An automatic analysis of Norwegian compounds

Janne Bondi Johannessen and Helge Hauglin[1]

1. Introduction

The University of Oslo is currently developing an automatic morphosyntactic tagger for Norwegian.[2] A very important module is one which can analyse compounds. Compounding is extremely productive in Norwegian, and it is futile to ever hope for a lexicon (dictionary) that will contain all or even most of the compounds that occr in actual texts. Since the tagger we are developing is based on the possibility of recognising words by the help of a lexicon, it is of great importance to have a module that recognises new compounds.

According to Munthe (1972), 10.4 per cent of all words in running text are compounds. Any text sample will contain a greatnumber of compounds. This statistics is true even for small samples. I took an arbitrary 440-word article from the newspaper *Aftenposten* from September this year, *Full penhet om passunion* (Full openness on passport union), and I quickly counted 47 compounds. Many of them are lexicalised, but there were 12 new compounds that could not be found in our lexicon (which contains a large number of compounds found in newspapers a few years ago). These included words like *flyktningepolitikken* (the refugee policy), *asylpolitikken* (the asylum policy), *passamarbeid* (passport cooperation), *justispolitikere* (politicians on legal matters), *hemmelighold* (secrecy keeping).

1 Johannessen is responsible for the linguistic aspects and Hauglin for the computational ones.

2 The tagger is one which disambiguates multiply tagged words, i.e. one whose goal is to solve ambiguities caused by homonymy in the language. There is widespread homonymy in Norwegian, not the least due to suffixes with multiple functions. We are using the Constraint Grammar formalism developed in Finland (see Karlsson et al. 1995).

We have developed a compound module that we will describe below.[3] We will not emphasise the computational aspects, but point at the various linguistic clues that constrain the compound analysis.

2. Types of compounds

From a morphological viewpoint, we can distinguish four different types of compounds. (In the following, a hyphen will be used to mark the joining between two stems.) The first type is one that is simply a juxtaposition of two stems:

(1) a. bil-syk — car sick
b. telefon-svarer — telephone answering machine

The second type contains a suppletive stem:

(2) a. kles-skap — clothes cupboard
Not: kl r- (clothes), klede- (cloth), klesplagg- (piece of clothing)

b. billed-spr k — picture language
Not: bilde- (picture)

The third type contains an epenthetic -s-:

(3) a. mors-binding — mother fixation
b. aluminiums-fabrikk — aluminium factory

The fourth type contains an epenthetic -e-:

(4) a. barne-trygd — child benefit
b. heste-ekvipasje — horse carriage

The four types of compund vary with respect to how widespread they are. The most common type is the stem compounding one. 75 per cent of all compounds belong to this type. Compounds containing an epenthetic -s- occur in 17 per cent of the cases, while those with

---

[3] There are still some parts that remain to be programmed, but they are marginal and do not affect evaluation results.

epenthetic -e- make up 8 per cent of all compounds. (These numbers are cited in Ak¿ 1989, and are due to Anne Golden.)

3. Why is it necessary to develop a separate compound-analyser?

Having seen that 75 per cent of all compounds are of a very regular type which can be analysed simply as stems being juxtaposed, one may ask whether there is a need for developing a potentially complicated compound-analyser. Could we not simply be contented with analysing the 75 per cent stem compounds, and make a simple guessing solution for the others? The answer is no, since there is no easy analysis even for lexical compounding - there is a need for an intelligent guesser even for this simples type of compounding.

Take as an example the word *kulturforskeren* the ethnographer . There are tens of possible analyses of this word, even if we restrict ourselves to requiring that each member apart from the last one must be uninflected (this is usually the case with compounds), and also require that each member is itself a possible Norwegian word:[4]

(5)	a.	kultur-forskeren	culture-the.researcher
	b.	kul-tur-forskeren	bump-trip-the.researcher
	c.	kultur-forske-ren	culture-research-clean
	d.	kultur-forsker-en	culture-researcher-one
	e.	kul-tur-forsker-en	bump-trip-researcher-one
		etc.	

It is important to find the correct analysis both from a semantic and a grammatical point of view. Semantically, we want to be able to find the correct interpretation:

(6)	melkeforretning:		
	a.	melke-forretning	milking shop
	b.	melk-e-forretning	milk shop

---

4 Here and below, we shall acstract away from the fact that homonymy makes many of the analyses ambiguous, so that *tur* means not only trip , but also luck , turn , celebrate , while *kul* means cool/nice/trendy , wait/rest in addition to bump .

Grammatically, a correct analysis is necessary to infer the correct part of speech and other features (the righmost member is the head and determines the features of the whole compound) :

(7) slottsvinduene:

a.	slottsvin-duene	castle-wine-the.pigeons
b.	slott-s-vinduene	castle-the.windows

4. A program for automatic compound analysis

The program we have developed has two main tasks. First, it must find all analyses of all compounds in a text. Second, it must be able to find the correct one amongst several possibilities. The evaluation results are very good: The analyser has made a wrong analysis in 1.1 per cent of the cases, and a partly wrong analysis in 1.3 per cent of the cases. We must add, however, that these results are not final, because when testing the analyser we used a lexicon that lacked certain stems that were used in the compounds of the test corpus. Since the analyser depends on a good lexicon, we decided not to count mistaken analyses based on this fact as actual errors.

As a wrong analysis we count any attempt by the analyser to suggest as its prioritised analysis one in which the final member is wrong. These analyses are unacceptable, since they give the wrong grammatical features for the compound as a whole.

(8) verdenscupseire:

a.	*verdenscup-s-eire	world cup become coated with verdigris
b.	verdenscup-seire	world cup victories

(9) vinduskitt:

a.	*vindu-skitt	window dirt
b.	vindu-s-kitt	window putty

(10) heleide:

a.	*hele-ide	whole idea
b.	hel-eide	fully owned

As a semi-wrong analysis we have counted compounds whose final member is correct even though the segmentation may give funny results with respect to meaning. Often, the semi-wrong analysis is due to the fact that the lexicon already contains several compounds:

(11) tilsynsorgan
   a. *til-synsorgan   to sight organ
   b. tilsyn-s-organ   supervision organ

(12) ikkespredningsavtale
   a. *ikke-spredningsavtale   non [expansion agreement]
   b. ikkespredning-s-avtale   [non expansion] agreeement

(13) takkformatentaler
   a. *takk-format-en-taler   thanks format one speeches
   b. takk-for-maten-taler   thanks for the.food speeches

5. What kind of knowledge does the compound-analyser use?

The compound-analyser makes use of different kinds of knowledge about the individual compound members. It uses grammatical knowledge, such as part of speech, inflection, type of stem (simple or complex) and derivation. It also uses phonological knowledge, such as phonotactic restrictions. In addition it takes quantitive considerations (length of compound members, number of members). Finally, it makes use of an extensive lexicon. Below, we shall see examples of the kind of judgements that the compound-analyser needs to make in order to determine the correct analysis.

5.1 Grammatical knowledge

5.1.1 Epenthetic -s- or lexical compounding?

The same phonemes that are used in epenthesis, -s- and -e-, are also extremely common in stems generally, and often cause ambiguity (the asterisk is for the incorrect, though in principle possible, interpretation). There is therefore a geunine need to distinguish between compounds containing stems plus epenthesis, and compounding with stems only (lexical compounding):

(14) lysmaskinen:
a. lys-maskinen the light machine
b. *ly-s-maskinen the shelter machine

(15) l¿vemanke:
a. l¿ve-manke lion mane
b. *l¿v-e-manke leaf mane

(16) ¿lskum:
a. ¿l-skum beer foam
b. *¿l-skum beer basin

Our investigations show that the following guideline can be used:

(17) Lexical compounding is preferable to compounding with epenthetic phones.

5.1.2 Epenthesis or lexical compounding - part of speech (a)

There are some exceptions to the above guideline:

(18) krigsmaske:
a. krig-s-maske war mask
b. *krig-smaske war (to) kiss

(19) aluminiumsnakke:
a. aluminium-s-nakke aluminium neck
b. *aluminium-snakke aluminium (to) talk

The crucial point with regard to such examples is that the incorrect analysis is one in which the second member is a verb:

(20) Epenthetic -s- is preferred to lexical compounding when the -s- can be ambiguous between epenthetic use and the first letter of a verbal last member.

5.1.3 Epenthesis or lexical compounding - part of speech (b)

There is a part of speech restriction on epenthetic -s-:

(21) oppslag poster, lexical entry
a. opp-slag up hit
b. *opp-slag up layer

(22) Epenthetic -s- can only follow noun stems.

5.1.4 Epenthesis or lexical compounding - morphological structure (a)

The morphological structure of the compound members also influences the choice between an analysis of a compound as containing epenthesis or not:

(23) lesesalsturer:
a. lesesal-s-turer reading room trips
b. *lesesals-sturer reading room moper

(24) storhavstang:
a. storhav-s-tang great sea seaweed
b. *storhav-stang great sea stick

(25) Epenthetic -s- is preferred to lexical compounding when the first member is itself a compound.

5.1.5 Epenthesis or lexical compounding - morphological structure (b)

There are restrictions with respect to the cooccurrence of epenthetic phones:

(26) barneskje children s spoon :
a. barn-e-skje child spoon
b. *barn-e-s-kje child kid (young goat)

(27) Epenthetic -s- cannot follow epenthetic -e- and vice versa.

5.1.6 Choosing between analyses - part of speech

There can be ambiguity also when there is no epenthesis involved. If two analyses have last members whose part of speech differ, this can be used. According to Heggstad (1982), the most frequent part of speech in Norwegian is the noun class - 67,5 per cent of all words.

(28) blomsterholder:

a. blomster-holder	flower holder (N)
b. *blomst-erholder	flower obtains (V present tense)

(29) hundyr:

a. hun-dyr	she animal (N)
b. *hund-yr	dog drizzle (V imperativ)

(30) If two analyses have the same number of members and there is no epenthesis involved, choose the one, if any, that is a noun.

5.1.7 Choosing between analyses - morphological structure

Sometimes two analyses are different only with respect to the morphological structure. This knowledge can be crucial:

(31) spisestueur:

a. spisestue-ur	[dining room] clock
b. *spise-stueur	eating [living-room clock]

(32) fagplanarbeid:

a. fagplan-arbeid	[technical plan] work
b. *fag-planarbeid	subject [plan work]

(33) If two analyses are equal with respect to epenthesis and part of speech, and one has a first member that is itself a compound, then choose that one.

5.2 Phonological considerations

5.2.1 Words with epenthetic -e-

Determining when a word contains epenthetic -e- and when an -e- is part of a stem will be very helpful in order to find the correct analysis. It turns out that there are quite strict limitations on when -e- can occur. Consider the following words:

(34) a. hest-e-ekvipasje   horse carriage
b. *tre-hest-e-ekvipasje   wooden horse carriage
c. tre-hest-ekvipasje

There rule turns out to be the following:

(35) Epenthetic -e- can only be attached to a stem that is monosyllabic.

This does not mean that epenthetic -e- must always make up the second syllable in a word, however:

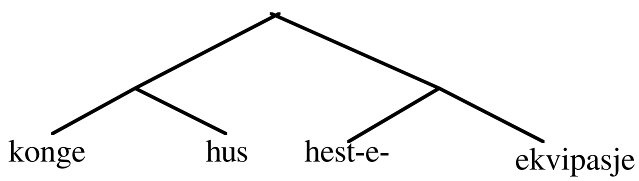
(36) a. konge-hus-hest-e-ekvipasje   royal house horse carriage
b. *konge-hus-hest-ekvipasje

It thus depends on what kind of internal structure there is prior to the one-syllable stem that precedes the -e-:

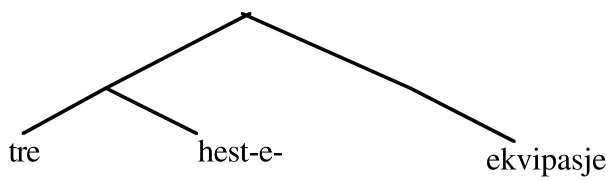
(37) Other possible stems can be prior to the stem preceding the -e-, if they do not form a compound with that stem.

This is illustrated in the below trees:

(38) a. OK



b. *



5.2.2 Epenthesis or lexical compounding - phonotactic issues

There are phonotactic restrictions for the occurrence of epenthetic -s-:

(39) buskspilling:
a. busk-spilling bush playing
b. *busk-s-pilling bush plucking

(40) Epenthetic -s- cannot occur after a sibilant or a final consonant sequence containing a sibilant (Ak¿ 1989)

However, this restriction can be overruled:

(41) eneb rbuskspilling:
a. eneb rbusk-spilling [juniper berry bush] playing
b. eneb rbusk-s-pilling [juniper berry bush] plucking

(42) ... except when the consonant sequence belongs to a compound.

5.3 Quantitative considerations

5.3.1 Choosing the best analysis of words with members unknown to the lexicon

There are a lot of loanwords, names and words belonging to various technical jargons that are not listed in most word-lists, however large. Such words do often occur, however, in actual texts, and often as part of compounds. Consider e.g. the following authentic examples, in which the first member is not in any of our large lexicons:[5]

(43) a. ibsenstykke Ibsen play
b. actionhelt action hero
c. gondoliermilj¿et the gondolier circles
d. geodatoppdrag geodat mission
e. couturevisningen the couture show
f. alkoroboten the alco robot

---

5 We use Bokm lsordboka (Landr¿ and Wangensten 1986) and IBM-ordlista (Engh 1989)

Many such words have several possible analyses, in which the first part of the word is unknown to the dictionary while the second word is known. Only one analysis should be analysed as correct:

(44)	a. couture-visningen	the couture show
	b. *couturev-isningen	the X frosting
	c. *couturevisnin-gen	the X gene

(45)	a. alko-roboten	the alco robot
	b. *alkoro-boten	the X fine

How can the compound-analyser know which analysis to prefer since the computer program has no world knowledge or semantic information? We have found one rule that applies generally:

(46) If the first member is unknown, choose the analysis with the longest last member.

5.3.2 Choosing the best analysis of words with members known to the lexicon

When the letters and letter-combinations in a word are very common the analyser may return several dozen possible analyses, all of which contain actual stems - sometimes with an epenthetic -s- or -e-:

(47)	a. lava-st¿vet	lava the.dust
	b. lava-s-t¿vet	lava the.nonsense
	c. la-va-st¿vet	let wade the.dust
	d. la-vas-t¿vet	let rubbish the.nonsense
	e. lav-as-t¿vet	lichen unrest the.nonsense
	f. lava-st¿-vet	lava steady knows
	g. lava-st¿v-et	lava dust eat
	h. lava-s-t¿-vet	lava melt knows
	i. la-va-st¿v-et	let wade dust eat
	j. lav-as-t¿v-et	lichen unrest nonsense eat
	etc.	

While these analyses are logically possible, the language user would never interpret them in all these ways. One reason is that compounds are always binary - they contain two members only (apart from possible epenthetic phones).

(48) Choose the analysis (or analyses) with the fewest compound members

6. Summary

We have presented an automatic compound-analyser that has been developed at the Text Laboratory at the University of Oslo. The analyser makes use of two lexicons - one for stems and one for fully inflected forms. It returns all possible analyses (whose members are stems and words found in the lexicons) of any given word not found in the lexicons, and then gives priority to one or a few analyses in particular. In order to choose one analysis before another, the analyser uses morphological (morphosyntactic and morphological), phonological and quantitative criteria.

References

Akø, Jørn-Otto. 1989. *Sammensatte ord: Bruken av s-fuge i moderne bokmål.* Master degree thesis, Department of Scandinavian languages, University of Oslo.

Heggstad, K. 1982. Norsk frekvensordbok. Universitetsforlaget, Oslo.

Karlsson, F., A. Voutilainen, J. Heikkil and A. Anttila. 1995. *Constraint Grammar.* A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin.

Landr¿, M. and B. Wangensteen 1986. *Bokm lsordboka*. Universitetsforlaget, Oslo.

Munthe, Synneve Kjuus. 1972. *Sammensatte ord. En kvantitativ undersøkelse av norsk litteratur og sakprosa.* master degree thesis, Department of Scandinavian languages. University of Oslo/Bergen. Not available for the public.

Contact addresses:

Janne Bondi Johannessen
The Text Laboratory
P.b. 1102 Blindern
N-0317 Oslo
Norway
e-mail: jannebj@ilf.uio.no

Helge Hauglin
The faculty IT department
P.b. 1102 Blindern
N-0317 Oslo
Norway
e-mail: helgeha@ilf.uio.no