Káñina, Rev. Artes y Letras, Univ. Costa Rica XL (Extraordinario): 49-58, 2016 / ISSN:2215-2636

MEASURING THE IMPACT OF COLLOCATIONAL KNOWLEDGE ON SENTENCE PARSING

*Eric Wehrli**

ABSTRACT

In this paper we focus on collocations, which have been studied in computational linguistics since they constitute a key factor when processing natural languages. For instance, they usually represent a challenge in automatic translation because the association of two terms is not easily computed. We proposed that the parser should be provided with a lexical database in order to make more effective the identification of collocations during the parsing process. We assessed this claim by using a corpus of 6'000 sentences retrieved from the British magazine *The Economist Espresso*. The corpus was parsed twice, first with the collocation detection component turned on and then with it turned off, and to make the comparison the Fips tagger was used. The results showed an improvement of the quality when the parser has access to collocation knowledge.
Key words: collocations, multiword expressions, sentence parsing, computational linguistics, natural language processing.

RESUMEN

En este artículo el análisis se centra en colocaciones, las cuales han sido estudiadas en repetidas ocasiones en lingüística computacional, ya que constituyen un factor clave durante el procesamiento de lenguajes naturales. Por ejemplo, por lo general representan un desafío cuando se trabaja con traducción automática debido a que la asociación de dos términos no es fácilmente calculable. Aquí se propone que el procesador sintáctico debería estar provisto con una base de datos léxica con el fin de hacer más eficaz la identificación de las colocaciones durante el proceso de análisis. Se evaluó esta afirmación utilizando un corpus de 6000 oraciones extraídas de la revista británica *The Economist Espresso*. El corpus fue analizado dos veces, primero con el componente de detección de colocaciones y luego sin este; para hacer la comparación se ha empleado el etiquetador Fips. Los resultados mostraron mejor calidad cuando el analizador sintáctico tenía acceso al conocimiento sobre las colocaciones.
Palabras clave: colocaciones, unidades léxicas pluriverbales, análisis sintáctico de oraciones, lingüística computacional, procesamiento de lenguaje natural.

* Université de Genève. Suiza. Correo electrónico: eric.wehrli@unige.ch
Recepción: 15/1/2016. Aceptación: 16/3/2016.

1. Introduction

Collocations, taken here in the relative broad sense of arbitrary and conventional association of two lexical items (not counting grammatical words) in a specific grammatical configuration – *white lie*, *cold case*, *to claim the life* – have been the focus of attention for a long time among computational linguists.

In this paper, I will briefly review (i) the different types of collocations, with examples from several languages, but mostly from English and French and (ii) why collocations matter for natural language processing (NLP) in general and more specifically for machine translation (MT). I will then turn to the crucia questions of "how to identify collocations" and "when collocation identification should occur in the complex task of language processing".

My answer to the first question is twofold: first, collocations must be "known" by the system. The arbitrary and conventional association of two terms, which constitute a collocation, cannot be guessed or computed. Therefore, collocations – as well as other multi-word expressions – must be part of the lexical knowledge part of the lexical database used by a parser. Second, collocations occur in spe- cific grammatical configurations, such as an adjective modifying a noun, a verb and its direct object, a noun modifying a noun, etc., which means that structural information is crucial for the processing of collocations.

As for the second question, I will argue that collocation identification should not occur before, nor after, parsing but during parsing, as soon as the second component of the collocation has been processed and attached to the syntactic structure. This, of course, adds some complexity to the parsing process, but I wil show that this added complexity is rewarded by better results.

The last section will present an evaluation of the impact of collocations on parsing. Parsing a sizeable corpus first with the collocation identification compo- nent turned on, and then parsing the same corpus with the component turned off.

I will show that the quality of the overall analyses obtained in the first run (col- location component on) is significantly better than the one obtained in the second run (collocation component off).

2. Collocations and other multiword expressions

Multiword expressions (henceforth MWEs) are lexical units made of more than one "word" (in the intuitive sense). MWEs do not constitute an homogeneous class. Although there is no general agreement among lexicographers and linguists about the precise partitioning of the class, we will assume the following subtypes:

- compounds ("word with spaces"
  by and large, *little by little*, *more or less*, *fer a` cheval*, *horse shoe*

- discontinuous words (e.g. particle verbs in English or German, pronominal verbs in Romance)
  *she **looked** this word **up***
  *der Zug **fa¨ hrt** um halb acht **ab*** (the train leaves at half past seven)
  L'homme ***s'est suicide´*** 'the man committed suicide'

- named entities
  John F. Kennedy, *European Central Bank*, *World Economic Forum*

- idiomatic expressions
  to kick the bucket, *bouffer du lion* (to be hyperactive), *meter la pata* (to make a blunder)

- collocations
  hot topic, *occupational hazard*, *risques du me´tier*, *black economy*, *cold case to command admiration*, *to take up a challenge*, *to claim the life*
  state of emergency, *bone of contention*, *casco di banane* (*bunch of bananas*)

- other fixed expressions, proverbs, etc.
  carpe diem, *last but not least*, *a` plus ou moins bre`ve e´che´ance*, *sooner or later*, *a pain in the neck*

From a syntactic viewpoint, compounds and named entities are lexical units of lexical category (noun, adjective, adverb, etc.). They behave just like simple lexical items but happen to contain spaces (or sometimes punctuation signs). We will consider that they belong to the lexical database[1]. Discontinuous words (e.g particle verbs) can also be considered as lexical units of lexical category (verbs in our examples), which happen to be made of two parts – the verb and the particle – which may not be adjacent to each other. It is the parser's task to recognize tha the two elements belong to the same lexical unit[2].

In contrast to compounds, named entities and discontinuous words, collo- cations and idiomatic expressions behave at the syntactic level not like lexica units but rather like syntactic units (phrases). They constitute noun phrases in the case of noun-noun, adjective-noun or noun-preposition-noun collocations, verb phrases in the case of verbal collocations (verb-direct object, verb-prepositiona object, etc.). Such MWEs must also be listed in the lexical database used by the parser – they cannot be guessed – for instance as associations of two lexemes (or *groupements usuels* 'usual phrases' as coined by Bally, 1909).

While many of our remarks and observations hold for all or many of the MWEs subclasses, we will mostly be concerned with collocations, taken here broadly as the association of two lexical units in a particular grammatical configu- ration. While idiomatic expressions often display semantic opacity (cf. *to kick the bucket* in the sense of dying) as well as restrictions on their syntactic behaviour such as no passive, no movement, no modifier, etc., collocations usually keep their usual syntactic properties, and are semantically relatively transparent.

2.1. Multiword expressions matter for NLP

The importance of MWEs for NLP applications, such as translation, is widely recognized. To undersand why, consider the three following points:

- most expressions cannot be translated literally (*dead loss*, *to make an appointment*, *to kick the bucket*)
- some compounds as well as some fixed expressions do not respect gram- matical rules, eg. *by and large*
- MWEs have a high frequency
  named entities constitute approx. 10% of newspaper articles few sentences do not contain any compound or collocation

As already pointed out, it is therefore necessary for most NLP applications to "know" and to properly identify MWEs. This, however, may turn out to be a complicated task if you consider what I will refer to as the syntactic flexibility of many MWEs, limited here (for collocations) to the three following cases (see Sag et al.(2002):

- Adjectival or adverbial modifiers can often be attached within a collocation, separating the two terms, eg. *a **school** of little **fishes***
- Several types of collocations can undergo grammatical processes which may modify the canonical order of the collocation (eg. passive, relativiza- tion, etc.)
- Occasionally, a noun in a verb-object or subject-verb collocation can be replaced by a pronoun

Syntactic flexibility is particularly important with verbal collocations such as verb-object or verb-prepositional object and subject-verb, where the two terms of the collocation can be separated by an arbitrary number of words; due to syntactic transformations, such as passive, relativization, interrogation, etc., they can also occur in a reverse order, which of course makes it difficult to identify them in a sentence. To illustrate, consider the following examples, in which the terms o collocations are in boldface.

(1)a. The scheme **addresses** one of America's prickliest **problems**.

b. The **problem** –that poor children do not get the chances that rich ones do– is a real one, but needs to be **addressed** earlier.

The Bangkok stockmarket plunged 4.5% in a single day after **news** of the possible human-to-human transmission **broke**.

Sentence (1a) contains the verb-direct object collocation (*to address a prob- lem*) with several words in-between the two terms *addresses* and *problems*. The same collocation occurs in sentence (1b), where the two terms are in reverse order due to passive and are separated by considerable material. Finally, sentence (1c) illustrates a subject-verb collocation (*the news breaks*) –a collocation type much less frequent than the verb-object type– with again several words separating the two terms.

Another transformation can affect collocations, pronominalization, as in the examples (2) below. Each of the two sentences in (2a) contains an occurrence of the collocation *to make a case*. Notice, however, that in the second sentence the direct object (*case*) has been pronominalized. In other words, the pronoun *it* which refers to the noun *case* of the previous sentence, validates the collocation The second example (2b) illustrates a similar scenario, with the pronoun *it* refer- ring to the noun *money*. Since the pronoun is the subject of the passive form *would be well spent*, it is interpreted as direct object of the verb and therefore stands for an occurrence of the collocation *to spend money*.

(2)a. Every Democrat is **making** this **case**. But Mr Edwards **makes it** much more stylishly than Mr Kerry.

b. ...though where the **money** would come from, and how to ensure that **it** would be well **spent**, is unclear.

2.2. Treatment of MWEs in a linguistically-based system

In this section we will briefly describe how our Fips parser[3], a multilingual grammar- based parser, handles MWEs. As explained above, we assume that MWEs must be "known", that is they are listed in the lexical database used by the parser. Com- pounds (and listed named entities) can be recognized during the lexical analysis of a sentence, just like plain words. As for the other types of MWEs, since their identification requires syntactic knowledge (cf. Seretan 2011), it should happen during the parse, as soon as the last term of the association (collocation or expression) is attached to the structure[4].

A collocation database has been added to our monolingual lexical databases using the collocation extraction system developed by Violeta Seretan and oth- ers at LATL (cf. Seretan & Wehrli, 2009; Seretan, 2011). This system extracts candidate-collocations from a corpus, filters those candidates using standard as- sociation measures and then let the linguist/lexicographer validate the best candidates, which are entered in the collocation database. The current content of the database for five European languages is shown in table 1 below.

TABLE 1

Number and types of collocations in the Fips lexical database

collocation type	English	French	German	Italian	Spanish
adjective-noun	2'803	5'391	482	1'325	1'615
noun-noun	5'342	429	2'436	131	66
verb-object	701	1'401	196	250	1098
others	1'302	10'139	370	1'453	1'569
total	10'148	17'360	3'484	3'159	4'348

The collocation detection component integrated in the Fips parser works as follows. It is triggered, during the parse, by the application of a right (or left) attachment rule. Governing nodes of the attached element are iteratively consid- ered, halting at the first node of major category (NP, VP, AP, AdvP)[5]. Then, the procedure checks whether the pair [governing item + governed item] corresponds to an entry in the collocation database.

This procedure will be illustrated by means of a simple example. We will return and refine it to handle more complex cases below.

Consider as first example the sentence (3a) with the verb-object collocation *to take up a challenge*. The structure, as assigned by Fips, is given in (3b) in the labelled-bracketing form, as well as in the more familiar phrase-structure repre- sentation in figure 1.

(3)a. Paul took up a new challenge

b. $[_{TP}\ [_{DP}$ Paul] $[_{VP}$ took up $[_{DP}$ a $[_{NP}\ [_{Adj}$ new] challenge]]]]

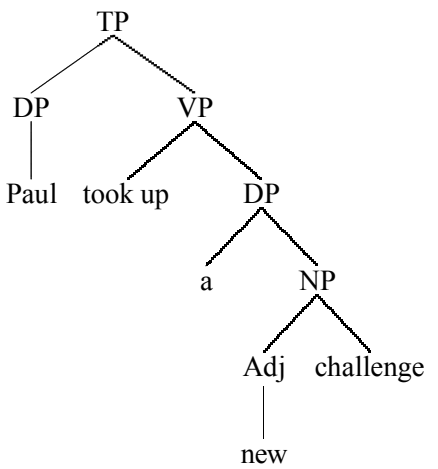

FIGURE 1

Phrase-structure representation of sentence (3a).

When Fips reads the word *challenge*, finding an adjective on its left, a left-attachment rule will create the noun phrase $[_{NP}\ [_{Adj}$ new] challenge], which can be attached as complement to the determiner phrase headed by the indefinite determiner *a*, itself governed by the verb *took up*. Given the strategy for collocation detection described above, going up the phrase-structure representation from the noun phrase, we find first the DP node and then the VP node. The latter being a major category node, the procedure halts and we check whether the pair [take up + challenge], with *take up* as a verb and *challenge* as a direct object, constitutes an entry in the collocation database.

This is indeed the case, so the collocation reading is assigned to the verb phrase.

Let us turn now to more complex cases, such as the ones involving syntactic movement, as in examples (4a-e).

(4)a. *wh*-interrogatives
Which **record** did Paul **break**?
[which record**]** did [Paul [break **[**e**]**]]]
CPDPTPVPDP

b. relative clauses
The **record** that Paul has just **broken** was very old.

c. *tough*-movement
This **record** seems difficult to **break**.

d. *wh*-interrogative + *tough*-movement
Which **record** did Paul consider difficult to **break**?

e. left dislocation + small clause + *tough*-movement
Ce **record**, Paul le conside` re difficile a` **battre**
"this record Paul considers very difficult to break"

[$_{CP}$ **[**$_{DP}$ ce record**]** [$_{TP}$Paul le$_{i}$ conside`re **[**$_{DP}$e**]**
[$_{FP}$ **[**$_{DP}$ e**]** difficile [$_{CP}$a` [$_{TP}$ [$_{VP}$ battre **[**$_{DP}$e**]**]]]]]]

How can the parser identify the collocation *to break a record* in such sen- tences? The answer is surprisingly simple if one takes into account the fact tha Fips assumes Chomsky's *wh*-movement analysis for such sentences (cf. Chom- sky, 1977). According to this view, *wh*-phrases (e.g. interrogative phrases or relative pronouns) bind an empty category in the position corresponding to their interpretation[6]. For instance, a *wh*-phrase interpreted as a direct object binds an empty category in that position. In the analysis of sentence (4a) returned by Fips, the empty category is represented as [$_{DP}$ e] and is co-indexed with the *wh*-phrase.

Given this analysis, to handle collocations involving *wh*-objects, the identi- fication procedure must be slightly modified in order to be triggered not just by the attachment of a direct object, but also by the attachment of an empty category (a trace) in the direct object position of a verb. Then the procedure will check whether the verb and the antecedent of the trace constitute a verb-object entry in the collocation database.

Finally, consider the case where the direct object of a verb-object collocation – or the subject of a subject-verb collocation – has been pronominalized, as in examples (5a-b).

(5) a. Paul set a new **record** last year and he hopes to **break it** this year.

b. Paul set a new **record** last year. He hopes to be able to **break it** again.

In such cases, to identify the collocation one has to consider the anteceden (the referent) of the pronoun. Fips uses an anaphora resolution component which tries to connect a pronoun with a preceding noun phrase in the sentence as in ex- ample (5a) or in the preceding sentence as in (5b)[7]. The collocation identification procedure has been updated, again, in order to be triggered by a pronoun attached for instance, in the direct object position. First the anaphora procedure attempts to identify the antecedent of the pronoun and then the collocation procedure verifies whether the verb and the antecedent of the pronoun correspond to a collocation in the database.

3. Evaluation

In order to assess the claim that collocational knowledge helps the parser, we have conducted an evaluation over a corpus of approximately 6'000 sentences taken from *The Economist*

Espresso, the daily on-line news service of the eponymous British magazine. The research questions are specifically

- What is the statistical significance of ambiguity resolution based on collocation knowledge?
- How frequently, in a given corpus, does the detection of a collocation help the parser make the "right" decision?

To answer those questions, we parsed the corpus twice; first with the collo- cation detection component turned on and then with the component turned off We then compared the results of both runs. To make the (manual) comparison easier, we used the Fips tagger, that is the Fips parser with part-of-speech output It is indeed much easier to compare POS-tags than phrase-structures. Table 2 and table 3 below illustrate the Fips tagger output for the segment in boldface of the sentence given in (6). The first table gives the results obtained with the collocation detection component turned on, and the second table the results obtained with the component turned off.

TABLE 2

Parser output with collocation knowledge

word	tag	position	collocation
the	DET	27	
total	ADJ	31	
worldwide	ADJ	37	
labour	NOUN	47	
costs	**NOUN**	54	labour costs

TABLE 3

Parser output without collocation knowledge

word	tag	position	collocation
the	DET	27	
total	ADJ	31	
worldwide	ADJ	37	
labour	NOUN	47	
costs	**VERB**	54	

(6) The researchers estimated **the total worldwide labour costs** for the iPad at $33, of which China's share was just $8.

The sentence segment *the total worldwide labour costs* is displayed in both tables with the words in the first column, the part-of-speech tag in the second column and the position – expressed as position of the first character of each word starting from the beginning of the sentence– in the third column. For the POS tagset, we opted for the universal tagset (cf. Petrov et al., 2012). As we can see the word *costs* is taken as a noun in the first analysis, as a verb in the second. The (correct) choice of a nominal reading in the first analysis is due to the detection o the collocation *labour costs*. In the second run, given the absence of collocationa knowledge, the parser opts for the verbal reading.

Both output files could then easily be manually compared using a specific user interface as illustrated in the screen shot given in the next page, where POS differences are displayed in red.

A summary of the results of the evaluation is given in table 4. The first line shows the number of complete analyses[8]. Collocational knowledge increases the number of complete analysis by approximately 0.5%, or about 30 sentences for our corpus of 6'000 sentences. 727 tags are different between the two runs. Of those, excluding differences which don't really matter (some words can be ana- lyzed either as predicative adjectives or as adverbs without much semantic differ- ences, etc.), in 382 cases the tags were better in the first run (with collocationa knowledge), and 106 cases better in the second run (without collocational knowl- edge). In other words, collocational knowledge helped the parser make the better decision four times more than it penalized it. Notice finally that

1668 colloca- tions were detected in the corpus (more than one in four sentences), which clearly stresses the high frequency of this phenomenon in natural language.

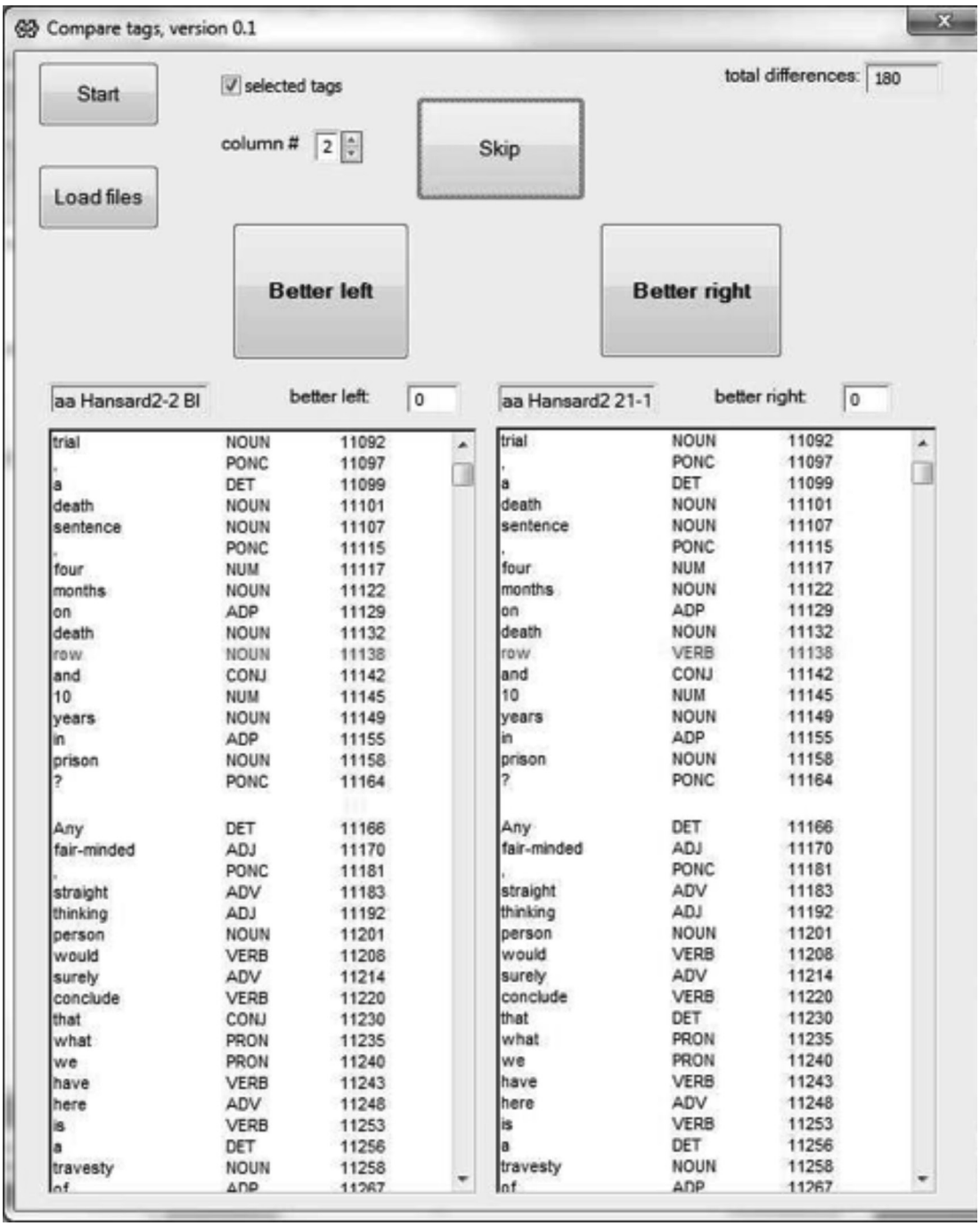

TABLE 4

POS-tagging with and without collocation knowledge

	with collocations	without collocations
complete analyses	73.41%	72.95%
POS-tag differences	727	
better tags	382	106
number of collocation	1668	0

4. Conclusion

Collocations, and more generally MWEs, constitute a fundamental property of natural language. They also have a considerable impact on NLP. We have shown how Fips, a grammar-based symbolic parser, can handle collocations in a wide- range of syntactic configurations, no matter how distant the two constituents of the collocation can be. Such examples clearly show that the identification of colloca- tions crucially depends on a very detailed syntactic analysis, including anaphora resolution. We have also argued that the identification of collocations should be done as soon as possible, during the parsing process, so that the parser can benefi from the collocation knowledge, for instance to disambiguate words. Our eval- uation, comparing analyses obtained with and without collocational knowledge showed a clear improvement of the quality when the parser has access to such knowledge.

Acknowledgments

The work reported in this paper has been developed over several years by many collaborators and graduate students at LATL, including Violeta Seretan and Luka Nerima, as well as Jorge Antonio Leoni di Leo´ n, Sharid Loa´ iciga and Mercedes Villalobos Cardozo, the three of them from the University of Costa Rica. Many thanks to Vasiliki Foufi for helpful comments.

Notes

1. In the case of named entities – a boundless class – most of them should probably be listed in domain-specific (or application-specific) lexicons. We will not pursue this matter further in this paper.

2. Notice that the identification of pronominal verbs in German or in Romance languages is quite similar. Here too, we have a particular lexeme constituted of two elements – the verb and the pronoun – which may not be adjacent.

3. See Wehrli (2007), Wehrli & Nerima (2015) for a description of the Fips parser.

4. Alternatively, one might consider that the identification could be delayed until the end of the parsing process. This, however, would prevent the parser from exploiting collocational knowledge for instance as heuristics to rank alternatives.

5. NP stands for 'noun phrase', VP for 'verb phrase', AP for 'adjectival phrase' and AdvP for 'adverb phrase'. The Fips grammar also uses the labels TP for 'tense phrase', DP for 'determiner phrase'.

6. In Chomsky's view, the *wh*-phrase moves from its "original" position to the initial position, leaving a trace (the empty category) behind.

7. See Wehrli & Nerima (2013) for more details about the anaphora procedure developed for Fips.

8. A complete analysis is one for which the parser manages to build a complete phrase-structure covering the whole sentence. When the parser cannot achieve a complete analysis, it outputs a sequence of chunks – usually 2 or 3 – covering the whole sentence. Notice that although a complete analysis doesn't necessarily mean a correct analysis, it is nevertheless a fairly good measure of the quality of the analysis.

References

Bally, Ch. 1909 [1951]. *Traite´ de stylistique franc¸aise*, Paris, Klincksieck.

Chomsky, N. 1977. "On wh-movement" in P. Culicover, T. Wasow & A. Akmajian (eds.) *Formal Syntax*, Academic Press.

Church, K. & R. Patil, 1982. "Coping with Syntactic Ambiguity or How to Pu the Block in the Box on the Table", *American Journal of Computationa Linguistics*, vol. 8, number 3-4, 139-150.

Petrov, S., D. Das & R. McDonald, 2012. "A Universal Part-of-Speech Tagset", *Proceedings of LREC-2011.*

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2002), "Multiword Expressions: A Pain in the Neck for NLP", *Proceedings of Cicling 2002* Springer-Verlag.

Seretan, V., 2011. *Syntax-Based Collocation Extraction*, Springer Verlag.

Seretan, V. & E. Wehrli, 2009. "Multilingual Collocation Extraction with a Syn- tactic Parser", *Language Resources and Evaluation* 43:1, 71-85.

Tutin, A. & F. Grossmann, 2002. "Collocations re´gulie` res et irre´gulie` res: esquisse de typologie du phe´ nome` ne collocatif", *Revue Franc¸aise de Linguistique Applique´ e, Lexique : recherches actuelles*, Vol. VII, 7-25.

Wehrli, E., 2007. "Fips, a deep linguistic multilingual parser" in *Proceedings o the ACL 2007 Workshop on Deep Linguistic Processing*, Prague, Czech Republic, 120-127.

Wehrli, E. & L. Nerima, 2013. "Anaphora Resolution, Collocations and Translation" in J. Monti, R. Mitkov, G. Corpas Pastor & V. Seretan (eds.) *Pro- ceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, Nice.

Wehrli, E. & L. Nerima, 2015. "The Fips Multilingual Parser", in N. Gala, R.

Rapp and G. Bel-Enguix (eds.) *Language Production, Cognition, and the Lexicon*, Text, Speech and Language Technology 48, Springer, 473-489.