	no quotation marks correct	26	1	**27**
GRQMTI02	double quotation marks correct	15	3	**18**
GRQMCI	Quotation marks citations etc	9	5	**14**
GRQMCI01	both quotation marks missing	2	1	**3**
GRQMCI02	left quotation marks missing	1	1	**2**
GRQMCI03	right quotation marks missing	4	2	**6**
GRQMCI04	quotation marks to be moved	2	0	**2**
GRQMCI05	direct speech - quotation marks to be removed	0	1	**1**
GRQMSK	Quotation after "så kallade" etc	5	0	**5**
GRQMSK01	quotation marks to be removed	5	0	**5**
GRQMOP	Other incorrect quotation marks	1	0	**1**
GRQMOP01	not in pairs - quotation mark to be removed	1	0	**1**
GRQMOP02	incorrect graphical signs used for single quotation marks	0	0	**0**
GRPA	Parentheses	**5**	**2**	**7**
GRPAPP	Parentheses not in pairs	0	0	**0**
GRPAPP01	parentheses of the same type	0	0	**0**
GRPAPR	Parentheses to be removed	1	1	**2**
GRPAPR01	both parentheses to be removed	0	0	**0**
GRPAPR02	left parenthesis to be removed	0	0	**0**
GRPAPM	Parentheses missing	4	1	**5**
GRPAPM01	both parentheses missing	1	1	**2**
GRPAPM02	left parenthesis missing	2	0	**2**
GRPAPM03	right parenthesis missing	1	0	**1**
GRTY	Typographical Errors	**49**	**0**	**49**
GRTYGC	Lower case and upper case characters	7	0	**7**
GRTYGC01	upper case characters => lower case characters	5	0	**5**
GRTYGC02	lower case characters => upper case characters	2	0	**2**
GRTYIT	Italic	10	0	**10**
GRTYIT01	emphasis	2	0	**2**
GRTYIT02	title	2	0	**2**
GRTYIT03	in caption	4	0	**4**
GRTYBO	Bold	13	0	**13**
GRTYBO01	bold missing	9	0	**9**
GRTYBO02	bold to be removed	4	0	**4**
GRTYFS	Font size	0	0	**0**
GRTYFS00	Font size	0	0	**0**
GRTYFO	Other font problems	19	0	**19**

Code	Error Type Code Description	UNT	SvD	TOT
GRTYFO00	Other font problems	19	0	**19**
GRTYMA	Margins	0	0	**0**
GRTYMA00	Margins	0	0	**0**
GROP	Other graphical errors	**2**	**4**	**6**
GROPHY	Hyphens	1	0	**1**
GROPHY01	hyphen to be removed	1	0	**1**
GROPAC	Accent	0	4	**4**
GROPAC00	Accent	0	4	**4**
GROPAP	Apostroph	0	0	**0**
GROPAP00	Apostroph	0	0	**0**
GROPOS	Other signs	1	0	**1**
GROPOS00	Other signs	1	0	**1**
SP	Style, Meaning, and Reference	**1049**	**397**	**1446**
SPPS	Preferred Spelling	**297**	**112**	**409**
SPPSPS	Preferred spelling	297	112	**409**
SPPSPS00	Preferred spelling	297	112	**409**
SPAB	Abbreviation	**144**	**87**	**231**
SPABCA	Choice of abbreviated form	98	82	**180**
SPABCA00	Choice of abbreviated form	97	82	**179**
SPABFE	Full expression preferred	46	5	**51**
SPABFE00	Full expression preferred	46	5	**51**
SPNS	Number Style	**153**	**97**	**250**
SPNSBS	Number beginning the sentence	0	0	**0**
SPNSBS00	Number beginning the sentence	0	0	**0**
SPNSSN	Small numbers	22	9	**31**
SPNSSN01	figures => letters	16	9	**25**
SPNSSN02	letters => figures	6	0	**6**
SPNSDN	Decimal numbers	24	2	**26**
SPNSDN01	measure	12	2	**14**
SPNSDN02	price, cost etc	10	0	**10**
SPNSDN03	time measure	1	0	**1**
SPNSDN04	space(s) to be removed at the comma	1	0	**1**
SPNSLN	Large numbers	40	58	**98**
SPNSLN01	space missing in large numbers	40	45	**85**
SPNSLN02	dot => space	0	13	**13**
SPNSLN03	other problems	0	0	**0**
SPNSAF	Approximate figures	4	3	**7**
SPNSAF00	Approximate figures	4	3	**7**
SPNSOR	Ordinals	0	0	**0**
SPNSOR00	Ordinals	0	0	**0**

Code	Error Type Code Description	UNT	SvD	TOT
SPNSYD	Year, date, time, etc	51	25	**76**
SPNSYD01	year beginning the sentence	0	7	**7**
SPNSYD02	definite article missing in date	8	14	**22**
SPNSYD03	definite article to be removed in date	10	0	**10**
SPNSYD04	definite form => indefinite form in date	5	0	**5**
SPNSYD05	alternative date expression	1	2	**3**
SPNSYD06	alternative time expression	17	0	**17**
SPNSYD07	alternative year expression	2	0	**2**
SPNSOP	Other problems	11	0	**11**
SPNSOP01	missing space(s) in telephone number	6	0	**6**
SPNSOP02	space misplaced in telephone number	4	0	**4**
SPNSOP03	space to be removed	1	0	**1**
SPWN	Correct Word Category but Wrong Word	**170**	**28**	**198**
SPWNAV	Adjectives	16	2	**18**
SPWNAV00	Adjectives	16	2	**18**
SPWNAB	Adverbs	13	1	**14**
SPWNAB00	Adverbs	13	1	**14**
SPWNCN	Conjunctions and conjunctional adverbs	14	5	**19**
SPWNCN00	Conjunctions and conjunctional adverbs	14	5	**19**
SPWNNN	Nouns	55	8	**63**
SPWNNN00	Nouns	55	8	**63**
SPWNPR	Prepositions	6	1	**7**
SPWNPR00	Prepositions	6	1	**7**
SPWNNN	Pronouns	55	8	**63**
SPWNNN00	Pronouns	55	8	**63**
SPWNVB	Verbs	58	6	**64**
SPWNVB00	Verbs	58	6	**64**
SPWNIN	Injectives	1	0	**1**
SPWNIN00	Injectives	1	0	**1**
SPCW	Choice of Words and Expressions	**142**	**27**	**169**
SPCWCW	Choice of Words and Expressions	139	26	**165**
SPCWCW00	Choice of Words and Expressions	139	26	**165**
SPCS	Choice of Signs	**40**	**8**	**48**
SPCSCD	Dash => colon	1	0	**1**
SPCSCD00	Dash => colon	0	0	**0**
SPCSDS	Colon => dash(es)	0	0	**0**
SPCSDS00	Colon => dash(es)	0	0	**0**
SPCSSL	Dash => slash	3	0	**3**
SPCSSL00	Dash => slash	3	0	**3**
SPCSPE	Points in list	34	7	**41**
SPCSPE00	Points in list	34	7	**41**

Code	Error Type Code Description	UNT	SvD	TOT
SPCB	Choice of Sentence Boundaries	**27**	**11**	**38**
SPCBOT	One sentence => two sentences	24	9	**33**
SPCBOT00	One sentence => two sentences	24	9	**33**
SPCBTO	Two sentences => one sentence	2	2	**4**
SPCBTO00	Two sentences => one sentence	2	2	**4**
SPSC	Choice of Syntactic Construction	**15**	**6**	**21**
SPSCOM	Omitted auxiliary "ha"	11	6	**17**
SPSCOM01	missing "ha" in perfect infinitive	10	4	**14**
SPSCOM02	missing auxiliary "har"/"hade" in subordinate clause	1	2	**3**
SPSCOR	Omitted of relative pronoun	0	0	**0**
SPSCOR01	som to be removed	0	0	**0**
SPSCSR	The adverb "så"	4	0	**4**
SPSCSR00	The adverb "så"	4	0	**4**
SPCN	Consistency	**0**	**2**	**2**
SPCNDI	Definite/indefinite inflectional form	0	0	**0**
SPCNDI00	Definite/indefinite inflectional form	0	0	**0**
SPCNNB	Number style	0	0	**0**
SPCNNB00	number style	0	0	**0**
SPCNNR	Number	0	0	**0**
SPCNNR00	number	0	0	**0**
SPCNSP	Spelling / word form	0	0	**0**
SPCNSP00	spelling / word form	0	0	**0**
SPRD	Redundancy	**32**	**14**	**46**
SPRDRD	Redundancy	32	14	**46**
SPRDRD00	Redundcy	32	14	**46**
SPRP	Referential Problems	**20**	**4**	**24**
SPRPNP	NP and NP	12	4	**16**
SPRPNP01	gender	1	2	**3**
SPRPNP02	number	5	1	**6**
SPRPNP03	semantic gender vs grammatical gender	2	0	**2**
SPRPNP04	noun => pronoun	3	1	**4**
SPRPNA	NP and AP	1	0	**1**
SPRPNA01	gender	1	0	**1**
SPRPCR	Clause and pronoun	0	0	**0**
SPRPCR00	Clause and pronoun	0	0	**0**
SPRPGS	General and specific reference	6	0	**6**
SPRPGS01	general => specific	6	0	**6**
SUMMA		**6798**	**2098**	**8896**

Code	Error Type Code Description	UNT	SvD	TOT

Bibliography and References

Gale, William A. & Church, Kenneth W. (1993): A program for aligning sentences in bilingual corpora. In: *Computational Linguistics*, 19(1), 1993.

Wedbjer Rambell, Olga (1998): *Error Typology for Automatic Proof-reading Purposes*. SCARRIE, Deliverable 2.1, version 1.1.