Morpheme-Based Recognition and Translation of Medical Terms

Alessandro Maisto(✉) and Raffaele Guarasci

Dipartimento di Scienze Politiche, Sociali e della Comunicazione,
Università degli Studi di Salerno,
Via Giovanni Paolo II, 84084 Fisciano (SA), Italy
{amaisto, rguarasci}@unisa.it

Abstract. In this paper we use Nooj to solve a recognition and translation task on medical terms with a morphosemantic approach. The Medical domain is characterized by a huge number of different terms that appear in corpora with very low frequencies. For this reason, machine learning or statistical approaches do not achieve good results on this domain. In our work we apply a morpho-semantic approach that take advantage from a number of Italian and English word-formation strategies for the automatic analysis of Italian words and for the generation of Italian/English bilingual lexicons in the medical sub-code. Using Nooj we built a series of Italian and bilingual dictionaries of morphemes, a set of morphological grammars that specify how morphemes combine with each other, a syntactic grammar for the recognition of compound terms and a Finite State Transducer (FST) for the translation of medical terms based on morphemes. This approach produces as output: a categorized Italian electronic dictionary of medical simple words, provided with labels specifying the meaning of each term; a Thesaurus of simples and compounds medical terms, organized in 22 medical subcategories; A an Italian/English translation of medical terms.

Keywords: Morpho-semantic · Medical domain · Translation · Recognition · NooJ · Finite state automatas

1 Introduction

The technical-scientific language of the medicine, provided with a number of technical lemmas that is larger than any other sub-code, is a part of the set of sub-codes that are organized in taxonomies and strong notional fields. Each term of this huge sub-dictionary, besides, occurs in texts with a very low frequency. For this reason, the majority of medical sub-domain terms could be defined as "rare events" (Möbius 2003). This phenomenon could has a negative impact on the performances of the statistical and the machine learning methods. In general, free lexical resources for the medical domain, are often few and incomplete for every kind of language. Multilingual resources, in addition, are very rare and have a crucial role in every NLP systems. The idea of the paper is to approach this large number of medical terms starting from a restricted dictionary (about 1000) of morphemes that, combined one another, allow the recognition of a huge number of terms, at least in two languages: Italian, English. This kind of approach, called Morpho-semantics, can be used to describe, in an analytical

T. Okrut et al. (Eds.): NooJ 2015, CCIS 607, pp. 172–181, 2016.
DOI: 10.1007/978-3-319-42471-2_15

way, the meaning of the words that belong to the same subdomain or to the same "morphological family" (e.g. words: *iper-acusia*, *ipo-acusia*; subdomain: *-acusia* "otolaryngology"; description: *ipo-* "lack", *iper-* "excess", etc.). We grounded the automatic creation of medical lexical databases on specific formative elements that are able to define a meaning in a univocal way, thanks to the regular combination of modules defined independently. Such elements do not represent mere terminations, but possess their own semantic self-sufficiency (Iacobini 2004). In order to build a multilingual medical thesaurus in which every lemma is automatically associated with its own terminological and semantic properties and with the respective English translations we created a small *NooJ* (Silberztein 2003) dictionaries of morphemes. Morphemes may belong to three morphological categories, *Prefixes*, *Confixes*, and *Suffixes*, which are provided with semantic annotations (explaining the meaning of the morpheme), terminological annotations (that refers to the medical class to which the morpheme belongs to) and with the translation of the morpheme in the other language (e.g. *iper*, "hyper"). A Morphological Grammar finds every possible combination of Prefixes, Confixes and Suffixes and annotates the recognized medical term separating it in different units, according with the morphemes that compose the words. Two corpora of Italian Medical Records have been analyzed with this resources configuration and, later, a syntactic translation grammar has been applied: for every combination of morphemes, the grammar transcribes as output the English transduction of the morpheme.

2 Related Work

Morpho-semantic approaches have been already applied to the medical domain in many languages. Works that deserve to be mentioned are Pratt (Pratt and Pacak 1969) on the identification and on the transformation of terminal morphemes in the English medical dictionary; Wolff (Wolff 1984) on the classification of the medical lexicon based on formative elements of Latin and Greek origin; Pacak et al. (Pacak et al. 1980) on the diseases words ending in *-itis*; Norton e Pacak (Norton and Pacak 1983) on the surgical operation words ending in *-ectomy* or *-stomy*; (Dujols et al. 1991) on the suffix *-osis*.

Between the nineties and the 2000, many studies have been published on the automatic population of thesauri, we recollect among others (Lovis et al. 1995), that derived the meaning of the words from the morphemes that compose them; (Lovis et al. 1998) that identified ICD codes in diagnoses written in different languages; (Hahn et al. 2001) that segmented the subwords in order to recognize and extract medical documents; and (Grabar e Zweigenbaum 2000) that used machine learning methods on the morphological data of the thesaurus SNOWMED (French, Russian, English). An advantage of the morphosemantic method is that complex linguistic analyses designed for a language can be often transferred to other languages. (Deléger et al. 2007), as an example, adapted the morphosemantic analyzer DériF (Namer 2009), designed for the French language, for the automatic analysis of English medical neoclassical compounds. (Amato et al. 2014) present a system for morpho-semantic classification of medical simple and compound terms that use Nooj dictionaries and Grammars in order to create a medical thesaurus.

As regards morphological approaches in machine translation tasks, we mention a lexical morphology based Italian-French MT tool (Cartoni 2009), that implemented lexical morphology principles into an Italian-French machine translation tool, to manage computational treatment of neologisms. We than consider (Toutanova 2008) and (Minkov et al. 2007), that proposed models for the prediction of inflected word forms for the generation of morphologically rich languages (e.g. Russian and Arabic) into a machine translation context. Furthermore, (Virpioja et al. 2007) exploited the *Morfessor* algorithm, a method for the unsupervised morph-tokens analysis, with the purpose of reducing the size of the lexicon and improving the ability to generalize in machine translation tasks. Their approach, which basically treated morphemes as word-tokens, has been tested on the Danish, Finnish, and Swedish languages.

(Daumake et al. 1999) exploited a set of subwords (morphologically meaningful units) to automatically translate biomedical terms from German to English, with the purpose to morphologically reduce the number of lexical entries to sufficiently cover a specific domain. (Lee 2004) explored a novel morphological analysis technique that involved languages with highly asymmetrical morphological structures (e.g. Arabic and English) in order to improve the results of statistical machine translations. In the end, (Amtrup 2003) proposed a method that involved finite state technologies for the morphological analysis and generation tasks compatible with Machine Translation systems.

3 Methodology

The morpho-semantic approach allows the analytical description of the meaning of the words that belong to the same subdomain or to the same "morphological family" (Jacquemin 1999).

Because of its frequency distribution (very large number of different terms that appear in texts with a very low frequency), medical terms could be considered as "rare event". This feature of the medical domain has a strong impact on the performances of the statistical and the machine learning methods, and, for this reason, the technical-scientific language of the medicine, rich of technical lemmas, in great part derived from neoclassical terms, is especially adapt to a morpho-semantic approach.

Our approach allow to manage a very large number of medical terms, starting from a restricted dictionary (about 1000) of morphemes pertaining to the domain. Combining these morphemes it is possible to recognize a huge number of terms and translate it into English.

In addition, finding *(almost-)synonym sets* (Namer 2005) on the base of the words that share morphemes endowed with a particular meaning (e.g. *-acusia*, hearing disorders), we can infer the domain of the medical knowledge to which the synonym set belongs (e.g. "otolaryngology") and, in the end, we can differentiate any item of the set by exploiting the meaning of the other morphemes involved in the words.

- synset: *iper-acusia*, *ipo-acusia*, *presbi-acusia*, *dipl-acusia*;
- subdomain: *-acusia* "otolaryngology'";
- description: *ipo-* "lack", *iper-* "excess", *presbi-* "old age", *diplo-* "double".

3.1 Lexical Resources

Thanks to the electronic version of the GRADIT (De Mauro 2003) it has been possible to collect three kind of morphemes related to the medical domain:

- Prefixes
  - quantitative description of the terms (*hyper-*, *hypo-*, *normo-*, *extra-*…)
  - qualitative description (*emo-*, *per-*, *peri-*, *pre-*, *pro-*, *trans-*…)
- Confixes
  - meaning of the single term (acusia, cancero, pulmono…)
- Suffixes
  - meaning of the term (*-oma*, *-asis*, *-itis*…)
  - grammatical category (*-able*, *-aceous*, *-atory*…)

The domain of medicine has been divided into 22 subcategories (cardiology, neurology, gastroenterology, oncology, etc.), and the majority of morphemes has been attributed to one of them.

A class "*undefined*" has been used as residual category, in order to collect the words particularly difficult to classify.

The morpheme dictionary, built with Nooj, has been enriched with other semantic information, concerning the meaning they express.

Each morpheme has been compared with the morphemes presented into the Open Dictionary of English by the *LearnThat Foundation* (https://www.learnthat.org/) and the respective English translation has been added to the NooJ Dictionary.

Furthermore, also other morphemes, that had not been treated by the GRADIT as medical ones, have been added to our list. Table 1 presents a list of morphemes types used in our dictionary.

Table 1. Number and types of morphemes of the Morphenita.nod dictionary

Manner of use	Category	Number	Translated
Medicine	Confixes	451	349
Medicine	Suffixes	14	13
Medicine	Prefixes	7	7
Anatomy	Confixes	45	27
General	Suffixes	19	18

The Nooj Medical Morpheme Dictionary specifies the category of the morpheme (PFX, SFX, etc.), and provides semantic descriptions about the meaning they confer to the words composed with them. Such semantic information regard the three following aspects:

- Meaning: introduced by the code "+Sens", this semantic label describes the specific meaning of the morpheme (e.g. *-oma* corresponds to the descriptions *tumori*, "tumors" and *-ite* to *infiammazioni*, "inflammations");
- Medical Class: introduced by the code "+Med", this terminological label gives information regarding the medical subdomain to which the morpheme belongs

(e.g. *cardio-* let the machine know that every word formed with it pertains to the subdomain of cardiology);
- Translation: introduced by the code "+EN", presents the corresponding translation of the morpheme in English.

The dictionary, compiled into the file *Morphenita.nod*, contains the three categories presented before (CFX for the confixes, SFX for the Suffixes and PFX for the Prefixes) and two new categories (as shown in Fig. 1):

```
#CONFIXES
cardio,CFX+SensCP=cuore+Med=CARDIO+EN=cardio
cerebro,CFX+SensCP=cervello+Med=NEUROEN=cerebro
epitelio,CFX+SensCP=tessutoInterno+Med=INTERN+EN=epithelio
patia,CFX+SensCP=malattia+EN=pathy
toraco,CFX+SensCP=torace+EN=thoraco

#CONFIXES BEFORE THE END
biosi,CFXE+SensCP=vita+EN=biosis
cardio,CFXE+SensCP=cuore+Med=CARDIO+EN=cardium
cerebro,CFXE+SensCP=cervello+Med=NEUROEN=cerebral
epitelio,CFXE+SensCP=tessutoInterno+Med=INTERN+EN=epithelium
toraco,CFXE+SensCP=torace+EN=thorax

#CONFIXES BEFORE SUFFIXES
bronco,CFXS+SensCP=bronchi+Med=PNEUMO+EN=bronch
carcino,CFXS+SensCP=cancro+Med=ONCOL+EN=carcin
cardio,CFXS+SensCP=cuore+Med=CARDIO+EN=cardi
toraco,CFXS+SensCP=torace+EN=thorac

#PREFIXES
emo,PFX+SensP=sangue+EN=hemo
iper,PFX+SensP=eccesso+EN=hyper
ipo,PFX+SensP=poco+EN=hypo

#SUFFIXES
ite,SFX+SensS=infiammazione+EN=itis
oma,SFX+SensS=tumoriInfiammazioniTumefazioni+Med=ONCOL+EN=oma
osi,SFX+SensS=lesione+EN=osis
```

Fig. 1. Extract of the dictionary

- CFXS, that includes all the Confixes that can appear before a suffix, with its correspondent English morpheme deprived of the final part, in order to avoid vocal repetition in case of suffixation. The word *Ateroscelrosi*, "Atherosclerosis", for example, that is composed by three morphemes, *atero*, *sclero* and *osi*, with the respective translation of morphemes, "athero", "sclera" and "osis"; when translated, produces the sequence "atherosclerosis". Since is not possible to operate directly on English morphemes, to prevent these kind of errors, the system contemplates the new category CFXS for Confixes that are followed by a Suffix. While the sequence of morphemes CFX-CFX-SFX produce "Atherscleroosis", a sequence CFX-CFXS-SFX translate correctly the medical term.
- CFXE, that includes all the Confixes that can appear at the end of a world, with the correspondent English morpheme modified ad it appear when close the world. For example, *Emotorace*, in English "Hemotorax", with a normal sequence of CFX-CFX, produce "Hemotoraco". With the sequence CFX-CFXE produce the correct translation.

3.2 Grammars

The creation of the Morphenita.nod dictionary represents the first step of the method: in order to automatically recognize and translate medical words in real text occurrences, needs the support of morphological and syntactic local grammars.

Morphological Grammars. For the recognition of medical terms we use seven parallel morphological grammars, called *Medita#.nom*, that automatically assign semantic tags to the simple words found in free texts, according to the meaning of the formative elements that compose the same words.

The seven grammars built with Nooj include the following combination of morphemes:

1. confixes-confixes or prefixes-confixes or prefixes-confixes-confixes;
2. confixes-suffixes or prefixes-confixes-suffixes;
3. confixes-confixes-suffixes or prefixes-confixes-confixes-suffixes;
4. nouns-confixes;
5. prefixes-nouns-confixes;
6. confixes-nouns-confixes;
7. nouns-suffixes;

In Fig. 2 is presented a sample of the morphological grammar: The code <+MEDICINA$1S$2S$ > allows the grammar to assign to the words the information inherited by the morphemes that compose them.

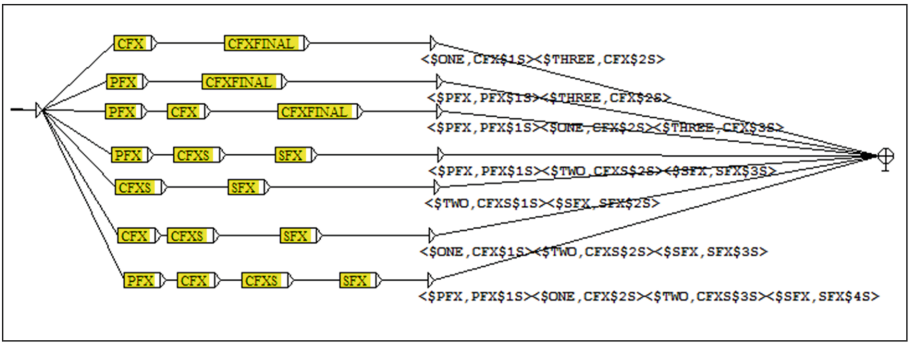

Fig. 2. Sample of morphological grammar *Medita1.nom*

To allows the automatic translation of medical terms, we built another morphological grammar with seven paths (corresponding to the seven grammars used for the classification task), called *MedItEn.nom*, that recognize each morphemes of the word as separate entities, and tag it independently as show in Fig. 3:

Syntactical Grams. In order to extract and classify multiword expressions, we exploited a Nooj syntactic grammar. The one designed for this work, called *MedClass.nog*, includes seven main paths based on different combinations of Nouns (N), Adjectives (A) and Prepositions (PREP) (Fig. 4).

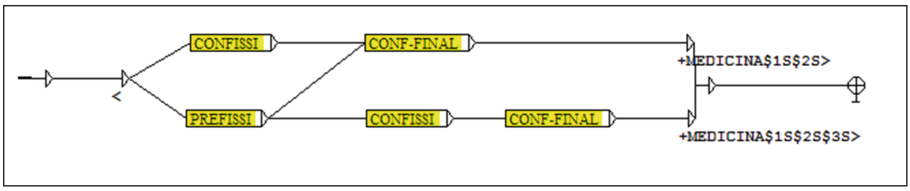

Fig. 3. The morphological Grammar *MedItEn.nom*

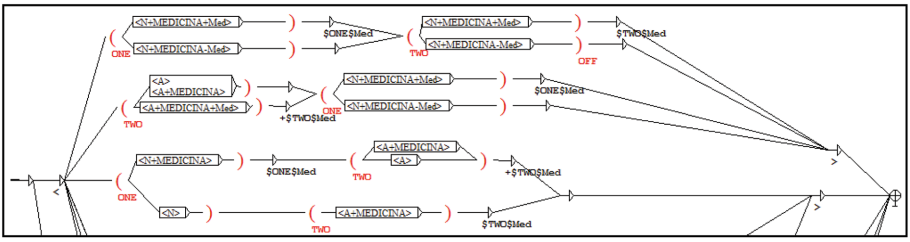

Fig. 4. Extract of the syntactic grammar *MedClass.nog*

1. Noun;
2. Noun + Noun;
3. Adjective + Noun;
4. Noun + Adjective;
5. Noun + Noun + Adjective;
6. Noun + Adjective + Adjective;
7. Noun + Preposition + Noun;

Every path attributes to the matched sequence the label that belongs to the head of the compound. In the case in which the head is not endowed with a semantic label, the compound receives the residual tag "undefined".

For the Translation task, we construct a different syntactic transducer called *Transiten.nog*, that simply consider the recognized morphemes and translate them into the respective English translation specified by the dictionary (Fig. 5).

4 Experimentation

In order to evaluate the precision of our morpho-semantic method, so for the classification task as for the translation task, we experiment them on two different corpora:

- a first corpus of about 5.000 simplified medical records, spliced into 20 subsections with a total of 64.360 tokens and 41.468 annotated word forms
- a second corpus of 330 complete Medical records. 38.696 tokens and 20.261 word forms.

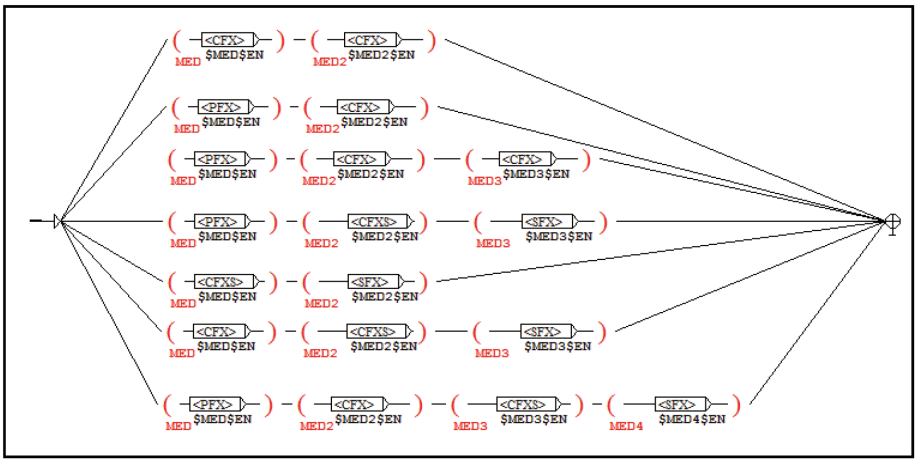

Fig. 5. The Finite State Transducer *Transiten.nog*

The classification task produce as output an Electronic Dictionary of simple medical words and a Thesaurus of simple and compound medical words.

Into the dictionary, the lemmas extracted from the diagnoses are systematically associated with their terminological ("\Med") and semantic ("\Sens") descriptions as in the example:

```
gastrite,N+SensCP=apparatoGastrico+Med=GASTRO+SensS=infiammazione
cardiopatia,N+SensCP=cuore+Med=CARDIO+SensCS=malattia
ipertensiva,A+SensP=eccesso+SensCP=tensione+Med=CARDIO
aortica,A+SensCP=aorta+Med=CARDIO
```

Into the thesaurus, medical terms recognized are grouped together on the base of their medical classes

The precision for the classification task is calculated by subdomain in order to underline strengths and weaknesses of the method in relation with a specific class or group of worlds.

As the Table 2 shows, best results has been obtained with Traumatology, Surgery, Pneumology and Gastroenterology classes. The Endocrinology class is the worst class due to problems with the recognition of the morpheme *Tiroido*, "Thyroid", that could be corrected in future works.

For what concern the translation task, the output is represented by a list of English Medical Terms preceded by the respective Italian words.

The evaluation of translation task has been carried out by comparing our method with google translate for 214 words presented in the corpus. Google obtain a 84,11 % of precision but our method achieve the 74,77 %, but with good performances with neologisms and scientific diseases terms (such as Hemorrhage, «bleeding» for Google Translate). Furthermore, our morpho-semantic method, in combination with Google Translate, reach the 93 % of precision, that is a very good results for a translation task.

Table 2. Precision values

Detected classes	Precision %
Traumatology	100
Surgery	97,82
Pneumology	95,83
Gastroenterology	89,18
Orthopedic	80,95
Urology	76,19
Intern Medicine	69,04
Cardiology	66,96
Endocrinology	23,80
Undefined	50,80
Tot	69,50

5 Conclusions

We presented a morpho-semantic approach for automatic recognition and translation of medical domain terms. For what concern the recognition task, we classify a great number of simple and compound terms with a good total precision value. Furthermore, it will be possible to improve this value working on a few number of morphemes pertaining at the Endocrinology sub-domain and increasing the number of medical morphemes presents into the dictionary. We automatically generate a Thesaurus of medical terms and a dictionary provided with description of the meaning of each term. In addition, this kind of approach do not suffer for the presence of neologisms into the medical corpora.

As seen in the evaluation phase, although the system do not reach the precision of Google in a translation task, it achieve good results in translation of neologisms or scientific disease terms. In future works it could be possible to extend the method to other languages such as Spanish by simply add the Spanish translation of the morphemes present into the dictionary. Moreover, improving the Finite State Translator and the morphological grammar, the precision of the translation task will grow.

References

Amato, F., Elia, A., Maisto, A., Mazzeo, A., Pelosi, S.: Automatic population of italian medical thesauri: a morphosemantic approach. In: 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 432–436. IEEE, Guangzhou (2014)

Amtrup, J.W.: Morphology in machine translation systems: efficient integration of finite state transducers and feature structure descriptions. Mach. Transl. **18**(3), 217–238 (2003)

Cartoni, B.: Lexical morphology in machine translation: a feasibility study. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 130–138. Association for Computational Linguistics (2009)

Daumke, P., Schulz, S., Markó, K.: Subword Approach for Acquiring and Crosslinking Multilingual Specialized Lexicons. Programme Committee (2006)

De Mauro, T.: Nuove Parole Italiane dell'uso, GRADIT, vol. 7 (2003)
Deléger, L., Naner, F., Zweigenbaum, P., et al.: Defining medical words: transposing morphosemantic analysis from French to English (2007)
Dujols, P., Aubas, P., Baylon, C., Grémy, F.: Morpho-semantic analysis and translation of medical compound terms. Methods Inf. Med. **30**(1), 30 (1991)
Hahn, U., Honeck, M., Piotrowski, M., Schulz, S.: Subword segmentation-leveling out morphological variations for medical document retrieval. In: Proceedings of the AMIA Symposium, p. 229. American Medical Informatics Association (2001)
Iacobini, C.: Composizione con elementi neoclassici, in La formazione delle parole in italiano, a cura di Grossmann, M., Rainer, F., pp. 69–95 (2004)
Jacquemin, C.: Syntagmatic and paradigmatic representations of term variation. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 341–348. Association for Computational Linguistics (1999)
Lee, Y.S.: Morphological analysis for statistical machine translation. In: Proceedings of HLT-NAACL 2004: Short Papers, pp. 57–60. Association for Computational Linguistics (2004)
Lovis, C., Baud, R., Rassinoux, A.M., Michel, P.A., Scherrer, J.R.: Medical dictionaries for patient encoding systems: a methodology. Artif. Intell. Med. **14**(1), 201–214 (1998)
Lovis, C., Michel, P.A., Baud, R., Scherrer, J.R.: Word segmentation processing: a way to exponentially extend medical dictionaries. Medinfo **8**(pt 1), 28–32 (1995)
Minkov, E., Toutanova, K., Suzuki, H.: Generating complex morphology for machine translation. ACL **7**, 128–135 (2007)
Möbius, B.: Rare events and closed domains: two delicate concepts in speech synthesis. Int. J. Speech Technol. **6**(1), 57–71 (2003)
Namer, F.: Acquisizione automatica di semantica lessicale in francese: il sistema di trattamento computazionale della formazione delle parole dérif. In: Thornton, A.M, Grossmann, M. (eds.) Atti del XXVII Congresso internazionale di studi Società di Linguistica Italiana: La Formazione delle parole, pp. 369–388 (2005)
Namer, F.: Morphologie, lexique et traitement automatique des langues (2009)
Norton, L., Pacak, M.G.: Morphosemantic analysis of compound word forms denoting surgical procedures. Methods Inf. Med. **22**(1), 29–36 (1983)
Pacak, M.G., Norton, L., Dunham, G.S.: Morphosemantic analysis of-ITIS forms in medical language. Methods Inf. Med. **19**(2), 99–105 (1980)
Pratt, A.W., Pacak, M.: Identification and transformation of terminal morphemes in medical english. Methods Inf. Med. **8**(2), 84–90 (1969)
Silberztein, M.: NooJ manual (2003). www.nooj4nlp.net
Toutanova, K., Suzuki, H., Ruopp, A.: Applying morphology generation models to machine translation. In: ACL, pp. 514–522 (2008)
Virpioja, S., Väyrynen, J.J., Creutz, M., Sadeniemi, M.: Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. Mach. Transl. Summit XI **2007**, 491–498 (2007)
Wolff, S.: The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. Methods Inf. Med. **23**(4), 195–203 (1984)