Summary of Ph.D. dissertation

Ágoston Nagy

Automatic Term Extraction from French Language Patent Descriptions with Rule-based and Statistical Methods

Supervisors:

Tamás Váradi, Ph.D.
senior research fellow

Zsuzsanna Gécseg, Ph.D.
associate professor

University of Szeged
Faculty of Arts
Doctoral School in Linguistics
Ph.D. Programme in French Linguistics

2012

Subject of the dissertation

Automatic term extraction (TE) is a subfield of computational linguistics: its aim is to elaborate an application that extracts automatically term candidates from written plain texts (Jacquemin & Bourrigault 2003). TE is used in document indexing during which expressions characterising a specific text file are extracted. The aim of this extraction process is to have a list of terms and statistics that help categorising text files and searching among them in the future (Enguehard 2005). Another aim of TE is to help the translation work in teams: a list of previously extracted terms, paired with their corresponding expression in the target language, can be really useful before the whole translation process. TE is also used in machine translation applications (e.g. Vasconcellos 2001), and in information extraction (e.g. Ahmad 2001).

Both Hungarian and international research focus on nominal terms because these have the most complex structure, that is why I also concentrated on nominal terms in the present dissertation. A computer programme is also made for the dissertation, which extracts this kind of terms from French language patent descriptions. In order to extract these terms, both rule based (i.e. linguistic) and statistical methods were used.

Theoretical background

The classic term definition is attributed to Wüster (1976, 1981). This is the point of view which is accepted and recognised by the terminological community when the notion of term have to be defined (Petit 2001). According to this approach, the terminological nature of a lexical unit can be determined on the basis of three criteria: (1) a term is attached to one and only one concept (2) that it denotes, and (3) it is connected to a scientific domain.

Patents are texts in which the traits (e.g. excessive usage of terms and that of impersonal structure) typical of scientific texts are dominantly present (Cabré 1999), therefore they should be an ideal corpus for TE. Descriptions are the most detailed parts of patents. Their aim is to describe patents in a brief and concise way, that is why terms occur frequently in these texts because one of the features of terms, according to Justeson & Katz (1995), is that they can rarely be substituted by other terms or shorter structures. On the basis of all these and Cabré et al. (2001), statistical methods may be applied to this kind of texts with good reliability.

According to Cabré et al. (2001) and Sauron (2002), TE applications perform the following steps: (1) extracting candidate terms from a corpus, (2) filtering these candidates and finally (3) validating the remaining candidate terms. Both term extraction and filtering can be achieved by rule-based methods (with predefined syntactic patterns) or by statistical methods. Most TE applications use hybrid methods, because completely rule-based term extractors provoke too much noise (i.e. the number of extracted terms is much higher than that of real terms), statistical methods lead to too much silence (i.e. the list of candidate terms may not contain numerous real terms). In the case of rule-based methods, terms can be extracted on the basis of their internal morphosyntactic structure, therefore compositions typical of terms (e.g. two nouns immediately following each other) have to be searched for. From among the hybrid methods, the preferred combination is to extract terms with statistical methods and to filter them with rule-based methods (e.g. Cabré et al. 2001, Ha et al. 2008).

TE applications use in general three kinds of statistical methods. The first type of statistical methods is constituted by *termhood* values which aim at determining whether a given term is connected rather to specialised language or to general language. In order to calculate this value for a specific term, its relative frequency in general language is also needed. The second type of value is *unithood*, which gives a probability value on the basis of the cohesion force between the elements of a multi-word candidate term, therefore it aims to determine whether it is the whole sequence which is a term, or only a part of it, or it is a part

of a bigger term. The third statistical measure gives a statistical value first to the tokens in the environment of real terms, and from these data, it can help to determine whether a sequence is a term on the basis of its surrounding elements (Wong et al. 2008).

The efficiency of my own TE application is measured by the values typically used in computational linguistics, namely precision, recall and F-value (e.g. Cabré et al. 2001). In TE, precision shows the proportion of real terms and that of all the extracted terms in the list of candidate terms: the higher precision is, the less the list of candidate terms contain non-terminological units. Precision is calculated by the following formula:

$$precision = \frac{number\ of\ real\ terms\ in\ the\ candidate\ term\ list}{number\ of\ extracted\ candidate\ terms}$$

Recall determines the proportion of correctly extracted terms and all existing real terms in the corpus: the higher recall is, the less terms were not recognised. Recall is calculated by the following formula:

$$recall = \frac{number\ of\ correctly\ extracted\ terms}{number\ of\ real\ terms}$$

In computational linguistics and mathematical statistics, F-value is frequently used in order to calculate efficiency, which is the harmonic mean of precision and recall, thus F-value is calculated by the following formula:

$$F\text{-}value = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Internal structure of French nominal terms

Since my own automatic term extractor works on predefined syntactic patterns in order to extract nominal candidate terms, the different syntactic structures of terms have to be precisely defined and the differences between common language NPs and terms have to be shown, especially with respect to prepositional complements (including the preposition+noun sequences of nominal compounds) and adjectival adjuncts. For this purpose, I rely on previous research (e.g. Nagy 2012a, 2012b) and L'Homme (2004).

It is not easy to define the notion of PP in French because it is not always evident whether a PP introduces a new entity inside the NP (1a) or it is associated to the nominal head with which it forms a complex noun (1b). In the case of nominal compounds (1b), the preposition is in general followed by a noun without a determiner because the presence of a determiner would imply a complex NP where the NP preceded by a preposition could be considered as an embedded NP having a separate reference (1a).

(1) a. *le* *moulin* *du* *village*
ART-DEF.MASC.SG mill DE+ART-DEF.MASC.SG village
'the mill of the village'

b. *le* *moulin* *à* *vent*
ART-DEF.MASC.SG mill À wind
'windmill'

Bosredon and Tamba (1991) differentiates the two different prepositional structures: they think that nominal compounds are simple nouns from a semantic point of view but they constitute a NP from a formal point of view. In this way, they distinguish the PPs (1a) from the preposition+noun sequences (1b) that are attached to a noun and they call them constituents and formants, respectively. However, the boundary between formants and

constituents are not so clear from a formal point of view. For example, there are nominal compounds that contain a PP with a determiner (e.g. 2a).

(2) a. *cancer de la peau*
cancer DE ART-DEF.FEM.SG skin
'skin cancer'
b. ?*cancer de peau*

In a previous study (Nagy 2009), it was showed that the proportion of NPs with internal determiner is nearly 7% in comparison with the totality of nominal terms but the proportion of NPs with determiner that can also appear without determiner was not calculated. Consequently, completely aware of the loss that it represents, terms with determiners were not considered as possible terms during the automatic extraction process.

As Cinque (1994) states, French, like most Romance languages, is an ANA language meaning that adjectives can either precede or follow the nominal head in a NP. On the contrary, Germanic languages, like English and German, are AN languages, that is adjectives can only precede the nominal head. The position of adjectives in French depends on their type.

From the four adjective groups (namely categorising, non-categorising, relational and ordinal adjectives according to Riegel et al. 2009), terms can only contain adjectives that designate the subtype of the nominal head they are attached to. From the four adjective types, it is only categorising and relational adjectives that can only follow the nominal head. This is exemplified in (3) in which the underlined part is the term: the second adjective *filaire* 'wired' is relational, the first one (*grand* 'big') is a frequently used, monosyllabic (thus prenominal) adjective, which is not so likely to be part of a term, but can modify it (3).

(3) *un* *grand* *réseau filaire*
ART-INDEF.MASC.SG big-MASC.SG network wired
'big wired network'

Nominal terms do not comprise the determinants preceding them, and contrary to common NPs, they do not contain subordinate clauses. This is a consequence of the three criteria of Sager (1990): (1) economy, (2) precision, (3) appropriateness, and of the fact that a term has to be repeated in the same form according to the definition of Wüster, therefore a nominal term corresponds to the N' projection of generative syntax. For example, if a term contains a relative subordinate clause, it is only possible term elements that have to be kept, for example in (4) from the patent corpus, only *poudre hygroscopique* 'hygroscopic powder' should be maintained as term.

(4) *La poudre, qui est hygroscopique, ...*
ART-DEF.FEM.SG powder which is hygroscopic
'The powder, which is hygroscopic, ... '

Aim of the research

According to Cabré et al. (2001), term extraction tools extract first candidate terms by means of statistical methods, and this list is then filtered with linguistic filters. However, in my term extraction tool, I chose the inverse direction: terms were extracted on the basis of their internal syntactic structure by rule-based extraction, and this list was filtered with statistical methods. By choosing rule-based extraction, I intended to prove that the point of view of scientific literature (e.g. Cabré et al. 2001) is not always applicable in all cases. This choice is confirmed by the following facts: (1) in French, terms tend to have internal structures that are not typical of common language nouns or nominal compositions, therefore rule-based approaches can be applied to them with more efficiency; and (2) that in the case of the most

cited term extractors used for French, extraction is rule-based: for example Acabit (Daille 1994), Lexter (Bourrigault 1994) and Fastr (Jacquemin 2001).

Our aim was to prove that rule-based term extraction achieves a higher recall with lower precision, and that the latter can be ameliorated by rule-based and statistical filtering. Our hypothesis was that if the threshold of statistical values is set to high, precision significantly increases and recall greatly decreases; if the threshold is set to low, F-values can slightly be increased.

Another aim of the dissertation was to show (1) to what extent the statistical values relying on different factors contributed to the increase of rule-based extraction and (2) to elaborate a method which creates a combined value from the three statistical measures used. Another aim was to find out what threshold value should be given to the combined value in order that precision be significantly increased without decreasing recall too much.

The results of the TE tool were compared to those of other programs carrying out similar tasks (e.g. automatic document indexing). The baseline programs were Fastr (Jacquemin 2001) and YaTeA (Aubin & Hamon 2006).

Corpus

French language patents were chosen as the corpus of the analysis because patents are written in a way to comply with the prerequisites of a specialised text, and terms can only be extracted from specialised corpora. A patent is divided into many units, like bibliographical data, summary, description and claims. From among these parts, our analysis is restricted to the description part of patents because (1) the description part is the most detailed and the longest part of a patent enumerating the advantages of the new invention and (2) as the description has to be as precise as possible, terms are frequently repeated in it as such without any modification. This leads to the presupposition that statistical methods may work well on these texts.

In our analysis, focus is given on patents of two different domains: one is the G06F patent class dealing with Informatics and the other is the A23L class which represents the Human necessities domain[1]. From these two areas, ten descriptions were chosen as samples on which the application was executed. These descriptions have in average 4000 word tokens, that is this number does not include for example punctuation marks. In order to measure the effectiveness of the rule-based extraction, of the rule-based and statistical methods, either separately or in different combinations, terms have manually been annotated, that is they have been marked as terms in these descriptions. Consequently, the term extractor can compare the list (and thus the effectiveness) provided by itself and that of the manually annotated text. In the G06F corpus the manual annotation marked 1752 terms, and in the A23L corpus this number was 2086.

Research method

In the application, TE is carried out by a hybrid method: the extraction phase is rule-based, the filtering phase is rule-based and statistical. Since rule-based term extraction requires the part-of-speech tags of all tokens in the text to be processed and statistical filtering requires the lemma of all tokens, a POS-tagging software (*Connexor*[2]) is applied first on the whole text, so the modules can rely on these data.

Before extracting candidate terms, proper names were filtered out from patent descriptions, which is required because they occur frequently in patent texts (e.g. when citing previous patents); in addition, these names are annotated as nouns by the POS-tagger

[1] In the present analysis, Patentscope was used, which is a patent search tool provided by WIPO (http://www.-wipo.int/pctdb/en/)

[2] http://www.connexor.eu/technology/machinese/demo/

application, therefore if they had not been filtered, numerous non-terminological units would have remained in the list. In the application, *OpenCalais Web Service API*[3] was used for this purpose, because it can be easily implemented in Java. This software made all person, organisation and location names invisible for the rule-based extraction phase.

Connectives were also filtered out in the text during the rule-based filtering process. According to Riegel et al. (2009), the role of these elements is to provide the cohesion in a text, like *en effet* 'in fact' or *par exemple* 'for example'. These have to be filtered out because these expressions containing at least one noun cannot be part of or cannot be a term. In this way, *exemple* 'example' and *effet* 'effect' words from the previous examples are excluded from the candidate term list.

This stopword list also comprises adjectives that has the same function, i.e. providing text cohesion. Such an adjective is for example *suivant* 'next' or *précédent* 'preceding'. Without filtering, the sequence noun-adjective-noun *acides gras suivants* 'following fatty acids' would be incorrectly added to the candidate term list, but if the last adjective is removed, the reduced candidate term (*acides gras* 'fatty acids') is appropriate, therefore both precision (minus one wrong candidate) and recall increase (one new candidate term is recognised).

The rule-based extraction module uses a finite-state automaton to recognise nominal terms. This automaton was created on the basis of previous studies (e.g. Nagy 2008, Nagy 2009a, Nagy 2009b), where almost all syntactic patterns related to terms were collected. For example, a typical term pattern is two nouns following each other (e.g. *accès internet* 'Internet access') or a noun followed by an adjective (e.g. *réseau filaire* 'wired network'). The used patterns contain syntactical compositions which do not only characterise terms, like simple nouns from which some are terms (e.g. *terminal* 'terminal'), some are not (e.g. *problème* 'problem'). Patterns that characterise mostly non-terminological units and the inclusion of which in the pattern list would deteriorate results were excluded. Such a pattern was structures containing internal determiners, like *mise à jour du (de+le) site web* 'update of the web site', in the case of which both *mise à jour* 'update' and *site web* 'web site' became separate candidate terms because of the determiner (*le*) before the latter. In the corpus, there were some terms containing an internal determiner (e.g. *état des (de+les) lieux* 'inventory of fixtures' having the structure noun-preposition-determiner-noun), but the number of these is not significant, therefore these patterns were excluded.

Using finite-state automaton instead of regular expressions is confirmed by the findings of our previous research (e.g. Nagy 2008): the former is simply more transparent. In the automaton, transitions between states were only labelled by part-of-speech tags, and it was made sure that the automaton accept only possible term patterns and that it recognise all intended patterns.

Rule-based filtering was followed by statistical measures, the aims of which were to increase even more precision and F-value without decreasing too much recall. From termhood values, it is the *weirdness* value of Ahmad et al. (1999) that was chosen: it calculates for all candidate terms their proportion of usage in specialised and general corpus. If a given expression occurs in a specialised domain more often than in a general corpus, it is more probable that it is a term. The reference corpus of general language was provided by documents available on Internet, more precisely in an on-line French newspaper, therefore termhood value was based on the results of an Internet search engine: queries were carried out automatically by the application that stored the number of documents in which a candidate term appeared.

Weirdness value is calculated in the following way:

[3] http://www.opencalais.com/calaisAPI

$$weirdness(w) = \frac{\frac{f_s(w)}{t_s(w)}}{\frac{f_g(w)}{t_g(w)}}$$

Where w is the examined element, $f_s(w)$ is the number of occurrences of w in specialised corpus, $f_g(w)$ is the number of occurrences of w in general corpus, $t_s(w)$ is the number of all tokens in the specialised corpus and $t_g(w)$ is the number of all tokens in the general corpus. In the case of words occurring at the same frequency in both the general and the specialised corpus, this value is nearly 1, while terms, which characterise more specialised texts, have a weirdness value more than 1.

From unithood values, it is the C-value of Frantzi & Ananiadou (1997) and Maynard & Ananiadou (2000) that was chosen: the aim of this metric is to determine whether a candidate term occurs more frequently as such or as a part of a bigger unit. C-value is calculated by the following formula:

$$C\text{-}value = \log_2|a| \cdot f(a) \qquad \textit{if a is not embedded}$$
$$C\text{-}value = \log_2|a| \cdot f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \qquad \textit{elsewhere}$$

Where a is the given candidate term, $f(a)$ is its number of occurrences, |a| is the length of the candidate term in words, T_a is the set of candidate terms that contain a, $P(T_a)$ is the number of terms that contain a and are longer than it.

In order to calculate this value, (1) the length of each candidate term, (2) the number of elements that contain it and are longer than it, as well as (3) the number of occurrences of the latter were needed.

From the values taking into consideration text environment, it is the weight value of Frantzi & Ananiadou (1997) that was chosen: for that purpose 100 terms were given as training. The application collected the tokens surrounding these 100 terms and examined their frequency with and without terms. The weight value of each environment token was calculated on the basis of the 100 terms in the following way:

$$Weight(w) = 0{,}5 \cdot \left(\frac{t(w)}{n} + \frac{ft(w)}{f(w)} \right)$$

In the formula, w is the context token, n is the number of training terms (100 in this case), $t(w)$ is the number of cases where a term occurs at least one time with the word w. The element $ft(w)$ demonstrates how many times the w word occurs together with terms, $f(w)$ is the number of occurrences of w in the corpus.

On the basis of weight, C- and weirdness values, a combined value is created, for which it is true that if it is bigger than a threshold value, it designates probably a term. For that purpose, (1) the three value had to be mapped into a common interval in order that they be compatible. Then (2) the weight of each value during the term extraction process had to be decided (e.g. *weirdness* value counts at 60%, weight and C-value at 20% in the combined value). The common interval into which the three values were mapped was the real interval between 0 and 1, because that is the one that is used with probability functions: 0 is for impossible events, 1 is for sure events.

Results

On the basis of Cabré et al. (2001), it was hypothesised that rule-based methods result in big recall with lower precision. This was confirmed by the results on patent descriptions: with only rules-based patterns, a recall of 0.8 and a precision of 0.53 could be achieved on the IT corpus (F-value: 0.64). The results are similar on the other corpus: 0.74; 0.54 and 0.62, respectively. With rule-based filtering, recall could only be slightly increased whereas precision values augmented more (with nearly 0.09).

Before applying statistical methods, an experiment was made on the first five IT texts. The three statistical measures were tried out in almost all possible combinations, and it was measured what the combined values were in the five documents for each term in each combination, and it was determined in which combinations the F-values were maximal in all documents. For that purpose, one-word and multi-word candidate terms had to be treated differently because C-value can only be applied to the latter whereas only weirdness and weight values can be used for the former.

One-word terms are determined more by weirdness value: maximal F-value can be achieved if the weight of the latter is set to 0.8 or 0.9. However, In the case of multi-word terms, it is weight value that is more dominant: the higher the weight this value is set to, the higher F-value becomes. Surprisingly, C-value did not contribute to the increase of F-value in most cases.

The second hypothesis, namely that statistical methods increase significantly precision but decrease recall, was also confirmed by the data: if the threshold of the combined value (CV) was set to high (i.e. if only candidate terms with high CV value were accepted), precision increased considerably but recall (and therefore F-value as well) significantly decreased. In the case of the IT corpus, if the threshold was set to 0.86, precision were 0.9 in average and F-value 0.04. Best F-values were obtained if threshold was set to low (0.03 and 0.05): they increased only a little, approximately with 1.5% on both corpora after rule-based filtering.

The results of the TE application made for this dissertation (own TE) were compared with other, freely accessible computer programs. The first one was Fastr (Jacquemin 2001), the other was YaTeA (Yet Another Term ExtrActor) by Aubin & Hamon (2006).

Table 1. *Comparison of the results of Fastr, YaTeA and my own term extractor on the patent description corpus*

Application	A23L			G06F		
	Recall	Precision	F-value	Recall	Precision	F-value
YaTeA	0,5711	0,3451	0,4270	0,5826	0,3045	0,3983
FastR	0,5764	0,4130	0,4806	0,5349	0,3962	0,4523
Own TE	**0,7614**	**0,6176**	**0,6809**	**0,8280**	**0,6367**	**0,7185**

Table 1 clearly shows that the results of the own application exceeded those of the two other applications with respect to both precision and recall. Although all three programs rely on rule-based methods, results can be different because Fastr and YaTeA do not use the same morphosyntactic analyser as the third but *TreeTagger*[4]. Another reason for the difference in efficiency may be that these term extractors were not created to extract terms specifically from French patent texts.

Sources of error, analysis of results

Enumerating recall, precision and F-value results cannot be sufficient: it is also worth examining at each stage of the extraction process what terms did not figure in the candidate

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

term list and what non-terminological units remained in the candidate term list in spite of the filtering. The main sources of error were the following: (1) POS-tagger errors, (2) in spite of filtering, many non-terminological nominal elements remained in the candidate term list.

In order to demonstrate the different error sources, one description was chosen from each of the two corpora (IT and Human Necessities), where all errors were analysed and categorised. From the G06F corpus, patent number FR2008051823[5] was chosen, from the A23L corpus document FR2007051158[6], and these were named *example corpus*. The first document contains 2760 word tokens and 193 manually annotated terms. The second document has 4372 word tokens and 462 manually annotated terms.

Since the term extractor relies on rule-based methods (i.e. on given morphosyntactic patterns), it is important that the automatic POS-tagger function as efficiently as possible. However, all annotator programs work with a more or less considerable error rate. In the example corpus, a frequent case was annotating nouns as adjectives, for example *solvant* 'solvent' was annotated as the adjective 'solvent'. In the IT example corpus, 5 cases of this kind were found within the 19 non-recognised terms, and 6 non-terminological units got into the candidate term list as noise from the 76 unintended elements. The ratio was different on the A23L example corpus: 31 terms were not recognised because they were incorrectly POS-tagged (total number of unrecognised terms is 99), and 17 non-terminological units got into the list (from among the 164 non-terminological cases).

Another frequent source of error was that the extracted candidate term was not really a term because there are patterns that characterise non-terminological units as well: for example simple nouns. In the IT example corpus, the candidate term list contains 39 candidates that are non-terminological units, in the other text, this number is 52. Such units are *place* 'place', *an* 'year', etc.: the proportion of these units is nearly 30% in the false positive cases.

Statistical methods could only increase F-value to a little extent. In the G06F example corpus, if the combined value was set to low (0.05), precision could be augmented by 4% (from 0.5315 to 0.5738) as compared to the rule-based phase while maintaining recall value. This means that removing elements with low combined value resulted in the increase of precision, that is the removed elements were really non-terminological units. The number of these non-terminological units is 15, for example *instant* 'instant', *lieu* and *place* 'place', *arrêt* 'stop'.

Further research aims

The TE application could be extended to patent descriptions written in other domains and eventually containing other types of patterns, therefore the list of patterns may be extended as well. It is also worth examining whether statistical filtering of terms could also be carried out with other metrics the list of which can be found in Section 5. The replacement of C-value with another unithood value may result in a slightly better result, therefore multi-word terms containing a postnominal adjective that is not part of the term may also be filtered out as well

---

5 Canu, S., Grilheres, B., Brunessaux, S. (2009) *Méthode et système d'annotation de documents multimédia* (Method and System for Annotating Multimedia Documents)
http://www.wipo.int/patentscope/search/en/detail.jsf?docId=WO2009053613&recNum=1&maxRec=4&office=&prevFilter=&sortOption=Pub+Date+Desc&queryString=FP%3A%28Method+and+system+for+annotating+multimedia+documents%29&tab=PCTDescription

6 Bourges, C. (2009) *Utilisation du safran et/ou du safranal et/ou de la crocine et/ou de la picrocrocine et/ou de leurs dérivés en tant qu'agent de satiété pour le traitement de la surcharge pondérale* (Use of Saffron and/or Safranal and/or Crocin and/or Picrocrocin and/or Derivatives thereof as a Sateity Agent for Treatment of Obesity)
http://www.wipo.int/patentscope/search/fr/detail.jsf?docId=EP15054358&recNum=1&office=&queryString=safran+safranal+crocine+picrocrocine+agent+de+sati%C3%A9t%C3%A9&prevFilter=&sortOption=Date+de+pub.+antichronologique&maxRec=2

as other problematic pattern parts. It would also be worth creating a corpus from the 20 patent descriptions where terms do not figure in a separate list but are annotated in the text. This corpus could be used for training purposes, which is not available yet for French for which there are only terminological databases where terms are simply enumerated. It should also be examined whether the same term extractor achieves the same results on other types of corpora: for example, a less specialised (that is more didactic) text contain more non-terminological units, therefore statistical measures should be more relied on there.

References

Ahmad, K., Gillam, L., Tostevin, L. (1999). University of Surrey Participation in Trec8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (Wilder). In *The Eighth Text REtrieval Conference (TREC-8)*.

Ahmad, K. (2001). The role of specialist terminology in artificial intelligence and knowledge acquisition. In Wright, S-E., Budin, G. (eds.) *Handbook of terminology management 2*, 809–844.

Aubin, S., Hamon, Th. (2006). Improving term extraction with terminological resources. *Advances in Natural Language Processing Lecture Notes in Computer Science*, 380–387.

Bosredon B., Tamba I. (1991). *Verre à pied, moule à gaufres* : préposition et noms composés de sous-classe, *Langue française* **91**: 40–55.

Bourigault, D. (1994). *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*, Ph.D. dissertation, Informatique Appliquées aux Sciences Humaines de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Cabré, M. T. (1999). *Terminology. Theory, methods and applications*, Amsterdam/ Philadelphia, John Benjamins.

Cabré, M. T., Bagot, R.E., Vivaldi Palatresi, J. (2001). Automatic term detection. A review of current systems. In Bourrigault, D., Jacquemin, Ch., L'Homme M-C. (eds.) *Recent advantages in Computational Terminology*. Amsterdam/Philadelphia, John Benjamins Publishing Co., 53–87.

Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In Cinque, G. (ed.) *Path Towards Universal Grammar. Studies in Honor of Richard Kayne,* Washington: Georgetown University Press, 85–110.

Connexor, http://www.connexor.eu/technology/machinese/

Daille, B. (1994). *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. Ph.D. dissertation. Université de Paris VII, Paris.

Enguehard, Ch. (2005). Un banc de test pour la reconnaissance de termes en corpus. In Williams, G. (ed.) *La linguistique de corpus*. Rennes, Presses Universiataires de Rennes, 273–286.

Frantzi, K.T., Ananiadou, S. (1997). Automatic term recognition using contextual clues. In *European Research Consortium for Informatics and Mathematics - Cross-Language Information Retrieval,* ERCIM-97-W003, 25–32.

Ha, L.A., Fernandez, G., Mitkov, R., Corpas, G., (2008). Mutual bilingual terminology extraction. In *Proceedings of LREC 2008* (CD-ROM), Marrakech, 1818–1824.

Jacquemin, Ch. (2001). *Spotting and discovering terms through natural language processing*, Cambridge, Cambridge(MA)/London, MIT Press.

Jacquemin, Ch., Bourrigault, D. (2003). Term extraction and automatic indexing. In Mitkov, R. (ed.) *The Oxford handbook of computational linguistics*, Oxford, Oxford university Press, 599–615.

Justeson, J. S., Katz, S. M., (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1): 9–27.

L'Homme, M-C. (2004). *La terminologie: principes et techniques*, Montréal, Les Presses de l'Université de Montréal.

Maynard, D., Ananiadou, S. (2000). Identifying Terms by their Family and Friends. In *Proceedings of COLING 2000*, Luxembourg, 530–536.

Nagy, Á. (2008). L'extraction terminologique : l'un des défis actuels de la linguistique informatique. In Kovács, K., Nagy, Á. (eds.) *Le Passé dans le Présent, le Présent dans le Passé.* Séminaire doctoral, Szeged, 2007. október 25–26., Szeged, JATEPress. 283–290.

Nagy, Á. (2009a). Kísérlet szintaktikai és statisztikai módszerekkel történő automatikus terminológiakivonatolásra francia nyelvű szövegekből [Attempt for automatic term extraction from French texts with syntactic and statistical methods]. In Sinkovics, B. (ed.) *LingDok 8. Nyelvész-doktoranduszok dolgozatai*. JATEPress, 2009. 71–86.

Nagy, Á. (2009b). La structure interne des termes techniques du français et leur reconnaissance par ordinateur. In Kieliszczyk, A., Pilecka, E. (eds.), *La perspective interdisciplinaire des études françaises et francophones*. Łask, Oficyna Wydawnicza LEKSEM, 117–123.

Nagy, Á. (2012a). Contrasting French nominal terms to common language NPs – towards a rule-based term extractor. In Surányi, B., Varga, D. (eds.) *Proceedings of the First Central-European Conference for Postgraduate Students*, Piliscsaba, Pázmány Péter Katolikus Egyetem, 191–210.

Nagy, Á. (2012b). La comparaison de la structure interne des groupes nominaux ordinaires et des termes nominaux: les groupes prépositionnels et les groupes adjectivaux. In Gécseg, Zs., Penke, O., Szász, G. (eds.) *Acta Romanica* (XXVIII). Szeged, JatePress, 7–18.

Petit, G. (2001). L'introuvable identité du terme technique, *Revue Française de Linguistique Appliquée* **VI**(2): 63–79.

Riegel, M., Pellat, J-Ch., Rioul, R. (2009). *Grammaire méthodique du français* (4th edition). Paris, PUF.

Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam & Philadelphia, John Benjamins.

Sauron, V. A. (2002). Tearing out the terms: evaluating terms extractors. In *Proceedings of Translating and the Computer 24*, London, Britain.

Vasconcellos, M. (2001). Terminology and Machine translation, In Wright, S-E., Budin, G. (eds.) *Handbook of terminology management 2*, 697–723.

Wong, W., Liu, W., Bennamoun, M., (2008). Determination of unithood and termhood for term extraction. In Song, M. and Wu, Y. (eds.) *Handbook of Research on Text and Web Mining Technologies*, IGI Global.

Wüster, E. (1976). La théorie générale de la terminologie, un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et la science des objets. In *Actes du colloque international de terminologie*, Québec 5–8 octobre 1975, Québec, L'Éditeur officielle du Québec.

Wüster, E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. In Rondeau, G. and Felber, H. (eds.) *Textes choisis de terminologie. Vol. I: Fondements théoriques de la terminologie*, Québec, Université Laval – GIRSTERM, 55-114.

Publications related to the PhD dissertation

Nagy, Á. (2012). Contrasting French nominal terms to common language NPs – towards a rule-based term extractor. In Surányi, B., Varga, D. (eds.) *Proceedings of the First Central-European Conference for Postgraduate Students*, Piliscsaba, Pázmány Péter Katolikus Egyetem, 191–210.

Nagy, Á. (2012). La comparaison de la structure interne des groupes nominaux ordinaires et des termes nominaux: les groupes prépositionnels et les groupes adjectivaux [Comparison of the internal structure of conventional noun phrases and nominal terms: prepositional and adjectival phrases]. In Gécseg, Zs., Penke, O., Szász, G. (eds.) *Acta Romanica* (XXVIII). Szeged, JatePress, 7–18.

Nagy, Á. (2011). Terminológiakivonatolás francia nyelvű szabadalmak leírásaiból [Term extraction from French patent descriptions]. In Váradi, T. (ed.) *V. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. Budapest, MTA Nyelvtudományi Intézet, 107–114.

Nagy, Á. (2011). Vers une définition opératoire du *terme technique* [Towards a working definition of technical terms]. In Oszetzky, É., Krisztián, B. (eds.) *Cahiers francophones d'Europe Centre-Orientale 14. (Mots, discours, textes – Approches diverses de l'interculturalité francophone en Europe Centre-Orientale*, Pécs, Pécsi Tudományegyetem, 213–220.

Nagy, Á. (2010). L'extraction terminologique automatique: quel domaine de terminologie [Automatic term extraction: which terminology domain]? In Kovács, K., Nagy, Á (eds.) *Acta Romanica. Tomus XXVII. Studia Iuvenum*, Szeged, JATEPress, 7–15.

Nagy, Á. (2010). Terminológiakivonatolás francia nyelvű szabadalmak leírásaiból különböző módszerek segítségével [Term extraction from French patent descriptions with different methods]. In Tanács, A., Vincze, V. (eds.) *MSZNY 2010*, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport, 375–378.

Nagy, Á. (2010). Reconnaissance de groupes nominaux: entre linguistique et intelligence artificielle [Recognition of noun phrases: on the boundary of linguistics and artificial intelligence]. In Malinovská, Z. (eds.) *Parenté/s*. Prešov, Filozofická fakulta Prešovskej univerzity v Prešove, 215–221.

Nagy, Á. (2009). Kísérlet szintaktikai és statisztikai módszerekkel történő automatikus terminológiakivonatolásra francia nyelvű szövegekből [Attempt for automatic term extraction from French texts with syntactic and statistical methods]. In Sinkovics, B. (ed.) *LingDok 8. Nyelvész-doktoranduszok dolgozatai*. JATEPress, 71–86.

Nagy, Á. (2009). La structure interne des termes techniques du français et leur reconnaissance par ordinateur [Inner structure of French terms and their automatic recognition]. In Kieliszczyk, A., Pilecka, E. (eds.) *La perspective interdisciplinaire des études françaises et francophones*, Łask, Oficyna Wydawnicza LEKSEM, 117–123.

Nagy, Á. (2009). Les premiers pas vers la création d'un extracteur automatique de groupes nominaux à partir de textes hongrois étiquetés [The first steps of creating an automatic noun phrase extractor on an annotated Hungarian corpus], In Gécseg, Zs., Penke, O., Szász, G., Kovács, K., Nagy, Á. (eds.) *Acta Romanica, Tomus XXVI*. Szeged, JATEPress, 41–48.

Nagy, Á. (2008). L'extraction terminologique: l'un des défis actuels de la linguistique informatique [Term extraction: an actual challenge in computation linguistics]. In Kovács, K., Nagy, Á. (eds.) *Le Passé dans le Présent, le Présent dans le Passé. Séminaire doctoral*, Szeged, JATEPress, 283–290.