Compound Nouns in English to Finnish Machine Translation

Hurskainen, Arvi

SALAMA - Swahili Language Manager
2018

Hurskainen , A 2018 , Compound Nouns in English to Finnish Machine Translation . in SALAMA - Swahili Language Manager : Technical reports on LT . Technical reports on language technology , no. 32 , SALAMA - Swahili Language Manager , Helsinki .

http://hdl.handle.net/10138/310991

unspecified
publishedVersion

Technical Reports in Language Technology
Report No 32, 20018
http://www.njas.helsinki.fi/salama

Compound Nouns in English to Finnish Machine Translation

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

The way how compound nouns are formed are different in English and Finnish. The differences occur in the way how members of the compound noun are ordered as well as in the way how they are written. English uses separate words, and Finnish joins the members into single words, without space in between. This is the general principle, but it also occurs that a noun compound in English is translated as separate words in Finnish. There are also several border cases, where Finnish compound nouns can be written together as single words or as separate words. The safest way to handle these would be to treat them as multiword expressions (MWE). Noun compounding is, however, such a productive phenomenon, that there is motivation to find also other, more general, methods for handling compound nouns. This report discusses those methods.

Key Words: *machine translation, compound nouns.*

1 Introduction

Verb English uses two methods in forming compound nouns. In one method, two or more nouns are written after each other as separate words, whereby the head of the compound is the last word, and the preceding word or words are modifying nouns. In case there are two or more modifying words, it is not always clear, whether the compound forms a hierarchical structure, where each modifying word modifies the whole structure to the right, or whether some modifying words form an internal compound, which is then joined to the whole compound. If this method for formulating compounds is used, it is sometimes impossible to know what the compound actually means.

The other method for formulating compound nous is to use a preposition, often *of* or *for*, for showing the relation between individual members. When this method is used, the confusion in relation to the structure of the compound is avoided.

Also Finnish uses two methods in noun compounding, but they are different than in English. Members of the compound are written after each other, and no prepositional structure is used. But Finnish uses both single-word compounds and compounds written as separate words. By using these two methods, Finnish avoids the ambiguity inherent in the first method of English. However, English provides no clue on whether the compound noun should be written as a single word or as separate words in Finnish. This is a major problem with this language pair. The safest way to handle the problem is to treat such compounds, which in Finnish are written as single words, as MWEs. Because noun compounding is very productive, there is motivation to seek also other methods for handling compound nouns. Below are examples on other methods.

2 Marking noun compound candidates in the lexicon

Because only the last member of the single-word compound in Finnish inflects, it is possible to add a separate entry for the preceding member(s) with the information, that this is a candidate for a non-final member in a compound, and that it should be joined to the following word as such, without possibility for inflection. The example in (1) illustrates the way how this is implemented.

(1)
```
"<apple>"
      "apple" { omena N11 } %A> N NOM SG
      "apple" { *apple N48 } %A> N NOM SG
      "apple" { omena-- COMP } %A> N NOM SG
"<juice>"
      "juice" { mehu N1 } %A> N NOM SG
      "juice" { mehu-- COMP } %A> N NOM SG
"<bottle>"
      "bottle" { pullo N1 } %NH N NOM SG
      "bottle" { --pullo N1 COMP } %NH N NOM SG
```

The word *apple* has three interpretations, one referring to the fruit *apple*, another referring to an American company, and the third referring to the possibility, that it is part of a compound. The words *juice* and *bottle* have two interpretations each. Note that the two dashes after *omena* and *mehu* indicate that they may be non-final members in a compound. The two dashes in front of pullo mean that the word can only be the last member in a compound noun. When we disambiguate the above example, we get the result as in (2).

(2)
```
"<apple>"
      "apple" { omena-- COMP } %A> N NOM SG
"<juice>"
      "juice" { mehu-- COMP } %A> N NOM SG
"<bottle>"
      "bottle" { --pullo N1 COMP } %NH N NOM SG
```

The final phase of the process is in (3).

(3)
omena-- mehu-- --pullo > *omenamehupullo*

The dashes show what should be done to the string. Note that the word breaks between *omena--* and *mehu--* and *mehu--* and *--pullo* are different. The word *mehu--* allows a noun on the right to be attached to it, but *--pullo* does not.

Let us see what happens with different combinations of these words (4).

(4)

```
(a)
"<juice>"
      "juice" { mehu-- COMP } %A> INDEF N NOM SG
"<bottle>"
      "bottle" { --pullo N1 COMP2 } %PCOMPL-S N NOM SG
(b)
"<apple>"
      "apple" { omena-- COMP } %A> INDEF N NOM SG
"<juice>"
      "juice" { mehu N1 } %PCOMPL-S N NOM SG
(c)
"<juice>"
      "juice" { mehu-- COMP } %A> INDEF N NOM SG
"<bottle>"
      "bottle" { --pullo N1 COMP2 } %PCOMPL-S INDEF N NOM SG
(d)
"<bottle>"
      "bottle" { --pullo N1 COMP2 } %PCOMPL-S INDEF N NOM SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<apple>"
      "apple" { omena-- COMP } %A> N NOM SG
"<juice>"
      "juice" { mehu-- COMP } %<P N NOM SG
(e)
"<bottle>"
      "bottle" { --pullo N1 COMP2 } %PCOMPL-S INDEF N NOM SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<juice>"
      "juice" { mehu-- COMP } %<P N NOM SG
```

The final result is in (5).

(5)
(a) omena-- mehu-- --pullo > omenamehupullo
(b) omena-- mehu > omenamehu
(c) mehu-- --pullo > mehupullo
(d) omena-- mehu-- --pullo > omenamehupullo
(e) mehu-- --pullo > mehupullo

If we change the order of words a bit and add a new word, we get as in (6).

(6)

```
(a)
"<juice>"
      "juice" { mehu-- COMP } %A> N NOM SG
"<apple>"
      "apple" { omena N11 } %NH N NOM SG
(b)
```

```
"<juice>"
      "juice" { mehu N1 } %A> N NOM SG
"<apple>"
      "apple" { omena-- COMP } %A> N NOM SG
"<basket>"
      "basket" { kori N5 } %NH N NOM SG
(c)
"<apple>"
      "apple" { omena-- COMP } %A> N NOM SG
"<basket>"
      "basket" { kori N5 } %NH N NOM SG
(d)
"<basket>"
      "basket" { kori N5 } %NH DEF N NOM SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<juice>"
      "juice" { mehu-- COMP } %A> N NOM SG
"<apples>"
      "apple" { omena N11 } %<P N PL NOM
```

In (a), the disambiguator has selected correctly the glosses. In (b), the word *mehu* cannot be joined to the following word, although it should. The reason is that the word *basket* does not have the tag for showing, that it could be the final member in the single-word compound. In (c), the disambiguation succeeds, because the structure has only two members. In (d), the prepositional structure also fails, because the word *basket* does not have the needed tag. The outcome is as in (7).

(7)
(a) mehu-- omena > mehuomena
(b) mehu omena-- kori > mehu omenakori
(c) omena-- kori > omenakori
(d) mehu-- omenoiden kori > mehuomenoiden kori

If we encode the word *basket* with the possibility to be the last member in a single-word compound, we get correct results as in (8) and (9).

(8)

```
(a)
"<Juice>"
      "juice" { mehu-- COMP } %A> CAPINIT N NOM SG
"<apple>"
      "apple" { omena-- COMP } %A> N NOM SG
"<basket>"
      "basket" { --kori COMP2 } %NH N NOM SG
(b)
"<Apple>"
      "apple" { omena-- COMP } %A> CAPINIT N NOM SG
"<basket>"
```

```
      "basket" { --kori COMP2 } %NH N NOM SG
(c)
"<basket>"
      "basket" { --kori COMP2 } %NH DEF N NOM SG
"<of>"
      "of" { NOGLOSS M-GEN } %<NOM-OF PREP
"<juice>"
      "juice" { mehu-- COMP } %A> N NOM SG
"<apples>"
      "apple" { omena-- COMP } %<P N PL NOM
```

(9)
(a) mehu-- omena-- --kori > mehuomenakori
(b) omena-- --kori > omenakori
(c) mehu-- omena-- --kori > mehuomenakori

Now we will test how the system works with various combinations in real sentences (10).

(10) Source text:
1. This is orange juice.
2. This is orange juice can.
3. This is orange juice bottle.
4. This is orange juice container.
5. This is a can of orange.
6. This is a can of orange juice.
7. This is a bottle of orange.
8. This is a bottle of orange juice.
9. This is a container of orange.
10. This is a container of orange juice.
11. This is a grape juice.
12. This is grape juice can.
13. This is grape juice bottle.
14. This is grape juice container.
15. This is a can of grape.
16. This is a can of grape juice.
17. This is a bottle of grape.
18. This is a bottle of grape juice.
19. This is a container of grape.
20. This is a container of grape juice.
21. This is carrot juice.
22. This is carrot juice can.
23. This is carrot juice bottle.
24. This is carrot juice container.
25. This is a can of carrot.
26. This is a can of carrot juice.
27. This is a bottle of carrot.
28. This is a bottle of carrot juice.

29. This is a container of carrot.
30. This is a container of carrot juice.
31. This is apple juice.
32. This is apple juice can.
33. This is apple juice bottle.
34. This is apple juice container.
35. This is a can of apple.
36. This is a can of apple juice.
37. This is a bottle of apple.
38. This is a bottle of apple juice.
39. This is a container of apple.
40. This is a container of apple juice.
41. This is pear juice.
42. This is pear juice can.
43. This is pear juice bottle.
44. This is pear juice container.
45. This is a can of pear.
46. This is a can of pear juice.
47. This is a bottle of pear.
48. This is a bottle of pear juice.
49. This is a container of pear.
50. This is a container of pear juice.
51. This is pumpkin juice.
52. This is pumpkin juice can.
53. This is pumpkin juice bottle.
54. This is pumpkin juice container.
55. This is a can of pumpkin.
56. This is a can of pumpkin juice.
57. This is a bottle of pumpkin.
58. This is a bottle of pumpkin juice.
59. This is a container of pumpkin.
60. This is a container of pumpkin juice.
61. This is in pumpkin juice.
62. This is in a pumpkin juice can.
63. This is in a pumpkin juice bottle.
64. This is in a pumpkin juice container.
65. This is in a can of pumpkin.
66. This is in a can of pumpkin juice.
67. This is in a bottle of pumpkin.
68. This is in a bottle of pumpkin juice.
69. This is in a container of pumpkin.
70. This is in a container of pumpkin juice.
71. I take it from pumpkin juice.
72. I take it from the pumpkin juice can.
73. I take it from the pumpkin juice bottle.
74. I take it from the pumpkin juice container.

75. I take it from the can of pumpkin.
76. I take it from the can of pumpkin juice.
77. I take it from the bottle of pumpkin.
78. I take it from the bottle of pumpkin juice.
79. I take it from the container of pumpkin.
80. I take it from the container of pumpkin juice.
81. I take it from the tin of pumpkin.
82. I take it from the tin of pumpkin juice.
83. I put it into pumpkin juice.
84. I put it into the pumpkin juice can.
85. I put it into the pumpkin juice bottle.
86. I put it into the pumpkin juice container.
87. I put it into the can of pumpkin.
88. I put it into the can of pumpkin juice.
89. I put it into the bottle of pumpkin.
90. I put it into the bottle of pumpkin juice.
91. I put it into the container of pumpkin.
92. I put it into the container of pumpkin juice.
93. I put it into the tin of pumpkin.
94. I put it into the tin of pumpkin juice.

Translation:
1. Tämä on appelsiinimehu.
2. Tämä on appelsiinimehutölkki.
3. Tämä on appelsiinimehupullo.
4. Tämä on appelsiinimehuastia.
5. Tämä on appelsiinitölkki.
6. Tämä on appelsiinimehutölkki.
7. Tämä on appelsiinipullo.
8. Tämä on appelsiinimehupullo.
9. Tämä on appelsiiniastia.
10. Tämä on appelsiinimehuastia.
11. Tämä on rypälemehu.
12. Tämä on rypälemehutölkki.
13. Tämä on rypälemehupullo.
14. Tämä on rypälemehuastia.
15. Tämä on rypäletölkki.
16. Tämä on rypälemehutölkki.
17. Tämä on rypälepullo.
18. Tämä on rypälemehupullo.
19. Tämä on rypäleastia.
20. Tämä on rypälemehuastia.
21. Tämä on porkkanamehu.
22. Tämä on porkkanamehutölkki.
23. Tämä on porkkanamehupullo.
24. Tämä on porkkanamehuastia.

25. Tämä on porkkanatölkki.
26. Tämä on porkkanamehutölkki.
27. Tämä on porkkanapullo.
28. Tämä on porkkanamehupullo.
29. Tämä on porkkana-astia.
30. Tämä on porkkanamehuastia.
31. Tämä on omenamehu.
32. Tämä on omenamehutölkki.
33. Tämä on omenamehupullo.
34. Tämä on omenamehuastia.
35. Tämä on omenatölkki.
36. Tämä on omenamehutölkki.
37. Tämä on omenapullo.
38. Tämä on omenamehupullo.
39. Tämä on omena-astia.
40. Tämä on omenamehuastia.
41. Tämä on päärynämehu.
42. Tämä on päärynämehutölkki.
43. Tämä on päärynämehupullo.
44. Tämä on päärynämehuastia.
45. Tämä on päärynätölkki.
46. Tämä on päärynämehutölkki.
47. Tämä on päärynäpullo.
48. Tämä on päärynämehupullo.
49. Tämä on päärynäastia.
50. Tämä on päärynämehuastia.
51. Tämä on kurpitsamehu.
52. Tämä on kurpitsamehutölkki.
53. Tämä on kurpitsamehupullo.
54. Tämä on kurpitsamehuastia.
55. Tämä on kurpitsatölkki.
56. Tämä on kurpitsamehutölkki.
57. Tämä on kurpitsapullo.
58. Tämä on kurpitsamehupullo.
59. Tämä on kurpitsa-astia.
60. Tämä on kurpitsamehuastia.
61. Tämä on kurpitsamehu.
62. Tämä on kurpitsamehutölkki.
63. Tämä on kurpitsamehupullossa.
64. Tämä on kurpitsamehuastiassa.
65. Tämä on kurpitsatölkissä.
66. Tämä on kurpitsamehutölkissä.
67. Tämä on kurpitsapullossa.
68. Tämä on kurpitsamehupullossa.
69. Tämä on kurpitsa-astiassa.
70. Tämä on kurpitsamehuastiassa.

71. Otan sen kurpitsamehusta.
72. Otan sen kurpitsamehutölkistä.
73. Otan sen kurpitsamehupullosta.
74. Otan sen kurpitsamehuastiasta.
75. Otan sen kurpitsatölkistä.
76. Otan sen kurpitsamehutölkistä.
77. Otan sen kurpitsapullosta.
78. Otan sen kurpitsamehupullosta.
79. Otan sen kurpitsa-astiasta.
80. Otan sen kurpitsamehuastiasta.
81. Otan sen kurpitsapurkista.
82. Otan sen kurpitsamehupurkista.
83. Panin sen kurpitsamehuun.
84. Panin sen kurpitsamehutölkkiin.
85. Panin sen kurpitsamehupulloon.
86. Panin sen kurpitsamehuastiaan.
87. Panin sen kurpitsatölkkiin.
88. Panin sen kurpitsamehutölkkiin.
89. Panin sen kurpitsapulloon.
90. Panin sen kurpitsamehupulloon.
91. Panin sen kurpitsa-astiaan.
92. Panin sen kurpitsamehuastiaan.
93. Panin sen kurpitsapurkkiin.
94. Panin sen kurpitsamehupurkkiin.

Note that a dash is added between two identical vowels in joining process. Therefore, such forms are produced as *omena-astia*, *porkkana-astia* and *kurpitsa-astia*.

3 Limitations in general production of single-word compound nouns

The examples above describe a very limited type of compound nouns. The system works when all non-final members are in nominative single. This is in fact the default in single-word compounds. There are, however, exceptions to this rule. Consider examples in (11).

(11)
pattern matching - hahmonsamaistus , hahmonparitus
pattern detection technique - hahmontunnistustekniikka
pattern recognition technique - hahmontunnistustekniikka
pattern matching technique - hahmonsamaistustekniikka , hahmonparitustekniikka

These examples resemble the earlier examples, but the first member *hahmon* is in genitive. To produce genitive for one of the non-final members would make the system unnecessarily complicated. In such cases the best and safest solution is to handle them as MWEs.

4 Compund nouns as separate words in Finnish

It is also common that Finnish uses separate words for compound nouns. In such cases it is common that the non-final members of the compound are in genitive. Consider the examples in (12).

(12)
(a) cold war fear - kylmän sodan pelko
(b) cold war - kylmä sota
(c) natural language processing - luonnollisen kielen prosessointi
(d) faculty member - tiedekunnan jäsen
(e) data science officer - datatiedevirkailija
(f) machine learning - koneoppiminen

All English examples in (12) have an identical form - a string of nouns separated with spaces. On the side of Finnish, (a), (c) and (d) have genitive in non-final members, they are written as separate words, and they do not inflect further. Example (b) has two members, written separately, and both inflect. Examples (e) and (f) have single-word implementations in Finnish.

In the above cases it is safe to use MWE implementation such as in (13).

(13)

```
(a)
"<It>"
      "it" %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" %+FMAINV V PRES SG3
"<cold_war_fear>"
       "cold_war_fear" { kylmän sodan pelko N1-D } %PCOMPL-S MW N
SG NOM
(b)
"<It>"
      "it" %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" %+FMAINV V PRES SG3
"<cold_war>"
       "cold_war" { kylmä N10 FRONT sota N10-F } %PCOMPL-S MW N SG
NOM
(c)
"<It>"
      "it" %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" %+FMAINV V PRES SG3
"<natural_language_processing>"
       "natural_language_processing" { luonnollisen kielen
prosessointi N5-J } %PCOMPL-S MW N SG NOM N-ING
(d)
"<He>"
      "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<is>"
```

```
      "be" %+FMAINV V PRES SG3
"<faculty_member>"
       "faculty_member" { tiedekunnan jäsen N32 FRONT } %PCOMPL-S
MW INDEF N SG NOM
(e)
"<He>"
      "he" %SUBJ CAPINIT PRON PERS NOM SG3
"<is>"
      "be" %+FMAINV V PRES SG3
"<data_science_officer>"
       "data_science_officer" { datatiedevirkailija N12 } %PCOMPL-
S MW N SG NOM
(f)
"<It>"
      "it" %SUBJ CAPINIT PRON NOM SG3
"<is>"
      "be" %+FMAINV V PRES SG3
"<machine_learning>"
       "machine_learning" { koneoppiminen N38 } %PCOMPL-S MW N SG
NOM N-ING
```

We see above that each compound noun in English has received an appropriate representation in Finnish. If the noun has an inflection tag after it, it will be inflected in all noun cases. If no onflection tag follows the noun, it cannot be inflected.

5 Conclusion

Noun compounding in English to Finnish MT is a complex process, and several methods should be used to make the process manageable. The safest way would always be to treat compound nouns as MWEs. From the viewpoint of processing speed, however, this is not ideal. If global disambiguation rules can be used for some phenomena, they should be used, because they greatly simplify the process. The examples discussed here are only one instance of developing global noun compounding. This is an area, which deserves more work.