Recreating Humorous Split Compound Errors in Swedish by Using Grammaticality

Jonas Sjöbergh and Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University
{js, araki}@media.eng.hokudai.ac.jp

Abstract

We present a program that recreates split compound errors with amusing effects in written Swedish. Two useful criteria for funniness is that the result should be grammatical and that the compound words should not be split into many short components.

1 Introduction

While humor is often used by humans, computational humor is an area of language processing that has seen relatively little attention. Most attempts have focused on language related humor, such as word play jokes. Languages with compounding where compound components are concatenated, i.e. combining several words into a single long word, have the possibility for compounding jokes, which seem fairly simplistic and thus achievable by computer.

Swedish is a language where compound components are concatenated and where compounding is very productive. Creating new quite long words from several shorter words is very common. If a word that should be written as one compound word is split up into several words, the meaning might be changed. Recently, many people have been annoyed by a perceived increase in this type of error in newspapers, signs and society. There was even a movement called *"skrivihop.nu"* (compound now!) which gathered over 25,000 members.

When a word is erroneously split so that the meaning is changed, the effect is sometimes amusing. Collections with examples of amusing mistakes from newspapers, restaurant menus, signs etc. are available on many humor sites on the Internet and seem to be an appreciated form of language humor.

We present a program that can recreate most examples from such humor collections given the intended (non-mistaken) text. It uses an automatic compound analyzer and an automatic grammar checking program. It turns out that grammaticality is a helpful criteria to tell if an erroneously split compound is amusing or not. To our knowledge, this is the first system for automatically generating this type of jokes, though other types of humor has been automatically generated (Binsted, 1996; Binsted et al., 2003; Stark et al., 2005; Yokogawa, 2001; Binsted and Takizawa, 1998; Sjöbergh and Araki, 2007). The system works on written Swedish, but a similar system for other languages with compounding, such as German or Norwegian, should be straightforward to create given a compound analyzer and a grammar checking program.

2 Description of the Program

Our program is quite simple. It uses a freely available program for analyzing compound words in Swedish (Sjöbergh and Kann, 2006) and an automatic grammar checking program for Swedish freely usable online (Domeij et al., 2000).

Given a sentence the program generates all possible compound analyses (according to the compound analysis program) of all the words. Each compound analysis of a word is then used to replace the word with the components of the compound separated into separate words. Compounds with several components or components that are in turn compounds need not be split fully, but can be. So given the word *"barnunderkläder"* (children's underwear), which can be analyzed as *"barn-under-kläder"* (children-below-clothes), the suggestions *"barn underkläder"* (children underwear), *"barnunder kläder"* (miracle-child clothes), *"barn under kläder"* (children under clothes) are generated.

Words with the character "-" are also processed in the same way, replacing the "-" with whitespace. The "-" has several uses, including some forms of conjunctions of compounds, e.g. *"hund- och katthår"* (dog [hair] and cat hair), and line breaks inside words.

From the complete sentence, new sentences are

generated by replacing one, two or three compounds with the suggested split variations. All combinations are generated. This of course generates very many variations for sentences with many compounds or compounds with many possible analyses or many components. Most of these are not amusing.

To remove unamusing sentences, two methods are used. The first is a powerful heuristic for removing over generation from the compound analyzer. If an analysis of a compound word results in more components than the analysis with the fewest components, the one with many components is ignored. Analyses with many short components are rarely amusing. There can of course be several different analyses with the same number of components.

The second method is to send the generated sentences to the automatic grammar checker. If the generated sentence is considered ungrammatical it is disregarded. The exception is the error type "split compound", which is ignored since the program is trying to add split compounds on purpose. The error type "no active verb" is also ignored, since many examples do not contain a verb, such as signs outside shops, e.g. *"Dansk fårost"* (Danish sheep cheese).

A small variation of this method is to remove suggested sentences for any grammar checking error, including the two exceptions above, if another suggestion for the same original sentence exists that has no errors at all. This was also tried.

For some sentences there is no suggested humorous variation from the program. This can be caused by either the compound analyzer failing to find any compounds to split in the sentence, or all suggestions being removed because they are considered ungrammatical.

3 Evaluation: Recreating Humor

A test corpus was created by searching the web for collections of amusing split compounds. Many examples are very similar, such as *"fryst kyckling lever"* (deep freezed chicken is still alive) with the intended sentence being *"fryst kycklinglever"* (deep freezed chicken liver) and *"djupfryst kyckling lever"* (same as previous example). In such cases, only one example from the set of variations of basically the same mistake was used in the corpus. This gave 230 examples of amusing split compounds from the real world. A few example sentences are given in Appendix A.

All examples were also fixed by hand, to recreate the intended meaning. Examples with only a word with no context are common in the split compound collections. These are often taken from signs with few words or leave out the original context because it was not very amusing. To get more information from the grammar checker, such examples were also given a simple context, so the grammar checker had something to base the analysis on. For example *"pris för slag"* (prices for being beaten), fixed to *"prisförslag"* (price suggestions), were at the same time put in the context *"Vi erbjuder: pris för slag."* (We are offering prices for being beaten.).

The program was run on the fixed sentences, with the goal of recreating the amusing split compounds. Not all split compounds can be recreated, since some of the corpus sentences contain words that are not covered by the compound analyzer. Several examples are words that are not strictly compound words, but were split into more than one word anyway by the original writer.

Generated sentences were classified as "Correct" if perfectly fitting the amusing original, "Almost" if almost fitting the original or "Wrong". "Almost" was used for sentences that found some but not all of the funny parts of a sentence, meaning that any compound which was split in the suggested sentence must be split in the same way as in the original, but if some compounds are correctly split it is an almost match if some compound is left untouched despite being split in the original. The sentence is thus a little bit funny, but has not achieved its full humorous potential. Sentences that are "wrong" are those that contain compounds that were split by the program that were either not split in the funny sentence or split in some other way than the one chosen by the program.

An example is *"datorn visar: fel meddelande, felkod 47"* (the computer is showing: the wrong message, error code 47) which is an almost match for *"datorn visar: fel meddelande, fel kod 47"* (the computer is showing: the wrong message, wrong code 47), both related to the error free sentence *"datorn visar: felmeddelande, felkod 47"* (the computer is showing: an error message, error code 47). Outputting *"datorn visar: fel med delande, fel kod 47"* (the computer is showing an error with dividing, wrong code 27) would be considered wrong, since the first compound is split in a different way than it should be.

The results are shown in Table 1. Both the few components heuristic and the grammar checking reduces the number of generated unamusing sentences considerably, while removing only one amusing sentence each. They can also be used together with even better effect, since the overlap in removed sentences is not very large. If suspicious suggestions are removed when a grammatical suggestion exists, many of the correct suggestions are removed. Thus the recall is decreased considerably, though precision is

	All	Gr.	Full Gr.	Few	Few, Gr.	Few, Full Gr.
Not found	16	17	31	17	18	26
Correct	214	213	199	213	212	204
Almost	43	40	37	40	40	37
Wrong	450	322	216	271	183	112
Recall (%)	93	93	87	93	92	89
Precision (%)	30	37	44	44	49	58
Precision (%, no A)	32	40	48	48	54	65

Table 1: "All", all suggestions from the compound analyzer. "Gr.", removing most ungrammatical suggestions. "Full Gr.", removing all ungrammatical suggestions. "Few", removing compound analyses with many components. "Precision, no A", the precision if sentences of type "Almost" are ignored.

increased.

It is also possible to increase the precision further by only using the suggestion with the most splits for each sentence. This reduces the number of generated suggestions drastically, since there is only one suggestion for each sentence, but while the precision rises to well over 70% the recall drops to about 70%, of course varying a bit depending on other settings.

Sentences for which the correct suggestion is not found generally contain split words that the compound analyzer does not consider to be compounds at all, often correctly. One example is *"Dagens prognos är öm som snö, slask och regn."* (Today's forecast is hurting like snow, slush and rain.), created from *"Dagens prognos är ömsom snö, slask och regn."* (Today's forecast is a mix of snow, slush and rain.). *"Ömsom"* is not a compound word, though it happens to become two words if a space is inserted in the right place. The fact that this word is not actually a compound word but was still split into two words by the original author and made sense is probably a large part of what makes this sentence funny (and thus made it appear in the joke collections the corpus is based on). This lack of recall could be mitigated by having a more aggressive compound analyzer, looking for any way to split a word that results in new words. This will however generate very many new suggestions. Most real life examples are split at the compound component borders, so the loss of recall from generating only such sentences is low.

Grammaticality is a useful filter. Only three of the real world sentences in the corpus are considered ungrammatical by the grammar checker. Requiring the amusing sentences to be grammatical is thus a good way to filter out bad suggestions with low risk of loosing actually amusing suggestions. It does however not remove as many bad suggestions as the heuristic for removing over generation from the compound analyzer. This heuristic is also very powerful, only removing one correct suggestion in the corpus while removing many faulty suggestions.

As a side note, the actual results are slightly funnier than what is suggested by Table 1. Several of the suggestions classified as "Wrong" are still funny, though in a different way than the real world example. Two examples are *"matt trea"* (fatigued letter three) instead of *"matt rea"* (fatigued sale) for *"mattrea"* (carpet sale) and *"brun stens batterier"* (the batteries belonging to a brown stone) instead of *"brunstens batterier"* (the batteries for when in heat) for *"brunstensbatterier"* (zinc-carbon batteries).

4 Evaluation: Creating New Humor

In the previous section, all sentences in the corpus had the potential to become funny. Taking a Swedish sentence in general, this is much less likely to be true. A (very) small test to get an indication of the potential on more general text was also performed. The front page of the Internet version of the Swedish newspaper Metro[1] was downloaded and the program run on the text. Removal of ungrammatical suggestions and suggestions with many short components was done.

The Metro text contains 335 sentences or phrases. From these, the program outputs 26 suggestions. Evaluating whether these are funny or not is of course subjective, though many cases where the program fails are easy to spot.

The following 10 suggestions were deemed (by the author, native speaker of Swedish) to be in the vein of the funny examples in the previous corpus, and somewhat funny (some are funny, some are very faintly funny):

- *En ja mes, men ingen Bond* (A yes saying wimp, but no Bond) *En James, men ingen Bond* (A James, but no Bond)

[1] http://www.metro.se/se/nyheter/

- *fort sätt* (quick way) *fortsätt* (continue)
- *Flykting mot tagande ska utredas* (Refugee against taking will be investigated) *Flyktingmottagande ska utredas* (Refugee welcoming procedures will be investigated)
- *Flyktingmott agande ska utredas* ("Butterfly refugee"-like beatings will be investigated) Original same as above.
- *Se fler bilder på Bil bo.* (See more pictures of Car nests.) *Se fler bilder på Bilbo.* (See more pictures of Bilbo (a lemur at the zoo).)
- *Tal man utmanar Bush i Syrien* (Speakable man challenges Bush in Syria) *Talman utmanar Bush i Syrien* (Speaker of the parliament challenges Bush in Syria)
- *Fri lans* (Free lance) *Frilans* (Freelance)
- *Quelle pastell le!* (Smile in the way of Quelle Pastell!) *Quelle pastelle!* (Quelle Pastelle!)
- *Play station 3 vapen i kampen mot Alzheimers* (The Playstation is 3 weapons against Alzheimer's) *Playstation 3 vapen i kampen mot Alzheimers* (The Playstation 3 as a weapon against Alzheimer's)
- *Pandaporren miss lyckades* (Miss Success, the panda porn) *Pandaporren misslyckades* (The panda porn failed)

Another six suggestions were deemed to also be in the vein of the funny corpus examples, though even less funny.

This means that a surprisingly high 10 or 16 of the 26 sentences were in some way joke like (though, as said before, the judgements are very subjective and were done by only one person). This means that about one suggestion in two was joke like, and about one sentence in 30 from the newspaper could be made into a joke. Of course, the attention seeking nature of a front page of a newspaper is still a fairly good source of funny formulations. Less successful results can probably be expected from other genres.

5 Conclusions

Amusing split compounds can successfully be recreated by a program, with very high recall. The program also generates sentences that contain split compounds that are not amusing. Grammaticality of the sentence is a good criteria for removing unamusing suggestions, filtering out many unamusing suggestions and only one of the amusing sentences. Another useful criteria is that the compounds should not be split into many short components. This also removes only one amusing suggestions while removing many unamusing ones. These two methods also complement each other, each removing many suggestions that the other method lets through. So, to be funny, be grammatical and don't overdo it!

With a recall of recreating 92% of the original amusing sentences, more than one suggestion in two is funny. At a quite high cost in recall, lowering it to 70%, it is possible to increase precision to over 75%.

While grammaticality seems to be almost a requirement for amusing split compounds, it is far from enough. Many texts can be split and still grammatical without amusing results. A small evaluation on the front page of a newspaper showed promising results on more general text, though. About half the generated suggestions were deemed amusing, and about one sentence in 30 from the newspaper could be turned into a joke.

Acknowledgments

This work has been funded by The Japanese Society for the Promotion of Science, (JSPS).

References

Kim Binsted and Osamu Takizawa. 1998. BOKE: A Japanese punning riddle generator. *Journal of the Japanese Society for Artificial Intelligence*, 13(6):920–927.

Kim Binsted, Benjamin Bergen, and Justin McKay. 2003. Pun and non-pun humour in second-language learning. In *Workshop Proceedings of CHI 2003*, Fort Lauderdale, Florida.

Kim Binsted. 1996. *Machine Humour: An Implemented Model of Puns.* Ph.D. thesis, University of Edinburgh, Edinburgh, United Kingdom.

Richard Domeij, Ola Knutsson, Johan Carlberger, and Viggo Kann. 2000. Granska – an efficient hybrid system for Swedish grammar checking. In *Proceedings of Nodalida '99*, pages 49–56, Trondheim, Norway.

Jonas Sjöbergh and Kenji Araki. 2007. Automatically creating word-play jokes in japanese. In *Proceedings of NL-178*, pages 91–95, Nagoya, Japan.

Jonas Sjöbergh and Viggo Kann. 2006. Vad kan statistik avslöja om svenska sammansättningar? *Språk och Stil*, 16:199–214.

Jeff Stark, Kim Binsted, and Benjamin Bergen. 2005. Disjunctor selection for one-line jokes. In *Proceedings of INTETAIN 2005*, pages 174–182, Madonna di Campiglio, Italy.

Toshihiko Yokogawa. 2001. Generation of Japanese puns based on similarity of articulation. In *Proceedings of IFSA/NAFIPS 2001*, Vancouver, Canada.

A Example Sentences

Here are some example sentences from the evaluation corpus, both the mistaken/funny versions and the intended versions are given.

- *Vi behöver tio öringar.* (We need ten salmon trouts.) *Vi behöver tioöringar.* (We need 10 "cent" coins.)
- *Vi skulle gärna vilja ha en flaggstång och några barn och vuxen cyklar också när vi ändå är på gång.* (Now that we are at it anyway, we would like a flagpole and some kids, and an adult is riding a bicycle.) *Vi skulle gärna vilja ha en flaggstång och några barn- och vuxencyklar också när vi ändå är på gång.* (Now that we are at it anyway, we would like a flagpole, bicycles for kids, and bicycles for adults.)
- *Vila under armarna mot skrivbordet.* (Rest below your arms on the desk.) *Vila underarmarna mot skrivbordet.* (Rest your wrists on the desk.)
- *Äldre dam eller herrcykel köpes billigt.* (Will buy cheaply: older lady or a bicycle for men.) *Äldre dam- eller herrcykel köpes billigt.* (Will buy cheaply: older bicycle, either men's or women's model.)
- *Behöver du extra knäck på lovet?* (Do you need more caramel during the vacations?) *Behöver du extraknäck på lovet?* (Do you need a part time job during the vacations?)
- *Brun hårig sjuk sköterska strök Herr skjorta.* (Brown, hairy and sick nurse ironed Mr. Shirt.) *Brunhårig sjuksköterska strök Herrskjorta.* (Brown haired nurse ironed a shirt [men's model].)
- *Dagens rubrik är svensk general agent för Kinaföretag.* (Today's headline is: Swedish general a spy for Chinese company.) *Dagens rubrik är svensk generalagent för Kinaföretag.* (Today's headline is: Swedish general representative for Chinese company.)
- *Dagens rätt är halvgrillad kyckling med kul potatis.* (Today's lunch is half grilled chicken with amusing potatoes.) *Dagens rätt är halvgrillad kyckling med kulpotatis.* (Today's lunch is half grilled chicken with round potatoes.)
- *Det finns en telefonservice som under normal arbetstid ger hjälp med svensk talande personal.* (We have a phone service that during normal working hours gives assistance with Swedish staff that can speak.) *Det finns en telefonservice som under normal arbetstid ger hjälp med svensktalande personal.* (We have a phone service that during normal working hours gives assistance with staff that can speak Swedish.)
- *Företaget bjuder samtliga anställda på Jullunch, för utom stående 50 kr* (The company treats all employees to a Christmas lunch, except people who are standing up who pay 50 kronor.) *Företaget bjuder samtliga anställda på Jullunch, för utomstående 50 kr* (The company treats all employees to a Christmas lunch, non-company people pay 50 kronor.)