NLP for writing: What has changed?

Koenraad De Smedt (University of Bergen)

Workshop on NLP for reading and writing — Resources, algorithms and tools. Stockholm, Nov. 20, 2008

It might appear that few advances have been made in proofreading technology since the 1980s[1]. On the one hand, spelling and grammar checking have become standard features in many kinds of applications that involve writing. On the other hand, a number of advanced research ideas and results from the 1980s do not seem to have been applied or further pursued in newer research. While there is continued research activity in the area of NLP for writing, the scale of projects in this area is not what it used to be. The present moment is therefore an opportunity to look back and reflect on what has been done so far and what has changed[2].

In the 1980s, several academic and commercial research groups in NLP started to turn their attention to automatic proofreading or *text critiquing*. One of the earliest large scale projects was the Writer's Workbench (Macdonald et al., 1982), followed by IBM's EPISTLE project (Heidorn et al., 1982), continued as CRITIQUE (Richardson and Braden-Harder, 1988), which was intended to check and correct the spelling, grammar and style of business letters in English. CRITIQUE uses a parser and grammar of English with relaxation and backoff, and applied lexical substitution to easily confused words. Figures 1 and 2 present screenshots from IBM terminals showing CRITIQUE feedback on mistakes in a business letter.

ESPRIT project OS-82 'Intelligent Workstation' was one of the earliest European applied IT projects that included the development of a proofreading tool. Under the name Author Environment, the tool was targeted at business letters in Dutch and English. Like CRITIQUE, Intelligent Workstation used a grammar and a parser with relaxation to correct grammatical

[1] In his summary submitted to the present workshop, Sjur Nørstebø Moshagen writes *"Utviklinga av grunnleggjande språkteknologiske verkty for vanlege brukarar, slik som gode stavekontrollar og presis orddeling, har i praksis ikkje gått framover sidan 1980-talet."*

[2] The present contribution has a limited scope and does not intend to present an encompassing overview of past work.

errors. It combined grapheme-to-phoneme conversion with trigrams so as to find similar-sounding spellings (van Berkel and De Smedt, 1988) and it provided single-click consultation of a dictionary and encyclopedia. The most advanced functionality consisted of the production of textual variants, not only by finding synonyms and related words, but also by changing from singular to plural and from active to passive and vice versa. The necessary changes were propagated throughout the document by means of a *grammar spreadsheet*. Figures 3 to 6 show examples of interaction with the Author Environment.

In the 1990s, some new techniques were explored and new insights were gained. Vosse (1994) built further on some techniques from OS-82, resulting in the comprehensive CORRIE system for Dutch spelling and grammar checking, which was also used as the basis for the SCARRIE project, supported by the European Commission and aimed at Danish, Norwegian and Swedish. Both CORRIE and SCARRIE offer advanced compound analysis, which is very important for the targeted languages. Parsing at sentence level was also included and functional, but the parser was not disambiguating, so that the number of ambiguities in authentic text remain a problem. GRANSKA (Domeij et al., 2000) for Swedish concentrated on grammar checking, using an HMM disambiguating tagger, tokenizer and rules, and generated a lot of exciting research, not only on techniques but also on user acceptance.

In the 1990s, commercialization by Microsoft, Lingsoft and other companies began to take a hold. Microsoft developed a grammar API and started to provide comments through red squiggles, dialogue boxes and the now discontinued paperclip 'Clippy' with a speech bubble. However, part of the targeted application area was moving faster than the technology. By the turn of the millennium, the typing of business letters was no longer a major office chore. Today, formal business letters have to some extent been replaced by communication through new channels such as email and web-based interaction, while also SMS must be mentioned as a new medium and voice input is starting to become a plausible option. The need for basic spelling and grammar checking remains, so that these functions have also become available in email and browser text windows, but the need for advanced functions like the *grammar spreadsheet* no longer seem important enough to justify their further development. Dictionaries, thesauri and encyclopedias have become available for free online, and Google can often be useful to check a word's spelling. Translation and summarization systems are also available online.

While the original target for the early dedicated proofreading systems had disappeared, the interest in the relation between NLP and the writing process remained strong and was explored in different ways. Experience with

CRITIQUE had already revealed that different groups profit differently: non-professional writers reported that more than 80% of CRITIQUE's suggestions to them were correct or useful, against 41% of professional writers. Domeij (1998) conducted a study and found that such tools can have a positive effect, but different writers cope differently with these tools. On the one hand, studies like these emphasize the importance of a thorough evaluation of NLP tools for writing in practical use. On the other hand, the larger cognitive and societal context in which writing takes place means that we must also consider the promotion of writing ability in the context of language learning and teaching and in relation to language policy issues.

Language learning and teaching started to become a target for NLP for writing relatively early. Research in proofreading had soon emphasized the distinction between mechanical and cognitive errors. Since the latter are in an obvious relation to language ability as the result of learning, they can be the target of various learning and teaching schemes. On the one hand, second language learners with gaps in their knowledge of the language may benefit not only from corrections but also from additional explanatory material that comes with good proofreading systems. On the other hand, native language learners are sometimes insufficiently aware of homophones with different spellings in different grammatical contexts, e.g. Norwegian *å* vs. *og* or French verbal forms ending in *-er*, *-ez*, or *-é*.

In the early 1990s, the Dutch company Cognitech developed several systems for spelling and grammar learning. Among these, SPELRAAM focused on spelling, and especially homophones, in syntactic contexts. The system is targeted at native speakers of Dutch and uses a decision tree to make learners aware of the grammatical choices that influence a word form. Figures 7–9 are screenshots of this system.

More recently, dedicated writing tools for second language learners were developed that combine proofreading with targeted pedagogical components. The Grim system (Knutsson, 2005) is a prime example of this line of research. By targeting the system to a specific audience, it is easier to optimize its usefulness. This presupposes empirical studies of writing processes and problems. As more data is becoming available, a systematic study of spelling and grammar problems in authentic writing situations is becoming feasible. The ASK project (Tenfjord et al., 2006) has collected a large number of Norwegian essays by students of Norwegian as a second language. These have been carefully error-coded and made searchable. Figure 10 shows a selection of the corpus revealing adjective form errors, while figure 11 shows the different distribution of some error types among different learner groups.

The second link concerns language policy, especially for languages that have complicated spelling systems. Public bodies governing language policy

tend to be very interested in promoting good spelling practice among language users. It is interesting that in the preparations for the Dutch spelling reform in the 1990s, consideration was given to NLP applications that would handle this spelling. Ultimately a simplification was achieved by establishing a single official spelling for each word, replacing preferred and less preferred variants. The even more complicated variation in Norwegian presented a headache for SCARRIE. Eventually, the Norwegian partners in SCARRIE solved this by establishing a limited set of subnorms and enabling adherence to a chosen subnorm though sophisticated dictionary and grammar codings. In the wake of this research, attention was drawn to the complications of the subnorms and the fact that many allowed lexical variants do not appear to be ever used (Rosén, 2000). A simplification of the variation in Bokmål was adopted by Norsk Språkråd in 2005 and there are plans for further empirical investigations of the situation. It should also be mentioned that political priorities have spurred the development of special writing tools to promote the participation of people with language-related disorders in social communication. In Norway, companies like Include and LingIT have been active in the development of such tools.

In conclusion, I would like to observe, firstly, that NLP for writing has been a research field that has seen important shifts in its intended application environments during the past couple of decades. Secondly, there are links between NLP for writing and other fields that directly or indirectly benefit from this research or vice versa, including language learning and teaching and language policy. Finally, a holistic approach to writing is needed, where NLP research better interacts with the study of cognitive aspects of the writing process (including first and second language learning and language disorders) and with an investigation of the changing environments for written communication and our appreciation of correctly written texts also in the new media.

References

Domeij, Rickard. 1998. Detecting, diagnosing and correcting low-level problems when editing with and without compuiter aids. *Text Technology* 8(1):12–25.

Domeij, Rickard, Ola Knutsson, Johan Carlberger, and Viggo Kann. 2000. Granska: An efficient hybrid system for Swedish grammar checking. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (NoDaLiDa)*.

Heidorn, George E., Karen Jensen, Lance Miller, Roy Byrd, and Martin Chodorow. 1982. The EPISTLE text-critiquing system. *IBM Systems Journal* 21(3):305.

Knutsson, Ola. 2005. *Developing and Evaluating Language Tools for Writers and Learners of Swedish*. Ph.D. thesis, Kungliga Tekniska högskolan.

Macdonald, Nina H., L. T. Frase, P. Gingrich, and S. A. Keenan. 1982. The Writer's Workbench: Computer aids for text analysis. In *IEEE Transactions on Communication (Special Issue on Communication in the Automated Office)*, vol. 30, page 105.

Richardson, Stephen D. and Lisa C. Braden-Harder. 1988. The experience of developing a large-scale natural language text processing system: CRITIQUE. In *Proceedings of the 2nd Conference on Applied Natural Language Processing, Austin, TX, 9–12 February 1988*, pages 195–202.

Rosén, Victoria. 2000. Er norsk et naturlig språk? In Ø. Andersen, K. Fløttum, and T. Kinn, eds., *Menneske, språk og felleskap*, pages 157–173. Novus forlag.

Tenfjord, Kari, Paul Meurer, and Knut Hofland. 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1821–1824.

van Berkel, Brigit and Koenraad De Smedt. 1988. Triphone analysis: A combined method for the correction of typographical and orthographical errors. In *Proceedings of the 2nd Conference on Applied Natural Language Processing, Austin, TX, 1988*, pages 77–83. ACL.

Vosse, Theo. 1994. *The Word Connection: Grammar-based Spelling Error Correction in Dutch*. Enschede: Neslia Paniculata.

Figures

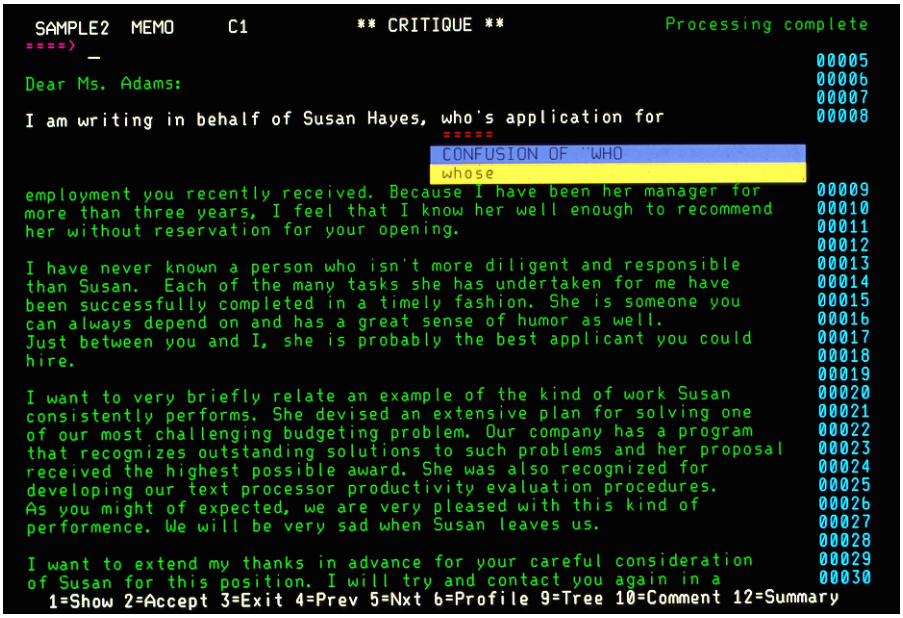

Figure 1: Screenshot of CRITIQUE proposing a correction (from a 35mm slide courtesy of Stephen Richardson).

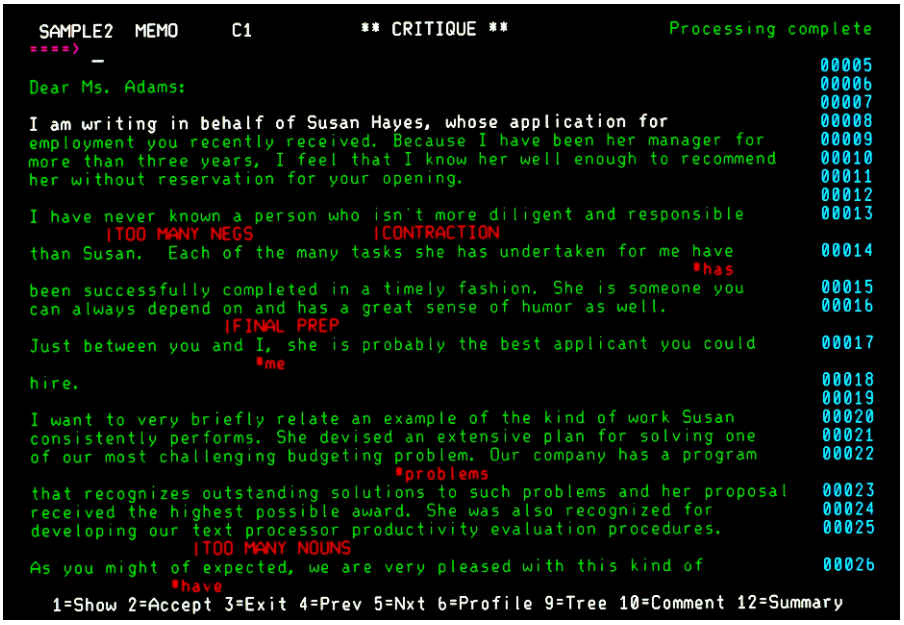

Figure 2: Screenshot of CRITIQUE highlighting suspected errors (from a 35mm slide courtesy of Stephen Richardson).

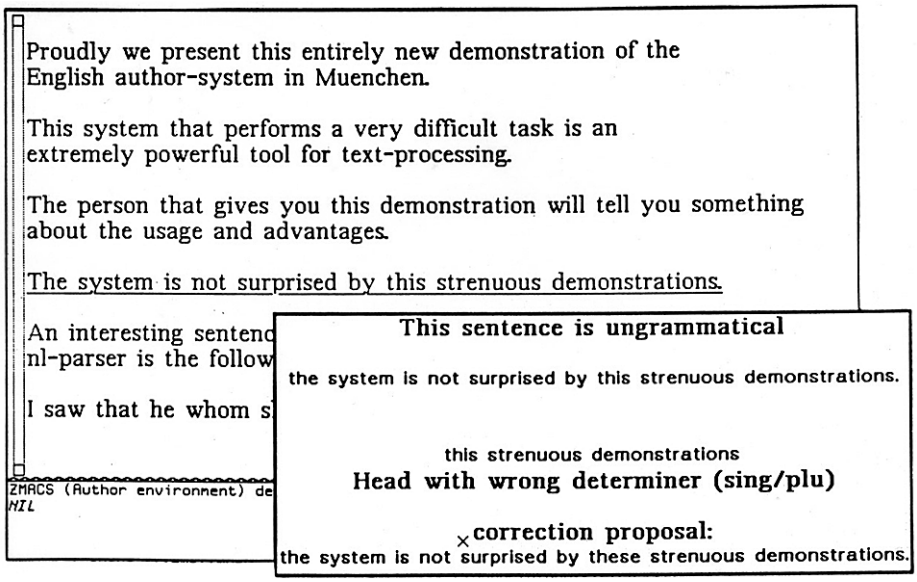

Figure 3: Screenshot of Author Environment proposing a diagnosis and correction.

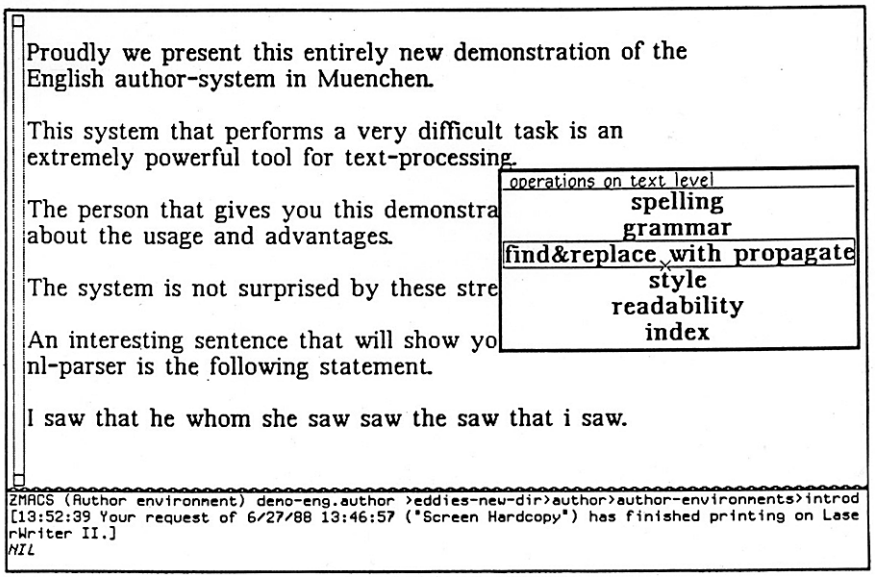

Figure 4: Screenshot of Author Environment menu including 'Find and replace with propagate'.

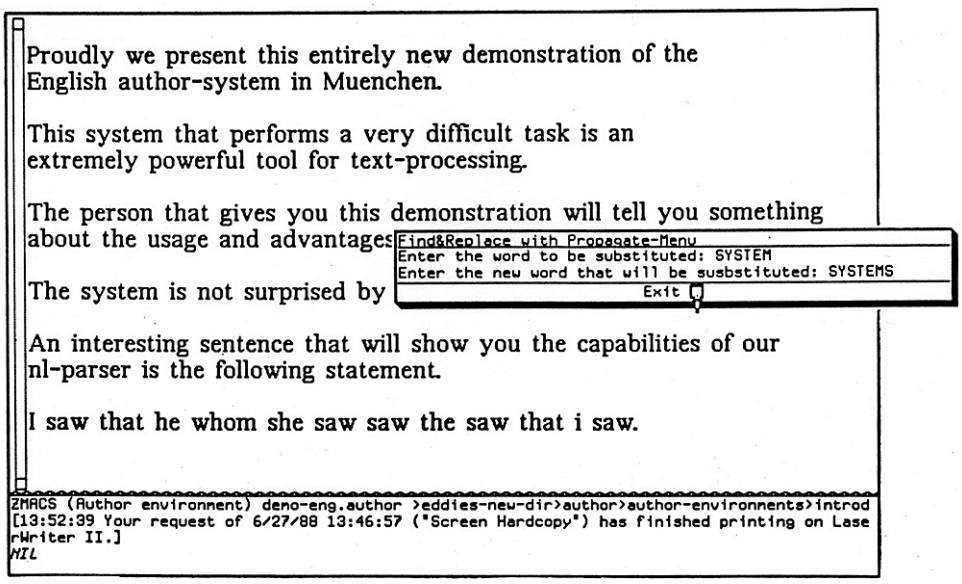

Figure 5: Screenshot of Author Environment where a word is being replaced.

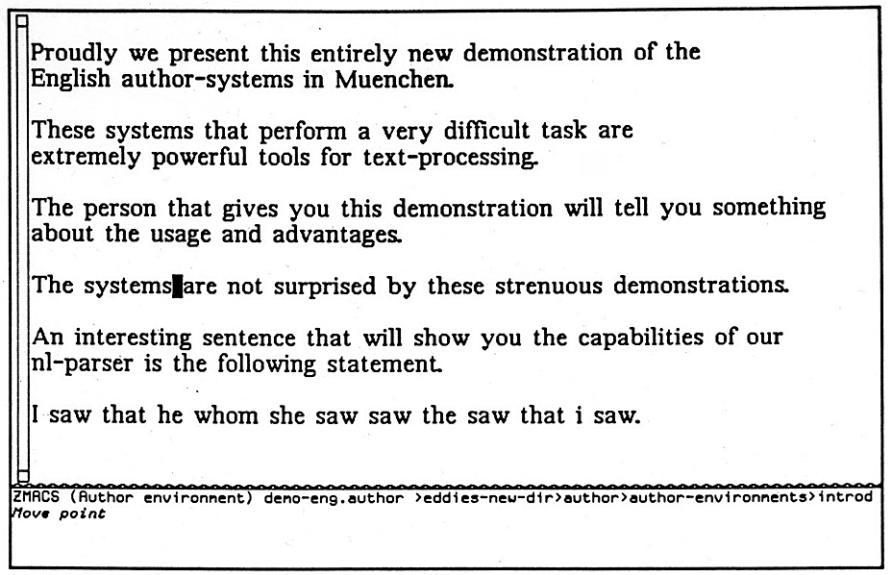

Figure 6: Screenshot of Author Environment showing the result of propagating a change from singular to plural.

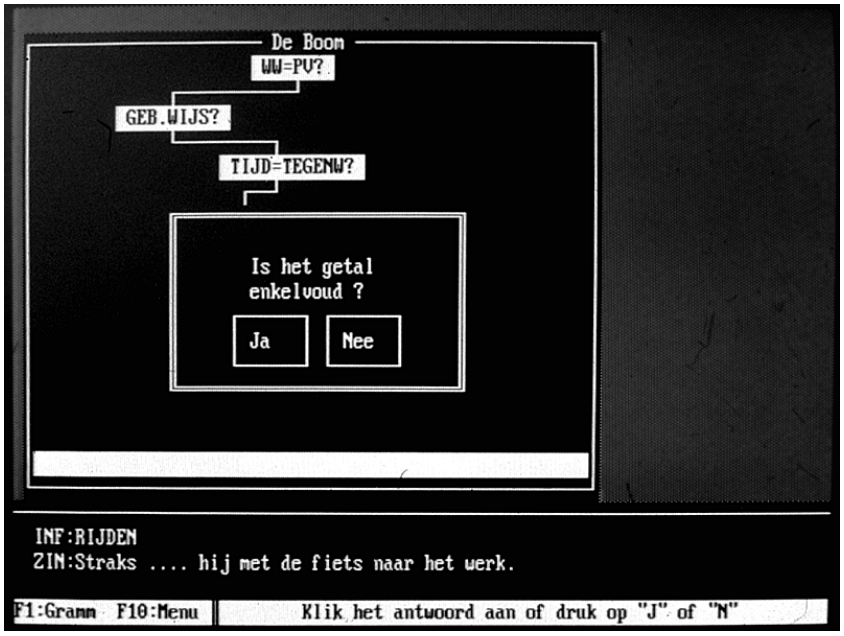

Figure 7: Screenshot of SPELRAAM showing how the user completes a decision tree (from a 35mm slide courtesy of Gerard Kempen).

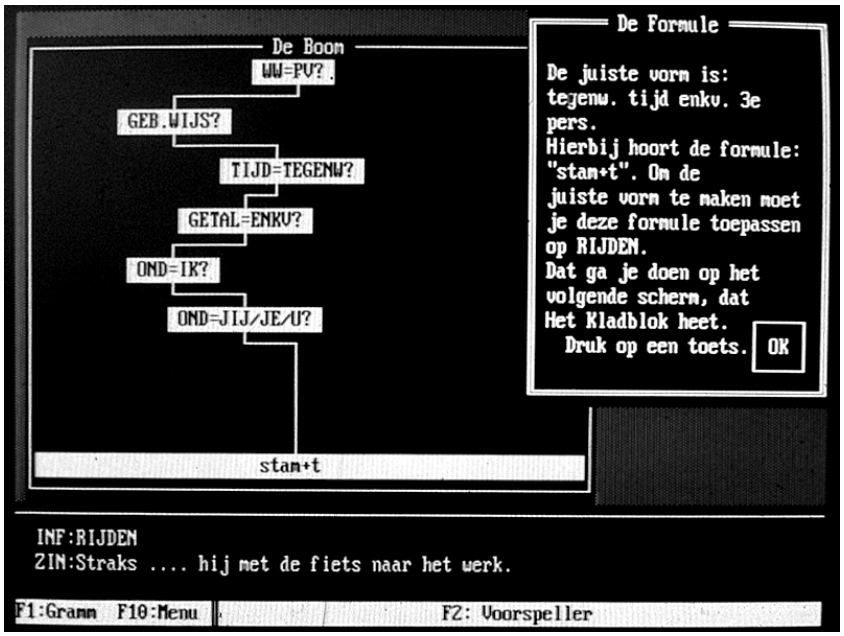

Figure 8: Screenshot of SPELRAAM showing a spelling rule for conjugation (from a 35mm slide courtesy of Gerard Kempen).

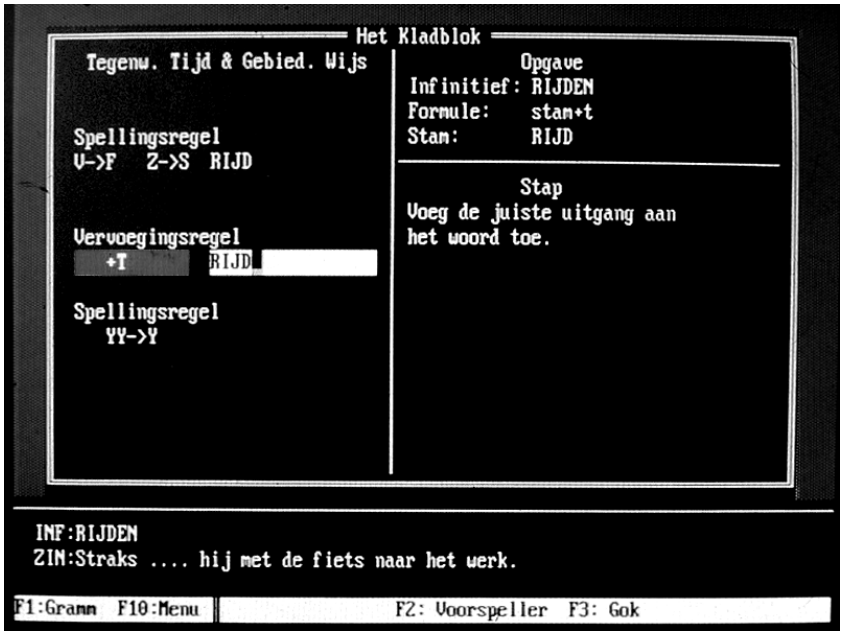

Figure 9: Screenshot of SPELRAAM giving spelling advice for conjugation by applying a rule (from a 35mm slide courtesy of Gerard Kempen).

Korpuset: ASK, **Søk:** @[type='.* F .*' & features='.* adj .*' & document!='no.*']

Treff 1 - 35 av 3911. | Vis kun ett treff per sic | neste 35 treff | treff: | KWIC | bredde: 250px | Last ned | Nytt søk | Hjemmeside

dokument	3 2 1 KWIC 1 2 3		feiltype	korreksjon
en200312-0338	år sonn til naboen min fikk mobiltelefon til	‹sic› **10års**	bursdag ‹/sic› sin. ‹/s› Hjemme, i STED, d	ORT F	|tiårs|tiårsdagen|
en200401-h0624	:astiske foreldre, kan det være f vanskelig.	‹s› ‹sic› **Adoptert**	‹/sic› barn har vanligvis lyst til å få vite on	F	|Adopterte|
po200401-h0612	nsket av mora si men også ble gitt bort av	‹sic› **adoptive**	foreldre ‹/sic›? ‹/s› Det kan skade barnet f	SPL F	|adoptivforeldre|adoptivfore
en200310-h0549	tror jeg at vi, enkelt og greit har blitt mer	‹sic› **aggresiv**	‹/sic› i våre kjøremåte og jeg tror at dette	ORT F	|aggressiv|aggressive|
ru200305-h0493	:ende vektet. ‹s› Vi er utsatt for et ganske	‹sic› **agressivt**	‹/sic› reklame av varer med høyt innhold	F	|agressiv|
vi200406-0852	for eksempel å gå på bussen uten kort er	‹sic› **akseptabel**	‹/sic›. ‹/s› Skilsmisse øker,... Mange vil no	F PUNCM	|akseptabelt|,|
ty200205-h0244	:p› ‹s› Jeg ser ingen sammenheng mellom	‹sic› **akti**	‹/sic› mennesker som ikke alltid har dårlig	ORT F	|aktiv|aktive|
ty199706-0965	å være nokk. ‹s› Og hvis man vil virk e lig	‹sic› **aktif**	‹/sic› vil bidra i dette projektet så kann m	ORT F	|aktiv|aktivt|
ru200205-h0290	· Den vinner som er den fortest, den mest	‹sic› **aktiv**	‹/sic›, som kan gjøre flere ting samtidlig. ‹	F	|aktive|
ne199706-0922	: kan være morsomt for å bruke denne tid	‹sic› **aktiv**	‹/sic›. ‹/s› Det er jo mye bedre at de kan g	F	|aktivt|
ty199706-0974	land. ‹s› Spesiellt i sommeren er vi veldig	‹sic› **aktiv**	‹/sic›. ‹/s› Vi bli delt i flere grupper fra ca.	F	|aktive|
en200305-h0441	nå lære på skolen hvordan å leve sunn og	‹sic› **aktiv**	‹/sic›. ‹/s› Hvis det er moro å være aktiv o	F	|aktivt|
en200305-h0441	nesker er mye opptatt, er de ikke fysikalsk	‹sic› **aktiv**	‹/sic›. ‹/s› Mange mennesker jobber i jobb	F	|aktive|
ty199706-0908	tetet med AI. ‹p› ‹s› Menneskerne som er	‹sic› **aktiv**	‹/sic› e med AI og andre menneskerettigh	F	|aktive|
en200310-h0579	ındrere. ‹s› Norske politiker må foreta noe	‹sic› **aktiv**	‹/sic› for å forandre på dette og gjevne de	F	|aktivt|
po200105-h0144	‹s› Ansatte i den bransjen bruker kroppen	‹sic› **aktiv**	‹/sic› hele dagen. ‹/s› De som er unge har	F	|aktivt|
ty200205-h0219	ny jobb. ‹s› For kvinner som ikke har vært	‹sic› **aktiv**	‹/sic› i yrkeslivet i flere år, er det nesten u	F	|aktive|
ne199706-0922	r og andre slektninger hvem som føler seg	‹sic› **aktiv**	‹/sic› kan gjøre noe. ‹/s› Kanskje du har s	F	|aktive|
ru200205-h0273	nå ha ansvar for barn? ‹s› De ville ha mer	‹sic› **aktiv**	‹/sic› liv, å bli samme medlem av samfunn	F	|aktivt|
se200105-h0147	: samfunn så må hele det samfunne være	‹sic› **aktiv**	‹/sic› med det. ‹/s› Folk kan kanskje bli hj	F	|aktivt|
ne200012-0396	‹s› Jeg er veldig opptatt med fotball, både	‹sic› **aktiv**	‹/sic› og passiv. ‹/s› Jeg leser nesten alt o	F	|aktivt|
se200205-h0285	en livsstil. ‹s› Den bruker både folk som er	‹sic› **aktiv**	‹/sic› og lever fort og folk som er passiv o	F	|aktive|
ru200310-h0554	› Når de vil finne seg en jobb, må de være	‹sic› **aktiv**	‹/sic› selv, vise interesse til arbeidsgiverer	F	|aktive|
en200306-0309	ne våre. ‹p› ‹s› Jeg liker dette for vi har et	‹sic› **aktiv**	‹/sic› sosial liv og vi finne vi føler ikke ens	F	|aktivt|
po200305-h0473	kke. ‹p› ‹s› Jeg tror at det er viktig å være	‹sic› **aktive**	‹/sic›. ‹/s› Folk skal drive med idrett. De sl	F	|aktiv|
en200210-h0331	erikansk kultur og mennesker. ‹s› De kan	‹sic› **aktivt**	bli en ‹/sic› del av samfunnet istedenfor b	F O	|aktiv|bli en aktiv|

Figure 10: Screenshot of KWIC search result for wrong forms of adjectives in ASK.

	target type	match lang	absolutt frekvens	relativ frekvens
☑	F	engelsk	634	0.16211
☐	F	tysk	484	0.12375
☐	F	spansk	480	0.12273
☐	F	nederlandsk	453	0.11583
☐	F	polsk	424	0.10841
☐	F	russisk	395	0.10100
☐	F	serbokroatisk	380	0.09716
☐	F	albansk	235	0.06009
☐	F	somali	227	0.05804
☐	F	vietnamesisk	207	0.05293

	target type	match lang	absolutt frekvens	relativ frekvens
☐	ORT	serbokroatisk	1976	0.12869
☐	ORT	polsk	1864	0.12139
☐	ORT	spansk	1825	0.11885
☐	ORT	engelsk	1746	0.11371
☐	ORT	albansk	1602	0.10433
☐	ORT	tysk	1589	0.10348
☐	ORT	russisk	1565	0.10192
☐	ORT	nederlandsk	1534	0.09990
☐	ORT	somali	1013	0.06597
☐	ORT	vietnamesisk	652	0.04246

Figure 11: Frequencies of two error types in ASK, grouped according to mother tongue: Wrong form of adjective (left); orthographical error (right).