Paraphrasing of Italian Support Verb Constructions based on Lexical and Grammatical Resources

Konstantinos Chatzitheodorou
Aristotle University of Thessaloniki
University Campus, 54124, Thessaloniki, Greece
chatzik@itl.auth.gr

Abstract

Support verb constructions (SVC), are verb-noun complexes which play a role in many natural language processing (NLP) tasks, such as Machine Translation (MT). They can be paraphrased with a full verb, preserving its meaning, improving at the same time the MT raw output. In this paper, we discuss the creation of linguistic resources namely a set of dictionaries and rules that can identify and paraphrase Italian SVCs. We propose a paraphrasing computational method that is based on open-source tools and data such as NooJ linguistic environment and OpenLogos MT system. We focus on pre-processing the data that will be machine translated, but our methodology can also be applied in other fields in NLP. Our results show that linguistic knowledge constitutes a 95.5% precision rate in identifying SVC and an 88.8% precision rate in paraphrasing SVCs into full verbs.

1 Introduction

NLP systems, particularly statistical MT (Brown et al., 1993) need very large corpora in order to produce high quality results. In less-resourced language pairs, many words may occur infrequently, so the estimation of the word alignments can be inaccurate. Furthermore, multiword expressions are still a *hot potato* area for an MT system either statistical or rule-based (Bannard and Callison-Burch, 2005).

A possible technique to resolve all those problems is to generate paraphrases. Paraphrases are alternative ways of expressing the same information within one or more languages (Callison-burch, 2007). The benefits of paraphrasing are multiple: the unknown words will be reduced, the MT output will be better understandable, the accuracy of the meaning will be the same etc.

In MT, paraphrases help to create a more fluent translation and are valuable in the evaluation of MT results (Zhou et al., 2006). Additionally, paraphrases encourage the end user to understand better the main idea of a given text and improve the linguistic level of the text in general, because it is better to express an idea using a full verb than a support verb that has no meaning and a noun.

In this paper, we focus our discussion on paraphrasing Italian SVCs and we propose a computational model for producing monolingual paraphrases. The sentence (1) is an example of a SVC, while the sentence (2) is its paraphrase. The sentence (1) consists of a support verb (*fare*) "make" and a noun (*viaggio*) "trip" that is the head of the sentence. In sentence (2) we observe that the SVC is replaced by a verb, which is the verbal form of the noun. Hence, the SVC of the sentence (1) semantically corresponds to the full verb of the sentence (2).

1. *Mario* ***fa un viaggio*** *negli Stati Uniti d'America.* "Mario **makes a trip** in the United States of America."

2. *Mario* ***viaggia*** *negli Stati Uniti d'America.* "Mario **travels** in the United States of America."

To generate this type of paraphrases, we use semi-automatic methods. On the one hand, the result will be improved and the whole procedure does not take long time to create the linguistic resources. On the other hand, it is not as simple as it may seem, taking into account many both decisions depend on the features of both the support verb and the nominalised verb.

The paper is organised as follows. Section 2 represents the past related work on paraphrasing. Section 3 describes the theoretical background on SVC and Section 4 the linguistic resources and tools used for creating the module. In Section 5, we state our method, explaining step-by-step how the SVC are identified and paraphrased, as well as the obtained results in Section 6. Finally, Section 7 concludes and discusses our work.

2 Related work

In literature, there are many published studies about paraphrasing SVCs. Research methods range from manual linguistic and lexicographic work to automatic NLP-oriented studies. Related work on paraphrasing includes MT, Question Answering, Information Extraction and Text Mining, Summarisation etc.

On the automatic side Bannard and Callison-Burch (2005) use statistical methods in order to acquire paraphrases that will improve the MT output. They use bilingual corpora for extracting the monolingual paraphrases by pivoting through phrases among the two languages. According to their method, if the X is an English phrase and Y its Italian paraphrase and T another possible paraphrase of Y, then, T is equal to X, so it is the paraphrase of X. Other studies (Barzilay and McKeown, 2001; Pang et al., 2003) have used monolingual parallel corpora, such as translations of classic novels in order to automatically generate the paraphrases.

Dictionary and ruled-based paraphrasing is less popular because it requires linguistic knowledge and time. However, Bareiro and Cabral (2009) present ReEscreve, a system that generates monolingual (in Portuguese) paraphrases using resources from OpenLogos MT system. Even if OpenLogos is an old MT system its lexical resources, grammatical rules and syntactic-semantic ontology (SAL) (Scott and Barreiro, 2009) can be applied in many fields in NLP. Other dictionary approaches that can be also used for paraphrasing are WordNet (Fellbaum, 1998; Green et al., 2001) and NOMLEX (Macleod et al., 1997).

3 Support verb constructions

SVCs are predicate noun complexes where the main verb has not a strong value (Gross, 1975). SVCs occur in many languages, such as Italian. For instance, in the Italian phrase *fare un viaggio* the verb *fare* is semantically reduced. In Italian, SVC include verbs like *dare* "give", *avere* "have", *prendere* "take", *essere* "be" etc.

A semantically weak verb is called support verb (Vsup) (Gross, 1975) or light verb (Polenz, 1963). One of its characteristics is that the predicative noun (Npred) is realised as head of a noun phrase. Identifying a SVC is not an easy task and several factors should be taken into consideration. Firstly, they are not frozen expressions because they can be syntactically splitted by a determiner, an adjective or an adverb. For example, *fare un lungo viaggio* "make a long trip". Secondly, there are constructions with the same structure but they are fake (pseudo SVCs). For example, *fare una banca* "make a bank" looks like a SVC but in that case *fare*'s semantic is not reduced.

Given that the meaning of the SVCs is mainly reflected by the nominal predicate, we paraphrase them by replacing the Vsup with a related full verb generated from the predicate noun. For instance, the phrase *faccio una telefonata a Maria* "make a call to Maria" can be simply paraphrased as *telefono a Maria* "I call Maria". The idea behind this methodology is to pre-process a text that will then be translated by a MT Engine so a better MT output will be archived.

4 Linguistic resources and tools

4.1 OpenLogos

OpenLogos is an open source program that machine translates from English and German into French, Italian, Spanish and Portuguese. The system was created by Scott in 1970 but then has been extended by the German Research Centre for Artificial Intelligence (DFKI). It is an old rule-based system MT, but its resources, such as the electronic dictionary, the rules and the SAL which is embedded in the dictionaries, are valuable (Barreiro et al., 2011).

In our work we use only the electronic dictionaries including the SAL, in order to implement a module that will identify and automatically paraphrase SVCs.

4.2 NooJ

As mentioned above, our goal is to implement linguistic resources, tools and methodologies that can be used in automatic processing of SVC and in exporting paraphrases. In this paper, we are presenting only SVCs that consist of the Vsup *fare*.

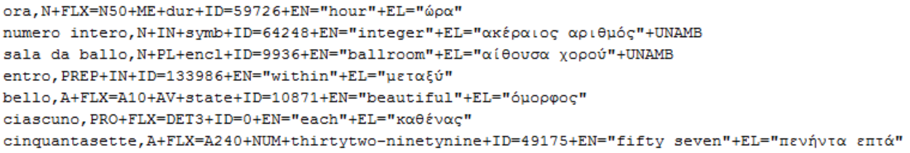
The main linguistic tool for recognising and paraphrasing SVCs is NooJ (Silberztein, 2003). NooJ is a freeware, linguistic-engineering development environment implemented for formalising various types of textual phenomena such as orthography, lexical and productive morphology, local, structural and transformational syntax. It contains several modules that include large coverage lexical resources such as dictionaries for specific purposes and local grammars that are represented by finite-state transducers for many different languages. Its electronic dictionaries contain the lemmas with a set of information, such as:

$$\text{lemma,(1)+(2)+(3)+}\ldots\text{(4)+}\ldots$$

where (1) corresponds to the category/part-of-speech (e.g. "Ver"), (2) to one or more inflectional and/or derivational paradigms (e.g. how to conjugate verbs, how to nominalise them etc.), (3) to one or more syntactic properties (e.g. "+transitive" or "+PREPin") and finally, (4) to one or more semantic properties (e.g. distributional classes such as "+Human", domain classes such as "+Politics").

Our module consists of specific local grammars and electronic dictionaries in order to recognise paraphrase and translate SVCs, such as *fare una presentazione* "make a presentation" → *presentare* "present". In order to process SVCs, we first converted the OpenLogos dictionary into NooJ format. Each lemma is associated with the category, the inflectional paradigm, the equivalent in English and attributes from SAL ontology. There are also some lemmas containing the Greek equivalent that will help for further research.

Figure 1 illustrates a sample of the electronic dictionariy that consists of 75509 entries. 20501 of them are nouns (2335 of them are proper names and toponyms), 10910 are verbs, 22193 are adjectives, 4621 are adverbs, 151 are conjunctions, 5 are determinatives, 295 are prepositions and 118 are pronouns. 14380 over 75509 lemmas are multiword expressions.

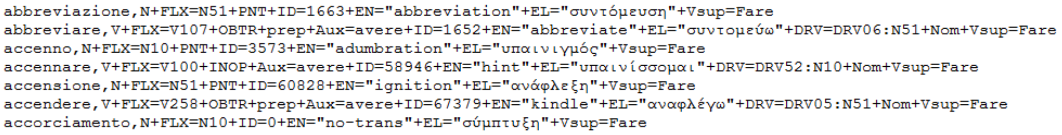


```
ora,N+FLX=N50+ME+dur+ID=59726+EN="hour"+EL="ώρα"
numero intero,N+IN+symb+ID=64248+EN="integer"+EL="ακέραιος αριθμός"+UNAMB
sala da ballo,N+PL+encl+ID=9936+EN="ballroom"+EL="αίθουσα χορού"+UNAMB
entro,PREP+IN+ID=133986+EN="within"+EL="μεταξύ"
bello,A+FLX=A10+AV+state+ID=10871+EN="beautiful"+EL="όμορφος"
ciascuno,PRO+FLX=DET3+ID=0+EN="each"+EL="καθένας"
cinquantasette,A+FLX=A240+NUM+thirtytwo-ninetynine+ID=49175+EN="fifty seven"+EL="πενήντα επτά"
```

Figure 1: NooJ electronic dictionary entries.

Additionally, for the verbs that can be nominalised we created manually its derivational paradigm and the Greek equivalent. Applying a derivational paradigm to a given word is possible to change its syntactic category but not its semantic value. In total, 78 derivational paradigms were created for 289 verbs. For instance, the affix *–zione* changes the verb *presentare* into the noun *presentazione* and the affix *–ata* change the verb *telefonare* to the noun *telefonata*. This is extremely important, in order to generate the paraphrases. Figure 2 illustrates dictionary verb and noun entries that are linked to SVC with the support verb *fare*.

Moreover, it was needed to create from scratch inflectional grammars and other syntactic grammars in NooJ format in order to disambiguate the Italian language.



```
abbreviazione,N+FLX=N51+PNT+ID=1663+EN="abbreviation"+EL="συντόμευση"+Vsup=Fare
abbreviare,V+FLX=V107+OBTR+prep+Aux=avere+ID=1652+EN="abbreviate"+EL="συντομεύω"+DRV=DRV06:N51+Nom+Vsup=Fare
accenno,N+FLX=N10+PNT+ID=3573+EN="adumbration"+EL="υπαινιγμός"+Vsup=Fare
accennare,V+FLX=V100+INOP+Aux=avere+ID=58946+EN="hint"+EL="υπαινίσσομαι"+DRV=DRV52:N10+Nom+Vsup=Fare
accensione,N+FLX=N51+PNT+ID=60828+EN="ignition"+EL="ανάφλεξη"+Vsup=Fare
accendere,V+FLX=V258+OBTR+prep+Aux=avere+ID=67379+EN="kindle"+EL="αναφλέγω"+DRV=DRV05:N51+Nom+Vsup=Fare
accorciamento,N+FLX=N10+ID=0+EN="no-trans"+EL="σύμπτυξη"+Vsup=Fare
```

Figure 2: NooJ electronic dictionary entries.

5 Automated processing of SVCs

5.1 Identification of SVCs

To identify and extract paraphrases for SVCs, we updated OpenLogos dictionaries with morfo-syntactic-semantic information and with derivational and distributional properties as well. This was necessary due to new words that were added in the Italian vocabulary in the last years. We have also created local grammars that are combined with the electronic dictionaries.

We firstly focused on identifying the SVC and updating the existing dictionaries. We obtained that by designing a simple local grammar, that recognises and annotates SVCs and their predicate nouns (see Figure 3). The grammar checks for a verb *fare* followed optionally by a determiner <DET>, adjective <A> or adverb <ADV> and a noun <N>, and annotates it as a SVC (<**SVC=+Pred=$N_**>).

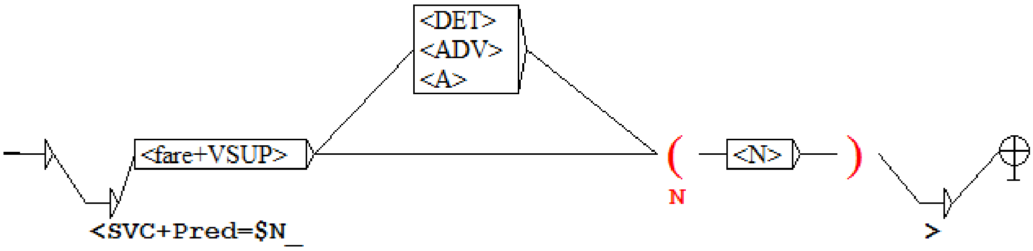
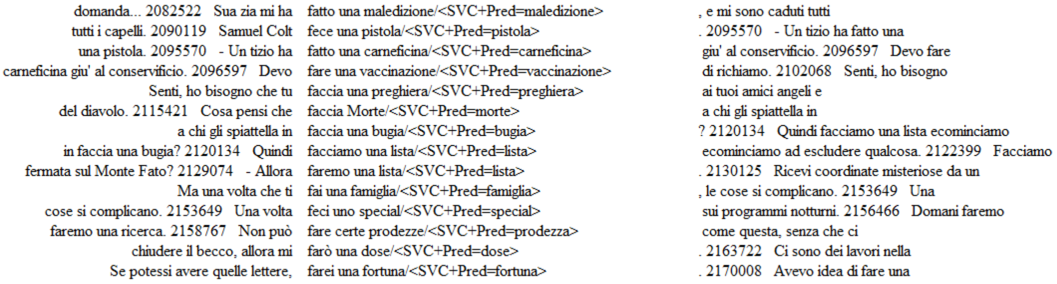


Figure 3: NooJ local grammar for recognizing and annotating SVCs and their predicates.

We applied that grammar to the Italian monolingual Europarl corpus (Koehn, 2005) in order to extract the lemmas of the predicate noun (**$N_**). Then, we updated manually the electronic dictionary by adding the new predicate nouns. We also associated every new predicate to a corresponding lexical full verb and every verb with a derivational paradigm (see Figure 4).



```
                 domanda... 2082522   Sua zia mi ha   fatto una maledizione/<SVC+Pred=maledizione>      , e mi sono caduti tutti
                 tutti i capelli. 2090119   Samuel Colt   fece una pistola/<SVC+Pred=pistola>               . 2095570   - Un tizio ha fatto una
                   una pistola. 2095570   - Un tizio ha   fatto una carneficina/<SVC+Pred=carneficina>      giu' al conservificio. 2096597   Devo fare
  carneficina giu' al conservificio. 2096597   Devo   fare una vaccinazione/<SVC+Pred=vaccinazione>     di richiamo. 2102068   Senti, ho bisogno
                         Senti, ho bisogno che tu   faccia una preghiera/<SVC+Pred=preghiera>         ai tuoi amici angeli e
              del diavolo. 2115421   Cosa pensi che   faccia Morte/<SVC+Pred=morte>                     a chi gli spiattella in
                              a chi gli spiattella in   faccia una bugia/<SVC+Pred=bugia>                 ? 2120134   Quindi facciamo una lista ecominciamo
             in faccia una bugia? 2120134   Quindi   facciamo una lista/<SVC+Pred=lista>               ecominciamo ad escludere qualcosa. 2122399   Facciamo
       fermata sul Monte Fato? 2129074   - Allora   faremo una lista/<SVC+Pred=lista>                 . 2130125   Ricevi coordinate misteriose da un
                              Ma una volta che ti   fai una famiglia/<SVC+Pred=famiglia>              , le cose si complicano. 2153649   Una
          cose si complicano. 2153649   Una volta   feci uno special/<SVC+Pred=special>               sui programmi notturni. 2156466   Domani faremo
            faremo una ricerca. 2158767   Non può   fare certe prodezze/<SVC+Pred=prodezza>           come questa, senza che ci
                       chiudere il becco, allora mi   farò una dose/<SVC+Pred=dose>                     . 2163722   Ci sono dei lavori nella
                   Se potessi avere quelle lettere,   farei una fortuna/<SVC+Pred=fortuna>              . 2170008   Avevo idea di fare una
```

Figure 4: NooJ concordance after annotation of SVCs and identification of the lemma of the predicate nouns.

5.2 Paraphrasing

After updating the electronic dictionaries, more monolingual paraphrases can be obtained easily. Figure 5 represents a local grammar used to recognise, generate SVCs and transform them into their verbal paraphrases. The grammar checks for the verb *fare* in present indicative tense followed by a <DET>, an <A> or an <ADV> and a noun, and generates the verbal paraphrases in the same tense. Furthermore, we restrict our research to Vsup *fare* but the same methodology can be apply to other SVCs. The same structure follow the grammars created for the other grammatical tenses and moods. The elements <**$V=:fare+PR+1+s**>, and **$N_PR+1+s** represent lexical constraints that are displayed in the output, such as specification of the support verb that belongs to a specific SVC. The predicate noun is identified, mapped to its deriver and displayed as a full verb while the other elements of the phrase are eliminated. Figure 6 shows a NooJ concordance were Italian SVCs are identified and paraphrased as full verbs.

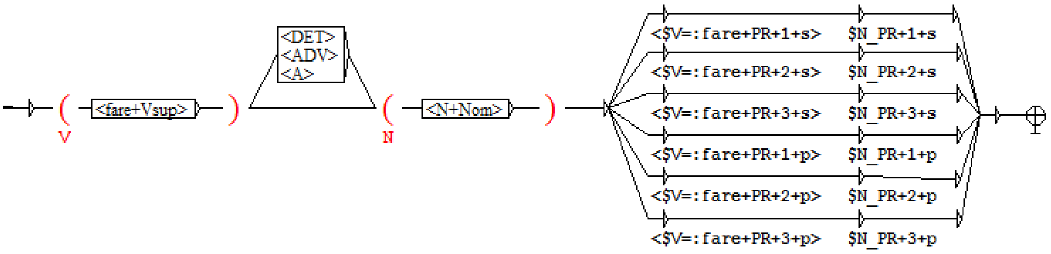
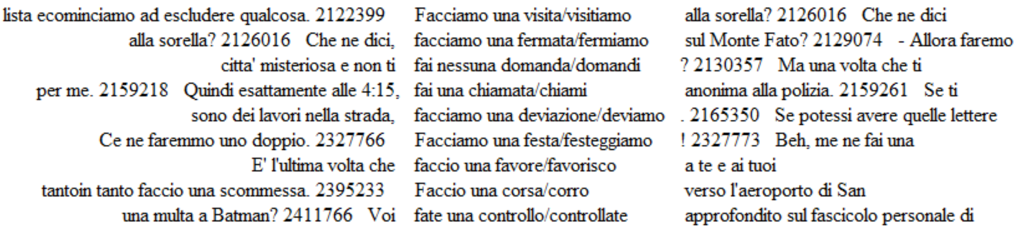


Figure 5: NooJ local grammar for paraphrasing SVCs.



lista ecominciamo ad escludere qualcosa. 2122399	Facciamo una visita/visitiamo	alla sorella? 2126016 Che ne dici
alla sorella? 2126016 Che ne dici,	facciamo una fermata/fermiamo	sul Monte Fato? 2129074 - Allora faremo
citta' misteriosa e non ti	fai nessuna domanda/domandi	? 2130357 Ma una volta che ti
per me. 2159218 Quindi esattamente alle 4:15,	fai una chiamata/chiami	anonima alla polizia. 2159261 Se ti
sono dei lavori nella strada,	facciamo una deviazione/deviamo	. 2165350 Se potessi avere quelle lettere
Ce ne faremmo uno doppio. 2327766	Facciamo una festa/festeggiamo	! 2327773 Beh, me ne fai una
E' l'ultima volta che	faccio una favore/favorisco	a te e ai tuoi
tantoin tanto faccio una scommessa. 2395233	Faccio una corsa/corro	verso l'aeroporto di San
una multa a Batman? 2411766 Voi	fate una controllo/controllate	approfondito sul fascicolo personale di

Figure 6: NooJ concordance for paraphrasing.

6 Evaluation

We performed a manual evaluation by judging the precision and the recall of 100 phrases that include the *fare*. We should notice that only 95 of them were containing SVCs while the other 5 contain the verb *fare* followed by a non predicate noun, hence they cannot be paraphrased. This test set was extracted radomly from the Italian OpenSubtitles corpus (Tiedemann, 2004). Table 1 details the evaluation results of recognition and paraphrasing of SVCs. We calculated the results for recognising and paraphrasing given that a recognised SVC is not always paraphrased correctly. We observe that our module can recognise 86 over 90 SVC that means a precision rate of 95.5%. Regarding recall, 86 over 95 SVCs were recognised so, an 90.5% rate was obtained. On the other hand, a precision rate of 88.8% (80/90) and a recall rate of 84.2% (80/95) were obtained for the generated paraphrases. The F-measure for recognising is 92.93 while for paraphrashing is 86.43.

According to Bareiro and Cabral (2009), MT performs better when translating full verbs over SVCs. We translated in Google Translate[1] the same test set both with SVC and its paraphrases and then we calculated the BLEU score (Papineni et al., 2002) having as reference the English version (with a single reference translation). Even if the test set is small for an automatic evaluation, results show an improvement of 0.6 BLEU points when we pre-process the data paraphrasing. In more detail, the obtained BLEU score for the original test set is 42.76 while for the paraphrased is 43.36.

	Precision	**Recall**
Identifing	86/90	86/95
Paraphrasing	80/90	80/95

Table 1: Human evaluation results.

[1] `https://translate.google.com/`.

The evaluation results clearly show that paraphrasing can improve the quality of MT. We expect that the low recall scores could be higher upon the improvement of the electronic dictionaries and local grammars.

7 Conclusions and Outlook

In this paper, we present a SVC-based paraphrasing framework that uses existing tools and technologies and hand crafted additions for purposes of increasing translation accuracy. Our methodology archived a precision of 95.5% and a recall of 90.5% in identifying and a precision of 88.8% and a recall of 84.2% in paraphrasing. We also applied our method in a freely available MT system and results show a significant improvement.

To make our paraphrasing methodology more accurate, further analysis and work on electronic dictionaries is needed. Especially, we need to work on the pseudo *fare* SVCs. Furthermore, our work should focus on paraphrasing SVC with full verb that is not associated to the predicate noun such as *fare una sigaretta* "make a cigarette" → *fumare* "smoke". Last but not least, the graphs should be extended in order to not discard the adverbs and adjectives that are included in the SVCs. In that case, the MT quality will be more accurate.

In future research, we are also willing to extend the local grammars and dictionaries in order to generate bilingual paraphrases in other languages such as Greek and English. For instance, *fare una presentazione → to present* in English or *fare una presentazione* → $\pi\alpha\rho o\upsilon\sigma\iota\alpha\zeta\omega$ in Greek.

References

Colin Bannard and Chris Callison-Burch. 2005. *Paraphrasing with bilingual parallel corpora.* In ACL-2005.

Anabela Barreiro, Bernard Scott, Walter Kasper and Bernd Kiefer. 2011. *OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization.* Machine Translation, volume 25 number 2, Pages 107-126, Springer, Heidelberg, 2011. ISSN: 0922-6567. DOI: 10.1007/s10590-011-9091-z.

Anabela Barreiro and Lus Miguel Cabral. 2009. *ReEscreve: a translator-friendly multi-pupose paraphrasing software tool.* MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT, August 29, 2009, Ottawa, Ontario, Canada.

Regina Barzilay and Kathleen McKeown. 2001. *Extracting paraphrases from a parallel corpus.* In ACL-2001.

Peter F. Brown and Vincent J.Della Pietra and Stephen A. Della Pietra and Robert. L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation.* In Computational Linguistics.

Chris Callison-burch. 2007. *Paraphrasing and Translation.* PhD Thesis, University of Edinburgh.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Rebecca Green, Lisa Pearl and Bonnie J. Dorr. 2001. *Mapping WordNet Senses to a Lexical Database of Verbs.* In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 244251, Toulouse, France.

Maurice Gross. 1975. *Mthodes en Syntaxe.* Paris: Hermann.

Philipp Koehn. 2005. *A Europarl: Parallel Corpus for Statistical Machine Translation.* MT Summit.

Catherine Macleod, Adam Meyers, Ralph Grishman, Leslie Barrett, Ruth Reeves. 1997. *Designing a Dictionary of Derived Nominals.* Proceedings of Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, September, 1997.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. *Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences.* In Proceedings of HLT/NAACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation.* In ACL-2002: 40th Annual meeting of the Association for Computational Linguistics.

Peter von Polenz. 1963. *Funktionsverben im heutigen Deutsch.* Sprache in der rationalisierten Welt, Dsseldorf, Schwann.

Bernard Scott and Anabela Barreiro. 2009. *OpenLogos MT and the SAL representation language*. In Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation / Edited by Juan Antonio Prez-Ortiz, Felipe Snchez-Martnez, Francis M. Tyers. Alicante, Spain: Universidad de Alicante. Departamento de Lenguajes y Sistemas Informticos. 23 November 2009, pp. 1926.

Max Silberztein. 2003. *NooJ Manual*. Available for download at: `www.nooj4nlp.net`.

Jorg Tiedemann and Lars Nygaard. 2004. *The OPUS corpus – parallel & free*. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal, May 26-28.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. *Paraeval: Using paraphrases to evaluate summaries automatically*. In Proceedings of HLT/NAACL.