Generation of Noun-Noun Compounds in the Spanish-English Machine Translation System SPANAM®

Julia Aymerich
Pan American Health Organization
525 23rd Street, N.W.
Washington, DC 20037 (USA)
aymericj@paho.org

Abstract

The translation of Spanish Noun + preposition + Noun (NPN) constructions into English Noun-Noun (NN) compounds in many cases produces output with a higher level of fluency than if the NPN ordering is preserved. However, overgeneration of NN compounds can be dangerous because it may introduce ambiguity in the translation. This paper presents the strategy implemented in SPANAM® to address this issue. The strategy involves dictionary coding of key words and expressions that allow or prohibit NN formation as well as an algorithm that generates NN compounds automatically when no dictionary coding is present. Certain conditions specified in the algorithm may also override the dictionary coding. The strategy makes use of syntactic and lexical information. No semantic coding is required. The last step in the strategy involves post-editing macros that allow the posteditor to quickly create or undo NN compounds if SPANAM® did not generate the desired result.

Keywords

Noun-Noun compounds, Spanish-English machine translation and SPANAM system

SPANAM® today

SPANAM® has been operational at the Pan American Health Organization since 1980. The program, along with its English-Spanish counterpart Engspan®, was ported from the mainframe to the PC in 1992 and then to the Windows environment in 2000 (León, 2000).

The basic architecture of the program hasn't changed since 1985: it is a transfer system with an ATN that generates a top-down, left-to-right sequential parse with chronological and explicit backtracking (Vasconcellos and León, 1988; Amores, 1996). Dictionary entries are rich in morphological, syntactic, and semantic information. The grammar and dictionaries have grown considerably in the past 15 years. SPANAM® dictionaries currently contain over 95,000 source entries and 85,000 translations. Source entries include single words (65,000 stem forms), multiple word entries or SUs[1] (11,000), analysis rules or AUs (8,500), and context-sensitive translation rules (10,000).

In several MT evaluation experiments sponsored by the Advanced Research Projects Agency (ARPA/SISTO) between 1992 and 1994, where 3 research systems and 5 commercial systems were evaluated for adequacy, comprehension, and fluency, SPANAM® consistently received the highest scores for all three criteria. The fluency scores were slightly lower than the other two.

The problem

Spanish makes extensive use of NPN constructions (*abuso de sustancias, abastecimiento de agua*) but does not use NN compounds. If the NPN structure is maintained in the English noun phrase, the resulting translation, although understandable, is usually too literal (*supply of water/water supply, abuse of subtances/substance abuse*), especially when several nouns are involved. Compare *center of control of hurricanes* with *hurricane control center*. Adequacy and comprehension may be acceptable, but fluency suffers.

A robust Spanish-English MT program should be able to generate NN compounds in English from NPN constructions in Spanish. It is crucial, however, that MT not overgenerate NN compounds and that it not create ambiguity or change the meaning. It is preferable to be overly literal than to produce a translation that sounds more natural yet confuses the human posteditor and ultimately produces a mistranslation. For example, if the noun phrase *difusión anual de resúmenes y documentos técnicos* is rendered as *annual summary and technical document dissemination*, the posteditor may misinterpret this as *dissemination of [annual summaries] and [technical documents]* or *dissemination of annual [summaries and technical documents]*. However, if the MT system produces *annual dissemination of summaries and technical documents*, the posteditor may choose to rearrange the English NP, but there will be no doubts as to the original meaning of the Spanish NP.

An early attempt at pattern-matching to generate NN compounds was eliminated from SPANAM® in 1979. Because the system did not have enough lexical or syntactic information to decide when to generate NN compounds, pattern-matching was considered counterproductive at the time (Vasconcellos, 1979).

Since SPANAM® had no way to handle NN compound formation, the cases that had come up in PAHO texts had been solved by adding the expression in the dictionary as a Substitution Unit (SU), i.e., an entry that subsumes its members under a single dictionary token and has a set translation. For example, *sistema de documentación* was a unit with the translation *documentation system*. This type of entry creates obvious problems in cases of modification or coordination. For example, in *sistema de documentación y referencia*, the parser conjoined [*sistema de documentación*] with *referencia* (instead of *documentación* and *referencia*) and rendered it as *documentation system and referral* instead of *documentation and referral system*. Another example is the expression *de salud*, which was added as an SU (adjective). This solved many cases but also created numerous problems when *de salud* was itself modified by another adjective (*de salud reproductiva*) or was

[1] The concept of SU and AU in the PAHO systems is explained below.

conjoined (*de salud y bienestar*). In an attempt to solve these problems at the lexical level, over 300 SUs had to be added that contained *de salud*.

Finding a solution

For the sake of efficiency, a strategy to produce NN compounds cannot be based solely on dictionary entries, since there are an unlimited number of NPN constructions that can occur in Spanish, and an equally unlimited number of NN compounds that can be generated in English (Wooley, 1997). It has been suggested (Brown, 1993) that there is no simple rule and that statistical methods are the only solution.

At PAHO, we use a combination of rules and analysis of real examples. As is customary in our development environment, we decided to follow a strictly empirical approach for the design of general rules. We analyzed dozens of documents processed with SPANAM® and collected a corpus of thousands of sentences containing NPN constructions. A thorough analysis of the data revealed that our strategy would have to involve:

1. Algorithm to form NN compounds on the fly, without the need for dictionary entries. The algorithm should only apply if there is a complete parse of the sentence. A conservative strategy is preferable to avoid the introduction of ambiguity.
2. Algorithm to filter out cases where NN formation should not be attempted
3. Dictionary coding for key words and expressions that block or allow NN formation. These are considered the exceptions to the general algorithm and override many of its filter conditions. As a first step, we had to locate NPN SUs and replace them with a new type of Analysis Unit (see below).
4. Postediting macros to create or undo NN compounds when SPANAM® produces the wrong result.

Coding in dictionary entries

English target entries
Certain English nouns always block or favor NN formation, as head noun, modifier noun, or both. These cases are coded in the English target dictionary. Example: *cause* never participates in NN compounds, neither as head (*causa de preocupación*: **worry cause*) nor as modifier (*explicación de causas*: **cause explanation*). On the other hand, *industry* is coded to favor NN formation as a head noun (*industria del petróleo: oil industry*). There are currently 214 target nouns coded, 93% of which block NN formation.

NPN Analysis Units
Most nouns do not behave so regularly. For instance, *curso* will front in some contexts (*curso de campo: field course*) but not in others (*curso de acción: course of action*). These expressions may be added as Analysis Units.

An AU is a rule stored in the source dictionary that resolves Part of Speech ambiguities, specifies alternate translations for one or more of its members, and indicates that its members together function as a certain type of phrase (Vasconcellos and León, 1988). A NPN AU additionally specifies whether the NN compound should be allowed or blocked. The words in an AU are parsed as individual tokens (which means that they may have their own modifiers and/or conjuncts) and the parser keeps them together in an NP. NPN AUs can be added for the following patterns:

- NPN (*agencia de viajes, canal a presión*)
- NPNPN (*agenda de negociación de paz*)
- NPNA (*centro de atención médica*)
- NAPN (*fuerza electrostática de restitución*)
- NPTN (*picadura de los insectos*)

Additionally, these AUs may:

- allow for wild card adjectives after the head noun. For example, the AU *sistema de salud* will be used for *sistema nacional de salud, sistema local de salud, sistema gratuito de salud*, etc.
- specify semantic features (classes of nouns) instead of specific lexical items. For example, the NN compound should be generated when *fractura de* is followed by a noun coded BODY PART: *fractura de cadera* (*hip fracture*), *fractura de brazo (arm fracture), fractura de huesos (bone fracture*), etc.
- select adjectival translations for the noun to be fronted (*medios de diagnóstico: diagnostic media*, not *diagnosis media*)

The SPANAM® dictionaries currently contain 4300 such entries (76% force NN compounding and 24% block it). Some 2000 AUs were converted from SUs. At the time when SUs were being turned into NPN AUs, it was decided that some should remain SUs because they were technical terms (*salida de mar: tidal wave, equipo de lectura óptica: optical scanner*), the translation was not a NN compound (*golde de Estado: coup d'etat, caja de Pandora: Pandora's box*), they were nested into larger SUs, or because they were collocations with extremely high frequency in PAHO texts.

PN Analysis Units
This second type of AU specifies a preposition + Noun that behave like an adjective and should always be fronted to form a NN compound, regardless of the noun that occurs as the head. Examples include *de calidad* (*quality*), *de emergencia* (*emergency*), *con sobrepeso* (*overweight*), *de cristal* (*glass*). These AUs may optionally select an adjectival translation for the noun: *de madera* (*wooden*), *de moda* (*fashionable*). Note that the words in the AU are parsed as individual words and the parse may undo the NN formation. Example: *programa de calidad*: *quality program*, but *programa de calidad dudosa: program of dubious quality*.

The SPANAM® dictionaries currently contain 73 such entries. Some 20 of these were converted from old SUs.

Algorithm to generate NN compounds with complete parse of NP

NN compound formation is attempted after the sentence has been parsed, contextual lexical selection has been determined, and the words have been looked up in the English target dictionaries.

The algorithm starts from the bottom of the parse tree, right to left. It locates NPs whose structure has been completely parsed and processes one NP at a time. The filtering conditions described in the next paragraphs are summarized in Table 1 below.

Within each NP with a noun head (no pronoun or adjective heads), postnominal descriptors are located and fronted. Adjectives, adverbs, and negators are fronted and the number of premodifiers fronted is recorded for future reference. If a hyphenated phrase is fronted, this NP is automatically blocked as a candidate for a NN compound.

Within the NP, the algorithm then looks for Prepositional Phrases introduced by OF (not followed by another preposition: *de entre los arbustos*). Most of these will be translations of "DE", but not all. PPs headed by a preposition other than OF are also accepted if they are part of an NPN AU (*seguro contra* accidentes: *accident insurance*).

The algorithm records the number of target words for the upper and lower nouns (some Spanish single-word nouns may translate as more than one word in English and viceversa). If there are more than two target words, NN compounding is automatically blocked.

NN compounding is also prevented if either upper or lower noun are not found in the dictionary, are proper names, are uppercase (unless they appear in a title or in a text that is all uppercase), are both Time nouns (*reunión de fin de mes*), or are coded in the dictionary to prevent NN compounding.

Evaluate the upper noun (N1). If it is coded to always allow fronting, no further conditions are tested. This code may come from the target entry or from an AU. Otherwise, the upper noun must pass the following tests in order to allow fronting:

1. it is a not a partitive, quantity noun, classifier, and is not preceded by a numeral (*ganancia de 8 millones de dólares, varios tipos de salmones, una mayoría de congresistas*)
2. it does not take DE as bound preposition
3. it is not conjoined. Exception: verb nominalizations are conjoined with other verb nominalizations if the PP headed by DE is parsed as the object of both nouns: *prevención y control de enfermedades*: *disease prevention and control*

Evaluate the lower noun (N2). It must pass the following tests in order to allow fronting:

1. it has not been parsed in a PP headed by a preposition bound to N1
2. it doesn't have a Relative Clause attached
3. it doesn't have a direct object (direct objects of nouns are NPs parsed in PPs headed by DE, when N1 is a transitive verb nominalization)
4. it is not preceded by a determiner. Exceptions: if lower noun is coded to block article insertion, it is coded to always be fronted, or the determiner has been blocked for translation elsewhere
5. it is not preceded by a numerative, demonstrative, or quantifier
6. it is not coded plural in the target dictionary or, if it is, its translation doesn't end in 'S'
7. it is not modified by a past participial adjective
8. it is not conjoined in a PP (*jefes de empresa y de comercios privados*)

If both upper and lower noun pass the tests, the following conditions must also be met:

1. the total number of descriptors and nouns does not exceed four
2. lower NP cannot have more than one descriptor, to avoid creating ambiguity: *solución de problemas de abastecimiento de agua potable*: *solution of problems of drinking water supply*, not *solution of drinking water supply problems*, or *drinking water supply problem solution.* Exception: if N2 is part of a NN compound, front a maximum of two descriptors.
3. N1 and N2 cannot both have descriptors, to avoid heavy embedding and to avoid creating ambiguity: *Campañas especializadas de información pública*: *specialized campaigns of public information*, NOT *specialized public information campaigns*
4. if N1 is a verb nominalization, N2 cannot have descriptors (*difusión de documentos técnicos*). Exception: there is an AU with adjective + Noun for N2. In this case, fronting is accepted because it is assumed that the adj + Noun collocation is a semantic unit that will not create ambiguity when fronted (*archivo de documentos normativos*: *policy document file*)
5. if N1 is a verb nominalization, has a direct object, and is conjoined, the noun with which it is conjoined cannot have any descriptors
6. if N2 is conjoined, front both nouns only if there are only two nouns and neither has modification

If all tests are passed, rearrangement occurs. Rearrangement involves:

1. fronting N2
2. blocking the translation of the preposition
3. fronting the descriptors of N2
4. removing all article insertion information for N2 (since it is now a modifier)
5. making N2 singular
6. if N2 is conjoined, fronting the conjoined noun also
7. if there is an upper conjunct sharing the direct object, moving N2 (and its modifiers) in front of upper conjunct: *monitoreo y control de la contaminación ambiental*: *environmental pollution monitoring and control*
8. marking the NP so that it will not undergo further fronting. This prevents more than one level of embedding. However, compounds spanning three NPs are allowed if there are no descriptors in either

NP: *sistema de abastecimiento de agua*: *water supply system*, or N1 has special coding to always allow fronting (either by itself or as part of an NPN AU)

If N1 and N2 were part of a NN AU but fronting could not occur because of other conditions, N1 is marked so that it is not fronted and thus separated from N2 (*cobertura de los servicios de imaginología y radiodiagnóstico*).

Algorithm to generate NN compounds with incomplete parse of NP

The algorithm cannot start from the bottom of the parse tree (i.e., from the end of the sentence to the front) because the information is not available. If there is no analysis, the algorithm starts from the beginning of the sentence and stops at the words that trigger NPN AUs. In other words, NNs are only generated if specified by an AU in the lexicon.

First, it looks for PN AUs (*de salud*). It blocks the translation of the preposition, changes the POS of the noun to adjective, and fronts it. If the noun is itself modified by another adjective, it is fronted as well (*estrategia de salud reproductiva*: *reproductive health strategy*). If the head noun is modified by another adjective, the algorithm makes sure that the "adjectival noun" is next to the head noun (*necesidad urgente de salud*: *urgent health need*, NOT *health urgent need*).

Then, it looks for NPN AUs and generates the NN compounds. The same rules apply as with complete parses, but rearrangement is based on look-back and look-head procedures rather than on information from the parser. This procedure ensures that the NN is always generated (as with SUs) but sometimes creates incorrect NN compounds because it does not take advantage of the syntactic context.

Upper Noun (N1)		Lower Noun (N2)	prep	Form NN	overridden by lexical coding that favors NN
lexical coding blocks NN	or	lexical coding blocks NN		NO	N/A
			not OF	NO	yes
			2 preps	NO	no
			Bound to N1	NO	yes
partitive, quantity noun, classifier, or preceded by numeral				NO	yes
Proper name	or	Proper name		NO	yes
has hyphenated phrase	or	has hyphenated phrase		NO	no
not found in dictionary	or	not found in dictionary		NO	no
uppercase (unless title)	or	uppercase (unless title)		NO	yes
conjoined (unless vnom conjoined with vnom that share the object)				NO	yes
		more than one target word		NO	yes
		has attached Relative Clause		NO	yes
		is vnom with direct object		NO	yes
		preceded by determiner (unless determiner blocked for translation)		NO	yes
		preceded by numerative, demonstrative, or quantifier		NO	no
		target entry plural and ends in S		NO	yes
		modified by past participle		NO	no
		conjoined in a PP		NO	no
		two or more descriptors		NO	yes (2 max)
has descriptors	and	has descriptors		NO	yes
verb nominalization	and	has descriptors		NO	yes
		conjoined with more than one noun		NO	no
		conjoined with noun that has descriptors		NO	yes
total number of words (adjectives and nouns) is more than 4				NO	no

Table 1: Filtering conditions

Implementation in postediting Macros

SPANAM® incorporates a set of postediting Macros for MS Word that help the posteditor speed up the process of correcting the raw output.

Two of these Macros involve NN compounds. With the first Macro, the posteditor places the cursor on any preposition and clicks on the Macro button. The words to the left and to the right are switched, and the preposition is deleted. With the second Macro, the posteditor places the cursor on the first noun of a compound and clicks on the Macro button. That noun and the noun to its right are reversed and the preposition *of* is inserted in between.

Posteditors are alerted about NN compounds and are advised to use these Macros when SPANAM® doesn't produce the desired result.

Testing and Conclusions

We need to be able to generate NNs on the fly now more than ever because we license the system to outside users.

To verify the improvement in fluency, we recently compared raw output of the current version of the software with the one used in the 1994 ARPA experiment (when NN compounds were only generated with SUs). A non-biased native speaker of English was asked to judge the difference in fluency between the 1994 and 2001 raw translations. The results of the comparison are summarized in Table 2.

Total candidates for NN (20,000 words of text)	388
NNs generated in 1994 (SUs)	18 (4.6%)
NNs generated in 2001	93 (23.7%)
with SUs	10
with NN AUs	33
with algorithm	48
with target coding	2
NNs that would have been generated with complete parse	142 (36.5%)
NNs not generated in 94 or 01	260
NN would have been better	58
NN would have been worse	202

Table 2: Testing results

In analyzing these results, we find that, all in all, SPANAM® has gained in fluency:

- no desirable NN compounds were lost
- of the new NN compounds, 73% produced more fluent translations (*social welfare system, television screen*), 24% generated errors that would have to be corrected by posteditors (*drug movement in the world* as opposed to *movement of drugs in the world*, *pension funds and retirements* as opposed to *retirement and pension funds*), and 3% were dubious (*environment of pressure/pressure environment*)
- another 58 desirable NNs were not generated in either version (59% were not generated because of the conservative algorithm and 41% because of the lack of a complete parse).
- the filtering conditions successfully rejected the undesirable NNs (202)

Our algorithm was fine-tuned over the years (1995-1997 approximately) as texts were being processed and analyzed in our translations department. The syntactic component of the algorithm is now stable. The dictionary coding of NPN and PN AUs, on the other hand, is an ongoing process that can never be said to be complete.

As an improvement to the current algorithm, we may consider the inclusion of semantic information for the resolution of some dubious cases of NN formation.

Acknowledgments

The original algorithm was designed by the author in collaboration with Mila Ramos-Santacruz. Graciela Rosemblat performed the conversion of SU entries into AUs. Maxine Siri evaluated the difference in fluency. The author would also like to thank Marjorie León for her comments and suggestions.

References

Amores, J. G. (1996) *Los sistemas de traducción automática ENGSPAN y SPANAM de la Organización Panamericana de la Salud.* Procesamiento del Lenguaje Natural 18, 87-10.

ARPA Workshop on Machine Translation Evaluation.20-22 November 1994. Sheraton Premier Hotel, Tyson's Corner, Vienna (VA)

Brown, P. (1993) *Candide: Overview of System Improvements*. ARPA MT Evaluation Workshop. Carnegie Mellon University, Center for Machine Translation

León, M. (2000) *A New Look for the PAHO MT System*, in J.S. White (Ed.) AMTA 2000, LNAI 1934, pp. 219-222

León, M. and L. Schwartz (1986) *Integrated Development of English-Spanish Machine Translation: from Pilot to Full Operational Capability: Technical Report of Grant DPE-5543-G-SS-3048-00 from the U.S. Agency for International Development*. Washington, DC: Pan American Health Organization

Vasconcellos, M. and M. León (1988) *SPANAM and ENGSPAN: Machine translation at the Pan American Health Organization*. Computational Linguistics 11, pp 122-136. Also in J. Slocum, editor (1988), Machine Translation systems, pages 187-236. Cambridge University Press

Vasconcellos, M. (1979) *A Further Note on the Rearrangement of Nouns*. Unpublished study

Woolley, R. (1997) *Compound Nominal Groups in the Machine Translation of Medical English: Lexical Units or Analysable Sequences?* Submitted in partial completion of the MSc degree in Teaching English. Aston University.