UPPSALA UNIVERSITY

SCARRIE

Project ref. no. *LE3-4239*

Project title *SCARRIE Scandinavian Proof-reading Tools*

Deliverable number *DEL 8.1.3*

Deliverable title ***Evaluation report for the Swedish prototype***

Number of pages *14*

WP/Task responsible *Department of Linguistics, Uppsala University, Box 527, S-751 20 Uppsala, Sweden.*

Author(s) *Anna Sågvall Hein, Leif-Jöran Olsson, Bengt Dahlqvist, Erik Mats, Department of Linguistics, Uppsala University, Box 527, S-751 20 Uppsala, Sweden*
E-mail: anna@ling.uu.se

EC Project Officer *Antonio Sanfilippo*

Keywords *Swedish prototype, user evaluation, system performance*

Abstract *The main objective of Work Package 8 of the Scarrie project is to test the system performance of the prototype. The testing is based on real-life text corpora. The Swedish validation was carried out in co-operation with the two Swedish newspapers Svenska Dagbladet and Upsala Nya Tidning.*

Executive summary

The main objective of Work Package 8 of the Scarrie project was to test the system performance of the prototype. The validation of the Swedish prototype has been carried out in co-operation with the two newspapers Svenska Dagbladet and Upsala Nya Tidning. The system's linguistic functionality was validated by running SCARRIE on a newspaper corpus of some 15,000 words from a randomly sampled set of articles. The validation process was supported by a Danish software, kraut, and by a Swedish software, scareval.

The spell checking recall obtained (98.0 % on lexical recall and 96.5% on error recall) seems to be very good. The vast majority of the real errors missed are outside the defined scope of the prototype, notably punctuation errors, in particular, errors in the use of the comma. Also a few errors in the use of the capital letter, and some typos (split words) are overlooked.

Spell checking precision is good, 41.3% good flags as compared to 20.0% in a comparative test with Swedish MS Word (see Evaluation Report - Del 7.2, and Dahlqvist, B., 1999). Many of the incorrect flags produced by SCARRIE are due to an unsatisfactory treatment of abbreviations, numbers, and typographical signs such as quotes. These areas have not been in focus in the project, but they should of course be covered in a commercial system. Unknown compound forms are as a rule successfully recognised by SCARRIE. Some over-generation remains to be handled, though.

Suggestion adequacy is the most problematic attribute for the spell checker of Swedish SCARRIE. Especially the high number of errors for which no replacement is suggested (60.6 %) is unsatisfactory. A large number of these are typos (some of them typos in compounds), again an area to which no great attention has been paid in the project. The measure of wrong replacements suggested (9.3 % of the cases) may also be improved. About half of these cases, in fact, are due to the fact that the user has wrongly split words at unexpected places.

The grammar checker of Swedish SCARRIE targets more than 30 error types. 8 of them were represented in the validation corpus. They refer to errors in the nominal phrase, errors in the verb phrase, word order errors, and erroneously split words. For these error types, an overall recall of 85.7 % and a precision of 92.3 % was obtained in a second run after a fine-tuning of the grammar. This seems to be quite good. It deserves mentioning that SCARRIE detected two error instances that were overlooked by the human proof-reader. Still additional validation studies on a larger corpus is called for, in specific for an assessment of the validity of the remaining error types targeted by Scarrie. The grammar checking validation results that were presented above are the first that we know of for Swedish. A grammar checking competitor was announced by the Finnish company LingSoft in October 1998 to be included in Office 2000. It will work together with the Word spell checker. As soon as it will be on the market, comparative studies will be made.

The Swedish Scarrie-prototype seems to be fairly fast, about 200 words/second, and using a little less than 25 MB of RAM. The memory usage will of cause change with the commercial Scarrie, adding the overhead of a user interface, but a total memory consumption not exceeding 32 MB seems reasonable.

Precision and recall are important keys to success for a system like this, but equally important is the interface and the way the flaggings and their replacements and comments are presented to the user. The Swedish user emphasised the need for convenient means for accepting or disregarding the suggestions or diagnoses offered by the system. They also proposed that the system be interconnected to a word processor application but running in its own window with three different panes (for target segment, suggestion, and explanation, respectively). The system should remember earlier responses to the same input throughout the session.

The lexical information must be easy to maintain; this issue has already been taken care of for Swedish Scarrie by the creation of a lexical database with a graphical interface *ScarrieLex* comprising all the lexical resources used by the system.

Uppsala University
Department of Linguistics

LE3-4239 SCARRIE PROJECT
Deliverable 8.1.3
Final

99-06-04

System Evaluation Report for Swedish Scarrie

Anna Sågvall Hein, Leif-Jöran Olsson, Bengt Dahlqvist, Erik Mats

Abstract

The main objective of Work Package 8 of the SCARRIE project is to test the system performance of the prototype. The testing is based on real-life text corpora. The Swedish validation was carried out in co-operation with the two Swedish newspapers *Svenska Dagbladet* and *Upsala Nya Tidning*.

Acknowledgements

The test corpora were manually proof-read by Daniel von Sydow from Upsala Nya Tidning (UNT), and Peter Johansson from Svenska Dagbladet (SvD). Bo Lövendahl (SvD) who is responsible for co-ordinating user contacts in the Scarrie project took an active part in the validation task and gave valuable input to the future design of the user interface. The users' work with the Scarrie prototype analyses for the system evaluation is thankfully acknowledged.

1 General

The main objective of Work Package 8 of the SCARRIE project was to test the system performance of the prototype. This was done in co-operation with the two user newspapers, Svenska Dagbladet and Upsala Nya Tidning. SCARRIE still being a research prototype, the evaluation had its focus on the linguistic functionality of the system and its efficiency, as well as on some aspects of the usability.

Validation was carried out by running Swedish SCARRIE on a corpus of newspaper text comprising some 15,000 current words from a randomly sampled set of articles. The total size of the validation corpus provided by the two newspapers amounts to 694,000 words.

For user interaction with the prototype a command line Unix-version was chosen as a more adequate alternative than the simple user interface that had been developed in the project; the reason for this being that the Swedish users don't as a rule work in a windows-based environment. The SvD user ran the validation texts on the computer at the developer's site, while the UNT user provided the developer with their material for him to run. With this approach user feedback could be taken into account, basically, on the run and the testing could be performed with continuously improved functionality.

2 Corpus statistics

Test corpus size

	UNT	SvD	Tot.
Words total	7853	6957	14810
Unique words	3218	2777	5995
Sentence total	628	546	1174
Average words per sentence	12.5	12.7	12.6

3 Approach and software

The linguistic functionality of the system is tested by comparing the system's log-file with the manually proof-read version of the text. The log-file of the Swedish Scarrie prototype contains errors detected by the spell checker, as well as errors detected by the grammar checker. The spell checking errors are marked by #n#, and the grammar checking errors are marked by error type code and the span in which the error was recognised as illustrated below.

Nu räcker det i princip med att den som tillhör #2#underrrepresenterat kön är tillräckligt kvalificerad

```
--> 2.underrepresenterat
```

för tjänsten för att positiv särbehandling skall kunna tillämpas.

An example of a spell checking error from the log-file

Ett viktigt motiv för de svenska EU-medlemskapet - också för socialdemokraterna - var just möjligheten att vara med i de fora där besluten fattas. $!43$

```
--> 43.intervall: 6 - 7 typ av fel: gpnpag01[1]: fel numerus
```

An example of a grammar checking error from the log-file

[1] See Appendix.

In the validation process, the log-file is compared with the manually proof-read version of the text and the original raw version. This process is supported by two kinds of validation software, one directed towards the output of the spell checker, i.e. Danish *kraut,* and one directed towards the output of the grammar checker, i.e. Swedish *scareval* (see Mats 1999).

Kraut searches the log-file for spelling errors (see e.g #2#underrrepresenterat in the example above), compares the results with the manually proof-read version of the text, and produces measures of recall, precision and suggestion adequacy (see e.g. --> 2. Underrepresenterat in the above example). In the evaluation, information about spell checking error type is also important. This is a problem, however, since the spellchecker does not distinguish between error types. All spelling errors were originally are marked by '#n#' in the log-fil, and this is what Kraut looks for. (On-going development of the spell checker at UU is directed towards a distinction between different spelling error types; so far, capital letter errors get a unique marking '$'). Thus the analysis of the different spelling errors had to be performed manually by an examination of bad flags, misses etc. in the log file. This is quite a laborious task, and the reason why only a minor part of the validation corpus (15,000 words) could be treated in the validation task. As a result of the manual analysis, the errors were then grouped into meaningful categories, and statistics were accordingly produced, see **4.1.2** below.

Scareval produces HTML tables of grammar error frequencies, based on the log-file, the user's corrected text, and the original raw text. Prior to the application of scareval, errors in the manually proof-read text have to be assigned error type codes in accordance with the Scarrie error typology; this text version is used by *scareval* as a facit ("golden standard"). For an analysis of the errors detected by Scarrie, scareval also presents the sentences in which the errors were found aligned with the original text. The program distinguishes between four different cases, i.e. 1) errors detected by SCARRIE and the human proof-reader and assigned the same error code, 2) errors detected by SCARRIE only, 3) errors detected by the manual proof-reader only, and, finally, 4) errors detected by the human proof-reader and SCARRIE and assigned different error codes. All the way through the error type codes are used. Examples of each case are given below. S denotes SCARRIE, H denotes Human, and F denotes Facit. 066 denotes the UNT corpus, and 000 the SvD corpus. Further, the sequential number of the sentence in the corpus is given.

1)
GPVFMF01 **066/570**
S: $GPVFMF01$Därefter var har han verksam som byggnadssnickare och även finsnickare.
H: $GPVFMF01$Därefter var har han verksam som byggnadssnickare och även finsnickare.
F: Därefter var han verksam som byggnadssnickare och även finsnickare.

2)
GPNPAG01 **000/570**
S: $GPNPAG01$Politiker och tjänstemän vill gå mer varsamt fram och ta hänsyn också till de positiva värdena som ofta finns i det här områdena.
H:
F:

3)
PUCOPH03 **000/322**
S: För riskpersoner till exempel äldre och hjärtsjuka med svåra symtom kan det dock vara rekommendabelt att uppsöka sjukhusvård.
H: $PUCOPH03$För riskpersoner till exempel äldre och hjärtsjuka med svåra symtom kan det dock vara rekommendabelt att uppsöka sjukhusvård.
F: För riskpersoner, till exempel äldre och hjärtsjuka med svåra symtom, kan det dock vara rekommendabelt att söka sjukhusvård.

4)
066/63
S: $SEOS02$ Jag lyssnade #idag (lördag) på #Ekonomiska klubben i radion, som gick igenom alla de #konjunkturprognoser som kommit från banker, konjunkturinstitut, kommun- och landstingsförbund, #arbetstagar - och #arbetsgivarförbund under våren.
H: $GRDWID01$Jag lyssnade idag (lördag) på Ekonomiska klubben i radion, som gick igenom alla de konjunkturprognoser som kommit från banker, konjunkturinstitut, kommun- och landstingsförbund, arbetstagar - och arbetsgivarförbund under våren.
F: Jag lyssnade i dag (lördag) på Ekonomiska klubben i radion, som gick igenom alla de konjunkturprognoser som kommit från banker, konjunkturinstitut, kommun- och landstingsförbund, arbetstagar- och arbetsgivarförbund under våren.

Validation measures are defined as follows:

Recall:

- no. of valid words accepted / total no. of valid words
- no. of errors flagged / total no. of errors

Precision:

- no. of correct flags / total no. of flags

Suggestion adequacy:

- no. of hits on initial suggestions / total no. of good flags
- no. of hits on non-initial suggestion / total no. of good flags
- no. of misses / total no. of good flags
- no. of times no suggestion is offered / total no. of good flags

4 Results

4.1 Spell checking

4.1.1 General functionality measures

Recall

	UNT	SvD	Tot.	%
Valid words	7797	6213	14010	
Valid words accepted	7651	6081	13732	98.0
Valid words rejected (bad flags)	146	128	274	2.0
Invalid words (real errors)	56	144	200	
Real errors spotted (good flags)	56	137	193	96.5
Real errors missed	0	7	7	3.5

Precision

	UNT	SvD	Tot.	%
Flaggings	202	265	467	
Good flags	56	137	193	41.3
Bad flags (false positives)	146	128	274	58.7

Suggestion adequacy

	UNT	SvD	Tot.	
Good flags	56	137	193	%
Hits on initial suggestion	19	37	56	29.0
Hits on non-initial suggestion	2	0	2	1.0
Misses (suggestions offered, none correct)	0	18	18	9.3
No suggestions offered	35	82	117	60.6

Many of the incorrect flags produced by SCARRIE were due to an unsatisfactory treatment of abbreviations, phrases, proper names, markup codes, numbers, and typographical signs.

4.1.2 Categorised functionality measures

Valid words rejected (bad flags)

Type of word	UNT	SvD	Tot.	%
Idiomatic expression (or part of one)	30	13	43	15.7
Compound form	8	31	39	14.2
Loan word	0	5	5	1.8
Numbers, dates, currency, units of measure	16	1	17	6.2
Proper names	35	13	48	17.5
Acronyms, abbreviations, symbols	12	24	36	13.1
Technical terms	1	0	1	0.4
Other	44	41	85	31.0
Total	146	128	274	100.0

Real errors missed

Type of error	UNT	SvD	Tot.	%
Capital letter error	0	3	3	42.9
Word formation error (hyphens, binding morphemes, etc.)	0	1	1	14.2
Spelling errors	0	0	0	0
Typing errors	0	3	3	42.9
Other problems	0	0	0	0
Total	0	7	7	100.0

Incorrect or no suggestions offered

Type of error	UNT	SvD	Tot.	%
Capital letter error	0	5	5	9.4
Word formation error (hyphens, binding morphemes, etc.)	0	10	10	18.9
Spelling errors	0	0	0	0

Typing errors	3	23	26	49.0
Other problems	0	12	12	22.6
Total	3	50	53	100.0

Problem areas that need to be further elaborated

1. Abbreviations
2. Colon-word formation
3. Mark-up code
4. Foreign words
5. Headlines
6. Proper names
7. Compounds
    - Position of the hyphen
    - Compounded phrases
    - Compounds with short segments – which segments should be allowed?

Unimplemented
Formatting of digital expressions

4.2 Grammar checking

4.2.1 Grammar checking measures

The grammar checker of Swedish SCARRIE targets more than 30 error types (see Appendix). 8 of them were represented in the validation corpus,i.e.

- errors in the nominal phrase: GPAPAG03, GPNPAG01, GPNPAG02, GPNPAG03
- errors in the verb phrase: GPVFMF01
- word order errors: GPWOAB03
- split words: SEWFSW01, SEOS02

For these error types, an overall recall of 76.9 % and a precision of 83.3 % was obtained. Two errors detected by SCARRIE were overlooked by the human proof-reader. These results seemed to be fairly good. Still an analysis of the contexts where SCARRIE had misbehaved indicated, that the figures might be further improved by fine-tuning the grammar. This was done, and a precision of 92.3 % and a recall of 85.7 % was achieved in a second run.

Two types of shortcomings with regard to the input were encountered during the validation process. One is due to cases where the spell checker has made an incorrect analysis of a word that is outside the dictionary. The other is due to wrong sentence segmentation. The system may be improved in both respects. However, in order to arrive at a reliable sentence splitting, the typographical markings in the newspaper articles have to be taken into account and presented to SCARRIE in a way that it can handle, e.g. in an SGML format.

Frequencies of error types represented in the validation corpus

First run

Code	Scarrie		Human		Precision	Recall
	Good	Bad	Good	Miss		
SEWFSW01	1	0	1	0	100.0 %	100.0 %
GPAPAG03	1	0	1	0	100.0 %	100.0 %
GPNPAG01	4	1	4	1	80.0 %	80.0 %
GPNPAG02	0	0	1	0	--	0.0 %
GPVFTS03	0	1	0	0	0.0 %	0.0 %
GPWOAB03	1	0	0	1	100.0 %	100.0 %
GPNPAG03	2	1	2	0	66.6 %	100.0 %
Sum:	10	3	10	2	76.9%	83.3 %

Second run

Code	Scarrie		Human		Precision	Recall
	Good	Bad	Good	Miss		
GPAPAG03	1	0	1	0	100.0 %	100.0 %
GPNPAG01	5	0	5	1	100.0 %	83.3 %
GPNPAG02	0	0	1	0	- %	0.0 %
GPNPAG03	1	0	1	0	100.0 %	100.0 %
GPVFMF01	1	1	1	0	50.0 %	100.0 %
GPWOAB03	1	0	0	1	100.0 %	100.0 %
SEWFSW01	1	0	1	0	100.0 %	100.0 %
SEOS02	2	0	2	0	100.0 %	100.0 %
Sum:	12	1	12	2	92.3 %	85.7%

The validity of the remaining error types currently targeted by SCARRIE has to be assessed by validation studies of a larger corpus.

The dominating error type outside the defined scope of SCARRIE is punctuation, notably errors in the use of the comma. Except for quite special contexts, the detection of errors in the use of the comma must be based on a reliable recognition of clause boundaries. This is feasible within the ScarCheck framework, but requires further development of the grammar to an extent that was outside the scope of the project. Another major break through into more error types would be feasible if valency information could be taken into account in a systematic way. So far, valency information in ScarCheck is limited to individual lexical items. Including valency aspects in a systematic way would require a substantial extension of the dictionary, also that outside the scope of the project, but inside the ScarCheck grammar checking strategy.

The grammar checking validation results that were presented above are the first that we know of for Swedish. A grammar checking competitor was announced by the Finnish company LingSoft in October 1998 to be included in Office 2000. It will work together with the Word spell checker. As soon as it will be on the market, comparative studies will be made.

5 Efficiency evaluation

The Swedish Scarrie-prototype seems to be fairly fast, about 200 words/second, and using a little less than 25 MB of RAM. The memory usage will of cause change with the commercial Scarrie, adding the overhead of a user interface, but a total memory consumption of not exceeding 32 MB seems reasonably ok.

6 Usability

Precision and recall are important keys to success for a system like this, but equally important is the interface and the way the flaggings and their replacements and comments are presented to the user.

The Swedish user emphasised the need for convenient means for disregarding, or accepting the suggestion or diagnosis or offered by the system. They also proposed that the system be interconnected to a word processor application but running in its own window with three panes for the target segment, the suggestion, and the explanation, respectively. The system should remember earlier responses to the same input throughout the session.

The lexical information must be easy to maintain.

References

Dahlqvist, Bengt (1999) *Protokoll över stavningskontroll med MS Word 97 på testtext fredag4.txt.* Uppsala universitet. Instituttionen för lingvistik. Arbetsrapport.

Mats, Erik (1999) *Programvara för automatisk utvärdering av Scarriesystemets prestanda*. Technical report. Uppsala University. Department of Linguistics.

Patrizia Paggio et al (1999) *Evaluation Report*. SCARRIE Project Report, Del. 7.2.

Wedbjer Rambell, Olga (1998) *Error Typology for Automatic Proof-reading Purposes*. SCARRIE Project Report, Del. 2.1.

Wedbjer Rambell, Olga, Dahlqvist, Bengt, Tjong Kim Sang, Erik and Hein, Nils (1998) *Error Database of Swedish.* SCARRIE Project Report, Del. 2.1.3.2.

Appendix

Error types targeted by Swedish Scarrie

SYSTEMATIC SPLIT COMPOUNDS: SEWF

SEWFSW01 Split compound	*Upplands kusten => Upplandskusten
SEWFSW13 Split compound	*IT fakulteten => IT-fakulteten

AGREEMENT ERRORS at CLAUSE LEVEL: GPAG

GPAGNA01 wrong number in the complement	*Tävlingen blev väldigt besvärliga.
GPAGNA03 wrong gender in the complement	*LO-distriktet i Stockholm är negativ och poängterar vikten av att alla elever uppnår Högskolekompetens.

ERRORS IN CONJUNCTIONS: GPCN

GPCNCC02	*Om glädjebeskedet som omvandlades till en chock som vände upp och ned på hela deras tillvaro och höll på att krossa såväl hälsa, äktenskap och ekonomi.

ERRORS IN THE NP: GPNP

GPNPAG01 Number agreement	*Efter förberedelser av sina nya utrikesminister, Mrs Albright, som hade ett möte med sin kollega Primakov, har den rullstolsbundne Clinton träffat Jeltsin i Helsingfors.
GPNPAG02 Gender agreement	*En eventuellt segerfest får vänta.
GPNPAG08 Number agreement: noun - apposition	*Thage G Pettersson har skyllt på sina företrädare Anders Björk.
GPPNPAG03 Wrong species in the head noun	*De kanske mest personliga områden är de som nu lyfts fram.
GPNPAG14 wrong species in certain adjectives	*Barnen får använda sin egna energi.

ERRORS IN THE AP: GPAP

GPAPAG01 Disagreement: parallel adjectives	*En upptrappad psykologiska krigföring väntar.
GPAPAG02 Disagreement: coordinated adjectives	*Saknade faktiskt och praktiska möjligheter att hävda sig.

ERRORS in PP: GPPP

GPPCOF01 Wrong pronoun case	*För de som verkligen använder katalogen var det bra.

ERRORS IN THE VF: GPVF

GPVFAI01 inf + inf => finite form + inf	*Om människor börja tro på en förändring, så blir allt bättre.
GPVFAM02 wrong verb form after modal	*Hur trygghet inte längre kan var statisk utan ligga i förnyelsen, utvecklingen och förändringen.
GPVFAM03 wrong verb form after auxiliary	*Polisen har hörde flera vittnen under kvällen och utredningen kommer att fortsätta under tisdagen.
GPVFIP01 Finite form after ”att”	*Han har lovat att i alla fall skall slå Turkiet.
GPVFMF01 Two finite verbs	*Det blev bytte dock namn i samband med den första privatiseringen under Thatcherepoken.
GPVFMF04 Infinite form in the predicate	*De avskedade kvinnorna få rådet att starta eget.
GPVFMF05 Supine instead of imperative	*Betänkt också de anläggningskostnader som tillkommer.
GPVFOP01 Double s-passive	*Saken har försökts tystas ner.
GPVFTS03 Double supine	*Vi hade velat sett en större anslutningstakt, säger Dennis.

WORD ORDER at CLAUSE LEVEL: GPWO

GPWOAB03 finite verb + adv => adv + finite verb in subordinate clauses	*Men vi måste ändå begränsa oss på grund av att det saknas framför allt tid i hallarna.
GPWOAB04 inf + adv => adv + inf	*Man kan tro inte sina öron.
GPWOIN01 inversion => no inversion	*Jag undrar vad gör de små busungarna.
GPWOIN02 no inversion => inversion	*Nu man kan testa de kommande versionerna av programvaran.

VERB VALENCY ERRORS: GPVV

GPVVIP01 "att" missing after some verbs	*Vad jag förstår kommer Hälsingborgshem skicka upp 12 miljoner till skatteministern.
GPVVIP02 "att" missing after preposition	*Vidare ska pengar omfördelas till bland annat satsningar på Internet för stödja myndigheters och företags miljöarbete.
GPVVIP03 "att" doubled	*Att Sveriges ekonomi är stark igen kommer att märkas i människors vardag och det kommer att att märkas i kampen för jobben.

GPVVIP04 "att" to be removed	*Sverige började att klassa kärnkraftsincidenter enligt den internationella standarden.
GPVVMV01 Finite verb missing	*Man kanske inte behov av större resurser.
GPVVPC05 Passive after some verbs	*Huset ämnar byggas.

PARENTHESES: GRPA

GRPAPP Parenthesis missing	* Nästa etapp innebar säkring av brottet, sprängning och utplanande av kalkmassorna (1994 hade 5 000 kubikmeter språngts!

PUNCTUATION: PUES

PUESEC03 Period instead of question mark	*Är det rättvist och solidariskt.