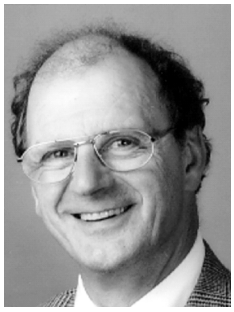
Benchmarking the Performance of Two Part-of-Speech (POS) Taggers for Terminological Purposes

Denis L'Homme*, Marie-Claude L'Homme**, Chantal Lemay***

Département de linguistique et de traduction, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal (Québec), H3C 3J7

*d.lhomme@sympatico.ca, **lhommem@ling.umontreal.ca, ***chantal-lemay@sympatico.ca



Denis L'Homme has a Master's degree in Economics and Astrophysics. He is currently a member of the research group Eclectik (Équipe de recherche en combinatoire lexicale, terminologie et informatique) at the University of Montreal, and has developed several programs for the processing of specialized corpora.

Marie-Claude L'Homme has completed a Ph.D. in Linguistics. She is currently an associate professor at the University of Montreal. She is responsible for the research group Eclectik (Équipe de recherche en combinatoire lexicale, terminologie et informatique). Her research interests are terminology and formal representations of terms.

Chantal Lemay is currently completing a Master's degree in translation at the University of Montreal. Her research interests are computational terminology and corpus linguistics. She is a member of the research group Eclectik (Équipe de recherche en combinatoire lexicale, terminologie et informatique) and is currently working on the evaluation of statistical methods for extracting terms.

D. L'Homme, M.-C. L'Homme, Ch. Lemay. (2002). **Benchmarking the performance of two part-of-speech (POS) taggers for terminological purposes.** *Knowledge Organization, 29(3/4).* 204-216. 19 refs.

ABSTRACT: Part-of-speech (POS) taggers are used in an increasing number of terminology applications. However, terminologists do not know exactly how they perform on specialized texts since most POS taggers have been trained on "general" corpora, that is, corpora containing all sorts of undifferentiated texts. In this article, we evaluate the performance of two POS taggers on French and English medical texts. The taggers are TnT (a statistical tagger developed at Saarland University (Brants 2000)) and WinBrill (the Windows version of the tagger initially developed by Eric Brill (1992)). Ten extracts from medical texts were submitted to the taggers and the outputs scanned manually. Results pertain to the accuracy of tagging in terms of correctly and incorrectly tagged words. We also study the handling of unknown words from different viewpoints.

KEYWORDS: POS tagger, terminology, medicine, specialized corpora

1. Introduction

Part-of-speech (POS) taggers – also called morphosyntactic taggers – are used in an increasing number of terminology tasks (Ahmad & Rogers, 2001; Pearson, 1998). Terminologists collect terms and contexts in which these terms appear in order to produce entries in specialized dictionaries or records in term banks. Most of this information is found in specialized texts in electronic form (texts on computing, law, tele-

communications, medicine, engineering, finance, etc.). The remainder of the information is obtained by consulting specialists in the field in question; however, most of the work is still carried out on corpora. Hence, taggers have recently become a part of terminologists' workstations, along with concordancers, term extractors, and so forth. They represent a very useful means to refine queries when browsing through texts and to reduce problems related to categorial ambiguity.

Up to now, taggers have been used for terminological purposes without questioning their actual performance on specialized corpora, that is, corpora composed of texts related to the same topic which typically contain a very precise vocabulary. Hence, terminologists still have an imprecise idea of how well taggers perform in their specific setting.

The selection of a tagger for terminological work can be quite challenging for a number of reasons listed and explained in section 2. This is due to the way taggers are developed, but also to the nature of terminological methodologies. In order to provide some grounding for the selection of a tagger for terminological purposes, we evaluated the performance of two part-of-speech taggers on specialized corpora. The first tagger is TnT, a statistical tagger developed at Saarland University (Brants, 2000). The second one is WinBrill, the Windows version of a POS tagger originally developed by Eric Brill (1992; 1993; 1995) and extended to French by the French Institute INaLF (INaLF 2002). We ran the taggers on medical texts (TnT on English texts only, WinBrill on French and English texts) and carefully analyzed their outputs. Our claim is that even though off-the-shelf taggers have been trained with corpora of a general nature, they are reliable enough to be used for specialized texts.

This paper is divided into 5 parts. Section 2 is devoted to general considerations on POS taggers, their use in terminology, and the aim of this evaluation. Section 3 describes the methodology used for testing the taggers. Section 4 gives quantitative as well as qualitative results obtained after the analysis of the outputs. A few concluding remarks are provided in Section 5.

2. Taggers and terminology

Taggers are well-known and widely used in disciplines involving the use of corpora, even in applied disciplines such as terminology. Taggers assign linguistic information to character strings in texts for disambiguation purposes, but the nature of the information itself can vary[1]. We will focus on part-of-speech (POS) taggers.

This section will present some methods currently used to develop POS taggers. We will also give some details on the specific characteristics of the two products dealt with in the paper, that is, TnT and WinBrill. Then, we will give some general considerations on the evaluation of taggers and why these evaluations are not always significant for terminology.

2.1 General considerations on POS taggers

Part-of-speech (POS) tagging is the assignment of a tag to each word of a text, the tag indicating the category to which the word belongs (adjective, noun, verb, etc.). In addition, tags can specify inflectional information, which is not extensive for English but can be quite elaborate for other languages, such as French or Spanish.

Whether done manually or automatically, POS tagging relies on two sources of information:

1 – lexical: the list of all possible tags for each word; and
2 – contextual: information about the category of surrounding words, which help determine the correct tag in the given context.

The advent of large manually tagged corpora and of powerful workstations in the last 10–15 years has contributed to the development and use of corpus/computer-based tagging. Corpus-based tagging relies on the fact that complex linguistic phenomena can be identified by observing the order and frequency of words in a text. Corpus/computer-based tagging is a two-stage process:

1 – Learning: during which a pre-tagged corpus is submitted to a learning algorithm that, among other things, generates a lexicon, a table containing all the words found in the corpus along with the associated tags. Other outputs are produced by the learning process and vary with tagger type, as we will see later in this section; and
2 – Tagging: during which an untagged corpus is submitted to the tagging algorithm, which makes use of the information produced by the learning process. Tagging also applies specific techniques in order to handle words not found in the lexicon.

Since the purpose of this article is not to review every type of POS tagger, we will only outline the main characteristics and differences of the two generic taggers used here, TnT and Brill[2]. The first is a statistical tagger, while the second is rule-based.[3]

a) Training of TnT

As mentioned above, the first stage of the tagging process consists of training the tagger. In the case of TnT, the training process generates a lexicon showing the number of times each tag appears associated with a given word. It also generates a table in which the tags are grouped as uni-, bi- and trigrams. Tables 1 and 2 illustrate the format of the lexicon and the table generated by the TnT training process.

old	21	JJ	21
Olsen	1	NP1	1
Olson	1	NP1	1
Olympic	1	JJ	1
Omaha	2	NP1	2
On	27	II	27

Table 1: Extract from lexicon generated from the Susanne training corpus[4]

NP1	4655		
	NP1	911	
		NP1	179
		AT	1
		JJ	4
		NN1	9
		TO	2
		VV0	1
		II	29
		IO	44
		NNJ1	4
		NN2	1
		VHZ	7
		YC	233
		CC	54
		YG	9
		YF	63

Table 2: Extract from n-gram file generated from the Susanne training corpus

b) Training of Brill

The training process of the Brill tagger does not produce any figures. The lexicon that is generated contains a list of all the words in the training corpus plus the associated tags, in which only the most likely tag, that is, the one that occurs most often, is placed in the first position. Table 3 shows an extract of a Brill tagger lexicon.

Laurance	NNP		
mg	NN	JJ	
expressing	VBG		
citybred	JJ		
Brestowe	NNP		
STARS	NNP	NNS	
negative	JJ	NN	
investors	NNS	NNPS	
mountain	NN		
mavens	NNS		
performing-arts	NNS		
car-care	JJ		
Athabascan	NNP		
founding	NN	VBG	JJ
oversold	VBN	JJ	VB

Table 3: Extract from the lexicon generated from training the Brill tagger on the Brown corpus

The Brill training process also generates transformation rules, that is, a set of rules for rearranging the tags of the words for which more that one possible tag exists; applying these rules results in a more accurate output. An extract from a contextual rules file is shown in Table 4. The first line of the sample file, for example, reads as follows: "change the tag IN to DT if the previous tag is IN."

```
IN DT PREVTAG IN
VBP VB PREV1OR2OR3TAG MD
IN RB WDAND2AFT as as
VBD VBN PREV1OR2TAG VB
RB JJ NEXTTAG NN
VBP VB PREV1OR2OR3TAG TO
POS VBZ PREVTAG PRP
NN VBP PREVTAG PRP
DT PDT NEXTTAG DT
IN WDT NEXTTAG VBD
JJ NN SURROUNDTAG DT IN
VBD VBN PREV1OR2TAG VBP
NNS VBZ PREVTAG PRP
IN WDT NEXTTAG VBZ
```

Table 4: Extract from contextual rule file generated from training the Brill tagger on the Brown corpus

c) Tagging algorithm in TnT

The tagging algorithm of TnT consists essentially of assigning the tag chain that maximizes the function $P(W|T) \times (P(T_i|T_{i-i}.....T_{i-n})$ where:

P(W|T) is the probability or relative frequency of a word given a tag, and
(P(T_i|T_{i-i}.....T_{i-n}) is the probability of tag i given tags (i-1 to i-n).

The default value of n in the TnT tagger is 3. The values P(W|T) and (P(T_i|T_{i-i}.....T_{i-n}) are obtained from the lexicon and table generated in the training phase. For words not found in the lexicon, the tagger uses suffixes of different lengths to generate words that might be in the lexicon and hence treated as known words.

veloped by Brill (1992): it includes an English and a French version. Table 5 presents the details of the different versions of the taggers we refer to in this article.

2.2 Performance of taggers

The efficiency of the algorithms implemented in part-of-speech taggers has attracted interest in the computational linguistics community. However, as in the case with most natural language processing applications, their evaluation poses a number of problems.

Tagger/version	Corpus used for training	Number of words	References	Versions evaluated
TnT-WSJ	Wall Street Journal (WSJ): press corpus	approx. 1,200,000 words	Brill (1993) Brants (2000)	X
TnT-Susanne	Susanne Corpus (SC)	approx. 150,000 words	Brants (2000)	X
Brill-WSJ	Wall Street Journal (WSJ): press corpus	approx. 1,200,000 words (the tagger was trained on 50,000 sentences)	Brill (1993)	
Brill-Brown	Brown corpus (BC): mostly contemporary literary texts	approx. 1,000,000 words)	Francis and Kucera (1979) and UCREL (2003)	
WinBrill-English	A combination of WSJ and BC			X
WinBrill-French	Frantext (FT): texts written during the 19th and 20th centuries, mostly literary texts but also a small number of scientific texts	approx. 400,000 words (the entire corpus has 180 million words)	Lecomte (1998)	X

Table 5: Corpora used for training the taggers

d) Tagging algorithm in Brill

In the case of the Brill tagger, tagging is done by first assigning the most likely tag found in the lexicon. Unknown words are first tagged as nouns and then prefixes, suffixes and infixes are used to guess the most likely tags. Transformation rules are then applied to improve the accuracy of the output.

Both TnT and Brill allow the addition of a file containing words that are not contained in the training corpus. This is designed to improve the accuracy of the unknown word tagging process (we will come back to this later in the article).

e) Corpora used for training the taggers

Existing versions of TnT and Brill and versions reported in the literature have been trained on different corpora, using different tagsets. Also, WinBrill contains the rules and the lexicons built from specific training processes used for the tagger originally de-

Interpreting figures related to the accuracy of part-of-speech tagging is not an easy task for a number of reasons.

First, several taggers have been trained on existing corpora, but, as was pointed out in Table 5, these corpora are not the same. Thus, the lexicons generated during the training stages and then integrated into the taggers differ from one product to the other.

Secondly – and this is another consequence of training the taggers on different corpora – the taggers use different tagsets. These vary in terms of number of tags; for example, the English tagset implemented in WinBrill comprises 45 tags compared to the TnT version trained on the Susanne2 corpus that comprises 62 tags. They also vary in terms of the nature of the linguistic information that is indicated: some taggers simply assign parts of speech such as "noun," "verb," "adjective"; others include information on inflection; others make fine-grained distinctions within otherwise general categories (types of verbs, of nouns, of determiners, etc.).

These first two characteristics make it almost impossible to compare POS taggers (Adda et al., 2000; GRACE 2002; Habert et al., 1997). The nature of the corpus, the number of tags and their granularity have direct consequences on the performance of the tagger, regardless of the strategy used (rule-based or probabilistic methods). Hence, an evaluation carried out on taggers in their current form would not only be an evaluation of their ability to disambiguate words, but also an evaluation of the coverage of their lexicon or their tagsets.

Up to now, evaluators have used two different strategies. First, a tagger is evaluated in isolation. This approach was taken by Brill (1993), for example. In the experiment based on the Wall Street Journal corpus (55,787 sentences representing 1,340,777 words), the corpus was divided into two sub-corpora: a training sub-corpus containing the first 50,000 sentences and a testing sub-corpus containing the remaining sentences. The training sub-corpus is itself subdivided in three parts:

1. an annotated lexical training corpus containing the first 1,000 sentences
2. an annotated contextual training corpus containing the second 1,000 sentences
3. an unannotated training corpus containing the remaining sentences.

The testing sub-corpus is, of course, annotated and is presumed error-free. This type of evaluation can be carried out on large samples automatically and fairly rapidly. However, it is only relevant for a specific tagger, and for the corpus it was trained on.

Secondly, taggers are compared to each other. This approach is the one taken by the Action GRACE (GRACE, 2002), a French project evaluating linguistic resources for corpus analysis. In order to achieve this comparison between taggers, a common set of tags was defined in order to convert those used in various products into a standardized set. This allows for a fair comparison of taggers since the evaluation thus focuses on their ability to disambiguate parts of speech. The drawback of this approach is the time it takes to plan it. Nevertheless, results are now available (GRACE, 2002).

2.3 Benchmarking the performance of POS taggers for terminology

Even if precise figures on the performance of POS taggers were available, the question of their relevance for terminological purposes would still not be clearly answered. This section is an attempt to explain the reasons for this state of affairs.

Terminologists always deal with specialized corpora. Normally, when they embark on a new project, they collect specialized texts on the subject under examination assuming that these will contain the terms they must describe and information on these terms. Hence, corpora assembled by terminologists are usually "project-specific"; they are much smaller than corpora used by lexicographers; on the other hand, they are very topical. Each new project calls for a new corpus.

Very seldom can terminologists rely on available tagged specialized corpora. As an illustration of this, the European Language Resources Association (ELRA, 2002) lists two corpora one could consider as "specialized," and only one subset is tagged. The University Center for Computer Corpus Research on Language (UCREL, 2003) lists three corpora. Table 6 presents the specialized corpora listed by ELRA and by UCREL.

Name of the corpus	Fields covered	Tagged
ECI - ELSNET Italian & German tagged subcorpus (ELRA)	Economy 17,000 words Politics 14,000 words Culture 18,000 words Sports 9,000 words Local events 8,500 words	yes yes yes yes yes
"Scientific" corpus of modern French (*La Recherche* magazine) (ELRA)	All articles published in 1998; various scientific fields	no
ETIO-63 Corpus French and English (UCREL)	Telecommunications 250,000 words both languages	yes
International Telecommunications Union (ITU) Corpus – CRATER French, English, Spanish (UCREL)	Telecommunications 1,000,000 words	yes
IBM Manuals Treebank English (UCREL)	Computing 800,000 words	parsed

Table 6: "Specialized" corpora listed by ELRA and by UCREL

Even if they are "specialized," these corpora cannot be directly used for any terminological project. The ECI-ELSNET corpus is simply a press corpus divided into topics. As for the *La Recherche* corpus, several disciplines are covered, but, even if the articles deal-

ing with a specific subject could be isolated, terminologists would still need to complement this first set of texts with others.

The corpora listed by UCREL could be useful for research on telecommunications or computing, but even in those fields the texts included in the corpora might not be varied enough in nature to constitute a balanced corpus.

In addition, terminologists cannot easily resort to large annotated corpora used by lexicographers. These contain sets of undifferentiated texts, such as novels, newspapers, and so forth. Of course, they are likely to contain specialized texts, but it is not always possible to isolate them and make sure they are relevant for a given terminological project.

This means that if terminologists want to use tagged texts for a given project, they must annotate them themselves. Technically, they could customize a POS tagger on the corpora at hand, since most taggers are trainable. This strategy would be unrealistic and very time-consuming in a terminological setting. Terminologists would need to tag the corpus manually since, as we said above, very few annotated specialized corpora are available. Moreover, they would need to repeat the process each time they embarked on a new project.

A more realistic approach is to use a resource that is available, that is, an existing tagger that was trained on a corpus of a different nature, and apply it to the specialized corpus. However, the question of the accuracy of tagging is raised, since specialized texts are likely to contain words that are not present in the lexicon built during the training stages of the existing tool.

This is precisely the approach we took. We submitted specialized texts – namely extracts of medical texts – to two different taggers (TnT and WinBrill) and evaluated the assignment of tags. We scanned the outputs manually in order to calculate the ratio of correctly versus incorrectly tagged words.

It is important to point out right away that this evaluation is not comparative, that is, we are not trying to find out which tagger performs best on medical texts. The results are given and interpreted for each tagger (and each version of the same tagger) individually. Even if we make comparisons between figures here and there, they should be interpreted as indicative. For reasons we listed in subsection 2.2 (different corpora used for training, different tagsets), TnT and Brill cannot be compared to each other.

Since the taggers were run on specialized corpora, we made the assumption that many tagging errors were due to unknown words, that is, words that are not listed in the lexicons of each tagger. We will also investigate this question and measure the coverage of a given tagger and how well it handles unknown words.

3. Methodology

The following subsections give the details of the methodology used for benchmarking the taggers. Details on the texts submitted to the products, on the tagsets of each tagger and each version of a given tagger, and on the decisions made when manually scanning the outputs are given.

3.1 The corpora

The samples submitted to TnT and WinBrill are extracted from French and English texts on medicine. The French texts deal with pharmacology, heart disease and pediatrics, while the English texts bear on neurology[5].

Ten extracts of 350 words each were selected for each language representing 3,500 words per language (this does not include punctuation marks). These extracts might seem short, especially when compared to the size of samples used for evaluations carried out automatically. However, we carried out a pre-evaluation on five texts and the results were similar to those given in section 4.

Le diagnostic différentiel se pose essentiellement avec des lésions traumatiques, en particulier le syndrome des enfants battus.
- Rubéole congénitale : les signes cliniques d'appel sont un retard de croissance intra-utérin, un purpura thrombopénique, des lésions oculaires, une cardiopathie, une anémie ou une hépato-splénomégalie. La microcéphalie est fréquente. Il existe un retard de maturation osseuse.
Les anomalies squelettiques sont le plus souvent latentes. Les signes radiologiques sont caractéristiques mais transitoires et disparaissent habituellement entre 1 et 3 mois. Les lésions osseuses touchent exclusivement les métaphyses, surtout fémorale inférieure et tibiale supérieure :
- bandes claires métaphysaires.
- irrégularité du bord métaphysaire.
- striations longitudinales radio-transparentes caractéristiques dites en « tige de céleri » parallèles au grand axe de l'os.
Il peut également exister un élargissement de la grande fontanelle.
- Autres : CMV, herpès, toxoplasmose.

Figure 1 : French extract

The most common blood vessel disease that causes stenosis is atherosclerosis. In atherosclerosis, deposits of plaque build up along the inner walls of large and medium-sized arteries, causing thickening, hardening, and loss of elasticity of artery walls and decreased blood flow. The role of cholesterol and blood lipids with respect to stroke risk is discussed in the section on cholesterol under "Who is at Risk for Stroke?".

A person with an arteriovenous malformation (AVM) also has an increased risk of hemorrhagic stroke. AVMs are a tangle of defective blood vessels and capillaries within the brain that have thin walls and can therefore rupture. Bleeding from ruptured brain arteries can either go into the substance of the brain or into the various spaces surrounding the brain. Intracerebral hemorrhage occurs when a vessel within the brain leaks blood into the brain itself. Subarachnoid hemorrhage is bleeding under the meninges, or outer membranes, of the brain into the thin fluid-filled space that surrounds the brain.

The subarachnoid space separates the arachnoid membrane from the underlying pia mater membrane. It contains a clear fluid (cerebrospinal fluid or CSF) as well as the small blood vessels that supply the outer surface of the brain. In a subarachnoid hemorrhage, one of the small arteries within the subarachnoid space bursts, flooding the area with blood and contaminating the cerebrospinal fluid. Since the CSF flows throughout the cranium, within the spaces of the brain, subarachnoid hemorrhage can lead to extensive damage throughout the brain

Figure 2 : Englisch extract

The extracts were selected from texts with a high density of terms specific to medicine. Figures 1 and 2 show examples of the extracts selected for each language.

Texts were then submitted to a spell-checker in order to eliminate spelling errors. They were also preprocessed as follows:

- files converted to ASCII format;
- texts arranged one sentence per line (for WinBrill) or one word per line (for TnT);
- words separated by spaces;
- punctuation separated from the previous word;
- for French texts, apostrophes stay with the previous word and are separated from the next word, and for English texts apostrophes stay with the 's' and are separated from the previous word;
- no sentences begin with a space;
- no blank lines;
- capitals at the beginning of sentences are changed to lower case.

These are the guidelines given in Brill documentation; all of these guidelines were followed, except for multi-word term preprocessing. Each word of multi-word terms (e.g., *pia mater membrane, subarachnoid space*) was tagged individually, but compounds (e.g., *self-consciousness*, *fluid-filled*) were taken into account and tagged as one word.

Once the preprocessing was done, a manual check was carried out in order to eliminate blanks in specific cases, for example, spaces inserted in abbreviations (between letters and periods), in numbered lists ending with a period, and so forth.

3.2. Submission to the POS taggers

The English extracts were submitted successively to WinBrill–English and TnT. WinBrill-English was trained on the Brown corpus (BC) and the Wall Street Journal (WSJ) with a set of 45 tags including punctuation. Two different versions of TnT were used for the evaluation, the version trained on WSJ and the one trained on the Susanne corpus (SC). The WSJ corpus has a tagset consisting of 45 different tags (the same as that used by WinBrill–English). Two sets of tags have been developed for the SC. We used the reduced set (Susanne2), which is comprised of 62 tags (this figure does not include punctuation marks).

The French extracts were submitted to WinBrill-French. This version was trained on the Frantext corpus (FT) and has a set of 47 tags. It is worth pointing out that the French version of WinBrill also includes a lemmatizer. The tags generated by the lemmatizer indicate, in addition to POS, the lemma and richer inflectional information. We did not activate this option when submitting the texts, so the results should be interpreted accordingly. Tables 7 to 10 are examples of outputs for each tagger and each version[6].

In WinBrill, unknown words were identified after an examination of the lexicons, which can be read by users. TnT does not authorize access to the lexicons but has a function that identifies unknown words automatically.

The extracts were submitted once again to certain versions of the taggers once unknown words were added to the lexicons. This allowed us to calculate the accuracy of tagging once all words were recognized by the tagger and compare it to the accuracy observed before the addition of the new words.

Word	Tag	Unknown word	Tagging Error	Explanation of the tag
in	IN			preposition
atherosclerosis	NN	*		noun singular
deposits	NNS			noun plural
of	IN			preposition
plaque	NN			noun singular
build	VB		X	verb, base form
up	RP			particle
along	IN			preposition
the	DT			determiner
inner	JJ			adjective
walls	NNS			noun plural
of	IN			preposition
large	JJ			adjective
and	CC			coordinating conjunction
medium-sized	JJ			adjective
arteries	NNS			noun plural
causing	VBG			verb gerund
thickening	JJ	*	X	adjective
hardening	NN	*		noun singular
and	CC			coordinating conjunction

Table 7: Output from TnT-WSJ

Word	Tag	Unknown word	Tagging Error	Explanation of the tag
in	II			general preposition
atherosclerosis	NN	*		common noun
deposits	NN			common noun
of	IO			of
plaque	JJ	*	X	general adjective
build	VV		X	verb, base form, present participle catenative, past participle
up	RP			particle
along	II			general preposition
the	AT			article
inner	JB			attributive adjective
walls	NN			common noun
of	IO			of
large	JJ			general adjective
and	CC			general coordinating conjunction
medium-sized	JJ	*		general adjective
arteries	NN			common noun
thickening	NN	*		common noun
hardening	VV	*	X	verb, base form, present participle catenative, past participle
and	CC			general coordinating conjunction

Table 8: Output from TnT-Susanne2

Word	Tag	Unknown word	Tagging Error	Explanation of the tag
in	IN			preposition
atherosclerosis	NN			noun singular
deposits	VBZ		X	verb, 3rd person singular present
of	IN			preposition
plaque	NN			noun singular
build	NN		X	noun singular
up	IN		X	preposition
along	IN			preposition
the	DT			determiner
inner	JJ			adjective
walls	NNS			noun plural
of	IN			preposition
large	JJ			adjective
and	CC			coordinating conjunction
and	CC			general coordinating conjunction
medium-sized	JJ	*		general adjective
arteries	NN			common noun
thickening	NN	*		common noun
hardening	VV	*	X	verb, base form, present participle catenative, past participle
and	CC			general coordinating conjunction

Table 9: Output from WinBrill English

Word	Tag	Unknown word	Tagging Error	Explanation of the tag
rubéole	SBC:sg			nom commun singulier
congénitale	ADJ:sg			adjectif singulier
les	DTN:pl			déterminant pluriel
signes	SBC:pl			nom commun pluriel
cliniques	SBC:pl			nom commun pluriel
d'	SBC:sg		X	préposition
appel	SBC:sg			nom commun singulier
sont	ECJ:pl			"être" (vb. ou aux.) conjugué plur.
un	DTN:sg			déterminant singulier
retard	SBC:sg			nom commun singulier
de	PREP			préposition
croissance	SBC:sg			nom commun singulier
intra-utérin	ADJ:sg			adjectif singulier
un	DTN:sg			déterminant singulier
purpura	SBC:sg			nom commun singulier
thrombopénique	ADJ:sg	*		adjectif singulier
des	DTC:pl			déterminant pluriel
lésions	SBC:pl			nom commun pluriel
oculaires	ADJ:pl			adjectif pluriel
une	DTN:sg			déterminant singulier

Table 10: Output from WinBrill-French

Finally, a number of decisions were made before calculating the number of errors. Every individual occurrence of an error was counted, even when it occurred several times in a text. Also, each unknown word was also counted as an error every time it occurred. For example, the word *hydrocephalus* was found three times in a text. In one case it was correctly tagged and in two cases it was incorrectly tagged. The word was counted as three unknown words and was twice considered as an incorrectly tagged word.

4. Results

This section presents the results observed after a detailed analysis of the outputs generated by each tagger (and within each tagger, each version). Subsection 4.1 gives the proportion of correctly versus incorrectly tagged words, regardless of the type of error. A few comments on frequent errors are also given. Subsection 4.2 studies the handling of unknown words from different viewpoints.

4.1 General results

The outputs generated by each tagger were checked manually and words classified as correctly tagged or incorrectly tagged. Table 11 presents the results obtained after the analysis of each output. It should be kept in mind that the total number of words for each language is 3,500. The total number of incorrectly tagged words is given, as well as the percentage of correctly tagged words. The percentage represents the tagging accuracy for each tagger.

Tagger and version	Number of incorrectly tagged words	% of correctly tagged words
TnT – Susanne	189	94.65%
TnT – WSJ	213	93.9%
WinBrill –English	191	94.55%
WinBrill –French	170	95.15%

Table 11: Correctly tagged words in the medical extracts

These figures show that both POS taggers perform well on medical texts. Even the lowest figure, that is, 93.9%, shows that they are reliable when applied on specialized corpora. Minor variations between figures can be observed and are due either to the coverage of the lexicon, the tagset, or the disambiguation strategy implemented in the tagger.

We can compare these figures in part to others given by the developers of the taggers evaluated. Brill (1995) ran his tagger on the Wall Street Journal (WSJ) corpus and obtained accuracies between 96.1% and 97.2% (depending on the size of the training corpus and the rules used during the tagging process). Brants (2000) carried out an evaluation of TnT (during various phases of the training process) and obtained accuracies between 96% and 97%.[7]

If we compare the highest score obtained by Brill on the WSJ corpus (97.2%) to the score obtained when the tagger was applied on English medical texts, the difference is 2.65%. Similarly, the differences between the highest score obtained by TnT (97%) on the corpora used for its training and the ones it obtained on the medical corpus range from 3.2% (TnT – WSJ) to 2.35% (TnT – Susanne2). Once again, this comparison is only indicative, since the results of these evaluations are obtained on versions of the taggers that are not fully comparable.

Hence, our preliminary conclusion is that POS taggers perform well, even when applied on specialized corpora. We can also point out that it is fair to assume that medical texts are among the most difficult to deal with, since they are likely to contain very specific vocabulary that would be unlikely to appear in general corpora used for training.

Frequent tagging errors that were encountered when analyzing the outputs are listed below.

We cannot make generalizations about the tagging errors found in outputs generated by WinBrill-French. The tagger has been trained on a single corpus (i.e., FT). However, it seems that several tagging errors affect nouns and adjectives. We give a few examples below:[8]

a) Nouns tagged as adjectives, pronouns
résulte/VCJ :sg *de*/PREP *plusieurs*/DTN:pl *effets*/SBC:pl
des/DTC:pl ***cadiotoniques***/**ADJ:pl**
les/DET:pl ***digitaliques***/**PRO:pl** *stimulent*/VCJ :pl *ainsi*/ADV
le/DTN:sg *système*/SBC:sg

b) Adjectives tagged as cardinals, nouns, past participles
système/SBC:sg *d'*/PREP *échange*/SBC***:sg*** ***transmembranaire***/**CAR**
radiologiques/ADJ:pl *sont*/ECJ:pl ***caractéristiques***/**SBC:pl**
mais/COO *transitoires*/ADJ :pl
Une/DTN :sg *action*/SBC :sg ***inotrope***/**ADJ2PAR** :sg *positive*/ADJ :sg

The most frequent tagging errors found in WinBrill-English affect nouns and verbs. The rules applied to unknown words by the tagger can explain why nouns with an *–ly* ending are incorrectly tagged. During the training process, the program applies a series of examinations of recurrent suffixes. Since, *-ly* is normally an ending for adverbs, the tagger assigns this tag to the word. Moreover, contextual rules might explain why *foot* was tagged as a verb: it follows *to* in the context. Some examples are given below.

a) Nouns tagged as adjectives, adverbs, verbs
is/VBZ *called*/VBN *an*/DT ***embolus***/**JJ** *and*/CC *often*/RB
what/WP *is*/VBZ ***holoprosencephaly***/**RB**
due/JJ *to*/TO ***foot***/**VB** *inversion*/NN *or*/CC *walking*/VBG

b) Verbs tagged as nouns
the/DT *disorder*/NN ***results***/NNS *in*/IN *an*/DT *abnormal*/JJ *skull*/NN

The most frequent tagging errors found in TnT, regardless of the training corpus, also affect nouns and verbs. Examples are given below.

a) Nouns tagged as adjectives, adverbs
*is*VBZ *called*VBN *an*DT ***embolus***JJ *and*CC *often*RB *forms*NNS (WSJ)
*since*IC *many*DA *pregnancies*NN *with*IW *a*AT ***fetus***JJ *diagnosed*VV *with*IW (Susanne)
*diagnosed*VBN *with*IN *HPE*NNP *may*MD *have*VB *a*DT *small*JJ *head*NN ***microcephaly***RB (WSJ)
b) Verbs tagged as nouns
*is*VB *variable*JJ *and*CC *may*VM ***range***NN *from*II *mild*JJ *elevations*NN *of*IO (SC)
c) Wh- determiner tagged as preposition
*other*JJ *drugs*NNS ***that***IN *help*NN *regulate*VB *digestion*NN *or*CC *reduce*VB (WSJ)

Some frequent errors can be observed in the outputs of both taggers for English (WinBrill-English and TnT). For example, *embolus* and *microcephaly* were incorrectly tagged by both products. As with WinBrill, words ending with *–ly* were incorrectly tagged as adverbs by TnT. However, other frequent errors were found with only one tagger. For instance, Wh-determiners were frequently tagged as prepositions by TnT but not by WinBrill-English.

4.2 Handling of unknown words

The handling of unknown words can be examined from different viewpoints. First, the lexicons used by different versions of the taggers lack a certain number of words encountered in the texts that are submitted to them. We can assume that the proportion of unknown words will be high for medical texts.

Table 12 gives an idea of the coverage of the lexicons by presenting the total of words unknown to each tagger. Column 1 contains the total number of unknown words (the total occurrences); column 2 presents the proportion of the total number of words (3,500) submitted to the taggers that were unknown).

These figures show that WinBrill-French (in fact, the FT corpus) has the widest coverage, since few words were not listed in the lexicon referred to during the tagging process. On the other hand, TnT–SC lacks several words found in the texts submitted to this version of the product.

	Total number of unknown words	Proportion of unknown words of the total number of words
TnT–Susanne	658	18.8%
TnT-WSJ	395	11.29%
WinBrill–English	328	9.37%
WinBrill–French	220	6.29%

Table 12: Coverage of lexicons

However, both WinBrill and TnT use specific techniques to deal with the tagging of unknown words. We can give an idea of how well unknown words are disambiguated by calculating the proportion of correctly tagged unknown words as compared to the total number of unknown words. The results for each version of the taggers are given in Table 13.

The percentage appearing in column 3 is a calculation of the number of incorrectly tagged unknown words on the total number of unknown words.

	Total number of unknown words	Unknown words correctly tagged	Unknown words incorrectly tagged	
TnT – Susanne	658	568	90	13.68%
TnT – WSJ	395	318	77	19.49%
WinBrill–English	328	241	73	22.56%
WinBrill–French	220	114	106	48.18%

Table 13: Handling of unknown words

Figures in the right-hand column show that the handling of unknown words is extremely variable from one tagger to the other. While all taggers assign correct tags to most unknown words, some appear to perform better in this area. For example, TnT-SC assigned 568 correct tags to a total of 658 unknown words, whereas WinBrill-French assigned 114 correct tags to a total of 220 unknown words.

These figures also show that although not all unknown words are correctly tagged, the taggers (except for WinBrill French) appear to perform relatively well. In addition – and this is rather surprising – there does not appear to be a direct relation between the coverage of the lexicon and the correct or incorrect tagging of unknown words. However, a closer examination of the nature of unknown words themselves would be needed to clarify this matter.

Finally, we recalculated the proportion of correctly versus incorrectly tagged words once unknown words had been added to the lexicons of each tagger.

	Accuracy of tagging before unknown words were added to the lexicon	Accuracy of tagging after unknown words were added to the lexicon
TnT–Susanne	94.7%	97.3%
TnT–WSJ	93.9%	97.2%
WinBrill–English	94.55%	96.3%
WinBrill–French	95.15%	97.95%

Table 14: Accuracy before and after adding unknown words

First, we can see that the addition of unknown words to the lexicons of taggers improves tagging accuracy slightly (from 1.75% to 3.9%).

These new figures become particularly interesting when compared to the figures given by the developers of the taggers and cited earlier. Recall that Brill (1995) reported accuracies between 96.1% and 97.2% (compared to our 96.3% accuracy). Brants (2000) reported accuracies between 96% and 97% for WSJ compared to the 97.2% we obtained.

Hence, the accuracy obtained when adding unknown words found in new corpora can be compared to accuracies obtained by submitting corpora used during the training of the tagger. The accuracy observed in medical texts once the unknown words have been added to the lexicon is very close to the figures given by developers.

5. Conclusions

Our quantitative and qualitative evaluation of part-of-speech (POS) taggers shows that different products perform well when applied on new corpora, and even on a specialized corpus composed of extracts from medical texts. Results obtained after applying the taggers without editing their lexicons show accuracies ranging between 93.3% and 95.15%. This proves that they are reliable tools for a terminological setting.

The accuracy of tagging can be improved even further by adding unknown words either to the lexicons or in a backup lexicon as allowed by the tagger. The addition of unknown words to the lexicons improved the assignment of correct tags by approximately 2%. This accuracy can be compared to the accuracies reported by the developers when running the taggers on "general corpora."

Acknowledgements

We would like to thank Elizabeth Marshman for her careful revision of the English text and the reviewers for their useful comments on a preliminary version of this paper.

Notes

1 In fact, some taggers assign syntactic and even semantic information to character strings. More information can be found in Habert et al. (1997) and on the UCREL website (UCREL 2003).
2 The description provided in this section is based on the original tagger developed by Brill. It is still valid for WinBrill, which is the tagger that we evaluated.
3 A detailed description of the TnT tagger can be found in Brants (2000) and that of Brill in Brill (1993; 1995).
4 The details of the different corpora referred to in the article are provided further in this section.
5 These texts are part of corpora comprising mainly specialized texts used at the Observatoire de linguistique Sens-Texte (OLST) at the University of Montreal.
6 We removed the punctuation marks since they were not considered in the evaluation, but it is worth pointing out that they are also produced in the output accompanied by a tag (e.g.. ,/,; ./.).
7 It should be reminded that we did not count punctuation marks. Figures given by Brill or Brants might include them.
8 We standardized the presentations of the outputs in the examples listed in this subsection. However, the tags are those assigned by each tagger.

References

Adda, G., Lecomte, J., Mariani, J., Paroubek, P., & Rajman, M. (2000). Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de parties du discours pour le français. In K. Chibout, J. Mariani, N. Masson, & F. Néel (Eds.). *Ressources et évaluation en ingénierie des langues* (pp.645-664). Paris : Duculot.

Ahmad, K., & Rogers, M. (2001). Corpus Linguistics and Terminology Extraction. In S.E. Wright, & G. Budin (Eds.), *Handbook of Terminology Management*, Vol. 2 (pp.725-760). Amsterdam / Philadelphia: John Benjamins.

Brants, T. (2000). TnT – A Statistical Part-of-speech Tagger. *Proceedings of the Sixth Conference on Applied Natural Language Processing, ANLP-2000*, Seattle, Washington. Retrieved 15 May 2002, from http://www.coli.uni-sb.de/~thorsten/publications/Brants-ANLP00.pdf

Brill, E. (1992). A Simple Rule-based Part of Speech Tagger. *Proceedings of the Third Conference on Applied -Natural Language Processing ANLP-1992*, Trento, Italy. Retrieved 12 November 2001, from http://www.cs.jhu.edu/~brill.

Brill, E. (1993). *A Corpus-based Approach to Language Learning. PhD Thesis.* Pennsylvania: Department of Computer Science, University of Pennsylvania.

Brill, E. (1995). Transformation-based error-driven Learning and Natural Language Processing. *Computational Linguistics*, 21(4), 543-565.

Brill, E. (1996). Transformation-based Error-driven Parsing. In H. Bunt, & M. Tornita (Eds.), *Recent Advances in Parsing Technology.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

De Loupy, C. (1995). « La méthode d'étiquetage d'Eric Brill », *Traitement automatique des langues (TAL),* 36(1-2), 37-46.

ELRA (2002). *European Language Resources Association*. Retrieved 15 March 2002, from http://www.icp.grenet.fr/ELRA/home.html

Francis, W.N., & Kucera, H. (1979). *Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers*. Providence: Department of Linguistics, Brown University.

Garside, R., Leech, G., & Sampson, G. (1987). *The Computational Analysis of English. A Corpus-based Approach*, London/New York: Longman.

GRACE (2002). *Grammaires et Ressources pour les Analyseurs de Corpus et leur Évaluation*, LIMSI, CNRS. Retrieved 15 May 2002, from http://www.limsi.fr/TLP/grace/

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus.* Paris : A. Colin.

Inalf (2002). Institut national de linguistique française. Retrieved 15 May 2002, from http://www.inalf.fr/

Lecomte, J. (1998). Le catégoriseur Brill14-JL5 / WinBrill-0.3, INaLF/CNRS. Retrieved 13 February 2003, from http://www.inalf.fr/winbrill/BRILL14-JL5_WinBrill.doc

Pearson, J. (1998). *Terms in Context*. Amsterdam/Philadelphia: John Benjamins.

Sparck-Jones, K., & Galliers, J.R. (1996). *Evaluating Natural Language Processing Systems*. Berlin: Springer.

UCREL (2003). *University Centre for Computer Corpus Research on Language*. University of Lancaster, United Kingdom. Retrieved 13 February 2003, from http://www.comp.lancs.ac.uk/computing/research/ucrel/

Van Halteren, H., Zavrel, J. & Daelemens, W. (2001). Improving Accuracy in Word Class Tagging Through the Combination of Machine Learning Systems. *Computational Linguistics* 27(2), 199-229.