Trab. Ling. Aplic., Campinas, (41):147-161, Jan./Jun. 2003

USING WORDSMITH TOOLS AND TAGGED CORPORA AS AN AID TO GRAMMAR LEARNING

LEONARDO JULIANO RECSKI
(Doutorando-UFSC)

RESUMO

Este artigo demonstra alguns métodos para o uso do software lingüístico WordSmith Tools aplicado a corpora sintaticamente anotados ao invés de corpora comum. Várias características e modelos são discutidos que possam ser aplicados na sala de aula como uma forma de auxiliar o ensino de gramática. Sugere-se que estes modelos possam auxiliar os aprendizes a gerar hipóteses e desenvolver habilidades para a resolução de problemas e que, assim, possam ajudar a criar um estilo de ensino centrado na autonomia do aprendizado do aluno. Os exercícios propostos neste artigo são adequados para o ensino de inglês para aprendizes de LE. Além disso, corpora sintaticamente anotados são recursos valiosos para professores prepararem suas lições utilizando-se de exemplos autênticos de uso da linguagem.

INTRODUCTION

Linguistic software used in personal computers, such as the *WordSmith Tools*, designed by Mike Scott (1996), have outstanding features to serve classroom needs. The use of concordancers to work on plain text as a classroom teaching aid has been widely discussed. Tribble and Jones (1990) offered a number of practical exercises on plain text, that is, texts which have no grammatical or syntactic tagging. Johns (1991, 1993, 1998) and Stevens (1991a, 1991b, 1995) also proposed a variety of effective classroom activities, from gap-filling to pattern searching and from vocabulary to grammar learning. Witton (1993), Gavioli (1996, 1998), Dodd (1998), and Davies (1999) discussed how to use concordancers with languages other than English. Louw (1991, 1998) showed how to use concordancers for critical literary appreciation.

This paper discusses some aspects of grammar learning by the application of *WordSmith Tools* on tagged or annotated corpora, in contrast to the results obtained when the software is run on other plain or raw corpora. Leech (1993: 275) defines corpus annotation as "the practice of adding interpretative (especially linguistic) information to an existing corpus of spoken and/or written language by some kind of coding attached, or interspersed with, the electronic representation of the language material itself". In other words, in a tagged corpus every word of the corpus is annotated automatically by a computer program which attributes a 'tag' to show the word class of each word in its context. The corpora used as demonstration data is a corpus of one million words automatically tagged with the computer program TOSCA tagger,

developed in the English Department at the University of Nijmegen, Netherlands. Both the TOSCA tagger as well as tagging schemes are discussed in detail in Section 2 below. In Section 3, five types of exercises are proposed with techniques and examples. The exercises are suitable for advanced students of English as a second or foreign language.

1. THE SOFTWARE

Wordsmith Tools is a relatively small, but undoubtedly useful, piece of software running on a personal computer. The programs in *WordSmith Tools* can handle virtually unlimited amounts of text. They can read text from CD-ROMs, so giving access to corpora containing many millions of words. The main advantage of *Wordsmith Tools* is that it displays the output directly on the screen. The output can also be saved as a file and printed out. *Wordsmith Tools* can be used not only on plain English texts, but also on texts in other languages, and on English texts with grammatical and syntactic encoding. The functions of the *WordSmith Tools* include frequency listing, alphabetical listing, keyword in context (KWIC) analysis, further searching on both sides of the keywords, and closer investigation of the target items in larger contexts.

To run *Wordsmith Tools* it is recommended 8MB of RAM, at least 5MB of hard disk space, an IBM-compatible PC with a 386 or better processor and at least a Windows™ 3.1 or more recent Windows versions.

2. DATA AND TAGGING SCHEME

The tagged data used in the present study is a corpus of one million words of written and spoken English horizontally tagged using the TOSCA Tagger (henceforth Corpus A). The written portion of the tagged corpus is divided in four samples of 200,000 tokens each, comprising four different registers: academic writing, fiction, business, and science. The spoken portion of the tagged corpus amounts to 200,000 tokens distributed within transcriptions of CNN Talk Shows and News Programs, transcriptions of White House Press briefings, and transcriptions of workplace interactions.

In a part-of-speech (or POS) tagged corpus every word is assigned a category tag, which is often complemented with a series of attributes. POS taggers are either rule-based (*ENGCG*), probabilistic (the *Birmingham Tagger, CLAWS*) or mixed, i.e. both rule-based and probabilistic in nature (*TOSCA*). The TOSCA tagger uses 256 tags, which means that a very refined analysis can be carried out. For example, the tagger distinguishes 22 word classes, many of which are subcategorized to give a total of 256 lexico-grammatical tags, 78 of which are for verb types alone and 15 punctuation and pause tags. What this means is that, for example, a concrete analysis such as the automatic retrieval of monotransitive verbs (verbs which only take a direct object) could not be performed on corpora tagged with a minimal tagset. There is thus a clear link between the refinements of the tagset, the precision of the analysis that can be carried

out and the benefits this accuracy may bring to language teaching and learning. An example of a sentence tagged with the TOSCA Tagger is:

The_ART(def) system_N(sing) is_VB(aux,pass,pres) based_VB(lex,montr,edp) on_PREP(ge) three_NUM(card,sing) categories_N(plu) of_PREP(ge) rules_N(plu) stored_VB(lex,montr,edp) in_PREP(ge) the_ART(def) computer_N(sing) 's_GENM memory_N(sing) ._PUNC(per)

The codes have the following interpretation

ART (def)	definite article
N (sing/plu)	singular/plural noun
VB (aux,pass,pres)	passive auxiliary verb in the present
VB (lex,montr,edp)	monotransitive lexical verb in the past participle
PREP (ge)	common preposition
NUM (card,sing)	singular cardinal number
GENM	genitive marker
PUNC (per)	punctuation - period
_	joins words to their grammatical tags

Rewriting the tags in parentheses, we can paraphrase the sentence above as follows:

The (definite article) system (singular noun) is (present of passive auxiliary verb) based (past participle of lexical monotransitive verb) on (common preposition) three (singular cardinal number) categories (plural noun) of (common preposition) rules (plural noun) stored (past participle of lexical monotransitive verb) in (common preposition) the (definite article) computer (singular noun) 's (genitive marker) memory (singular noun) . (period)

As we have seen, POS tagging assigns a tag to each word in a text to label the word class to which it belongs in context.

With any software learning tool it is necessary to ask to what extent the computer-based environment facilitates or inhibits access to learning. In the case of *WordSmith Tools* the keyboard and screen conventions are standard and familiar to users who have a functional competence in software, like word-processing systems, running on a Windows machine. The codes for tagging in the corpora, however, are not so immediately transparent. As far as tagging is concerned, it is established that some training needs to be undertaken to allow students to operate *WordSmith* confidently, and to have a better understanding of the corpora to be processed, together with the tagging scheme. The schemes developed by the grammarians with various considerations are not hard to learn. The tags can be assimilated with reasonable speed and ease. They tend to evoke the grammatical name of the category ("VB" = verb). It is not difficult to grasp that within "VB" (verb category) "edp" means "past participle of a verb", "ingp" means "present participle of a verb", and "infin" means "infinitive of a verb". Students have shown in practice that they can become functionally competent with both *WordSmith* and the tagging codes after a couple of hours of supervised practice.

3. SAMPLE ACTIVITIES FOR CLASSROOM PRACTICE

There are numerous activities for classroom practice when *WordSmith* is applied to tagged corpora. These include the study of new grammatical structures, frequency listing, gap-filling, and some imaginative pattern searching. The optimal role of *WordSmith* as a grammar-learning tool is an enhancement and addition to classroom-based learning.

3.1. The study of new grammatical items/structures

When introducing a new grammatical item/structure, the teacher may let the student look up examples of the item/structure in the corpora, so as to understand the grammatical item/structure better through exposure to examples in authentic texts. Let us suppose that the student is interested in the use of the verb "succeed". A concordance search for "succeed" in a POS tagged corpus yields examples like:

Concordance for "succeed" [partial]
Text: Corpus A – Business section (approx. 200,000 tokens)
societies_N(plu) would_VB(aux,modal,past) succeed_VB(lex,intr,infin) in_PREP(ge) meeting_VB(lex,mon,ingp)

"Succeed" is tagged as an infinitive intransitive lexical verb "lex,intr,infin" and it is followed by the preposition *in*. The students can then analyze all occurrences of the verb "succeed" in the corpus to investigate if this is the most common pattern associated to the use of this verb. Below is a sample of the first ten lines of a concordance search for "succeed":

Concordance for "succeed" [10/32 lines]
Text: Corpus A – Business section (approx. 200,000 tokens)

i,infin):1/2 to_VB(aux,semi,infin):2/2	succeed_VB(lex,intr,infin) in_PREP(ge) depressing_VB(lex,m
s_N(plu) would_VB(aux,modal,past)	succeed_VB(lex,intr,infin) in_PREP(ge) meeting_VB(lex,mon
_VB(aux,modal,past) not_ADV(neg)	succeed_VB(lex,intr,infin) in_PREP(ge) selling_N(sing) off_P
,pres):1/2 to_VB(aux,modal,pres):2/2	succeed_VB(lex,intr,infin) in_PREP(ge) partnering_VB(lex,m
ue_N(sing) will_VB(aux,modal,pres)	succeed_VB(lex,intr,infin) or_or_CONJUNC(coord) fail_N(si
mma) desire_N(sing) to_PRTCL(to)	succeed_VB(lex,intr,infin) ._PUNC(per) <sent3958
es) not_ADV(neg) yet_ADV(ge,pos)	succeeded_VB(lex,montr,edp) in_PREP(ge) getting_VB(lex,m
kets_N(plu) have_VB(aux,perf,pres)	succeeded_VB(lex,montr,edp) ._PUNC(per) <sent8406>
>_MARKUP She_PRON(pers,sing)	succeeded_VB(lex,intr,past) rather_ADV(ge,pos) more_ADV
by_PREP(ge) actually_ADV(ge,pos)	succeeding_VB(lex,intr,ingp) in_PREP(ge) the_ART(def) take

The verb "succeed" occurred 32 times in the Business section of Corpus A. The first ten concordance lines show that "succeed" mostly occurs as an intransitive verb and that it can be used in the infinitive "infin", past participle "edp", past "past", and present participle "ingp" tenses. It is also possible to observe that in five of the ten sentences the verb "succeed" is followed by the preposition "in" + the present participle of lexical monotransitive verbs "lex,montr,ingp". All together, intransitive uses of the verb "succeed" accounted for 28 occurrences while monotransitive uses accounted for only

four occurrences. The pattern *succeed + in + v-ing* occurred 11 times (34%) in the Business section.

The concordance search described above shows how to retrieve examples of specific lexical and grammatical information from a tagged corpus. The students may want to know what structure a certain sequence of parts of speech can form. Investigating whether a sequence of part of speech tags can be found in the corpora will give some idea of how common the pattern is. However, it is important to bear in mind that corpora are a limited source of language data, so we should be careful when drawing conclusions. For example, we can look up words of a sequence of part of speech tags, like "ART(def) ADJ(ge,*) N(sing)" (i.e., "ART(def)" = definite article "the", "ADJ(gen,*)" = superlative, comparative, and positive adjectives, "N(sing)" = singular noun). The operational instructions to *WordSmith* are:

(1) in the *Concord* tool open the "Getting started" menu
(2) click on "Chose the texts now" and select the texts you want to analyze
(3) choose "Specify search word"
(4) fill in the input box with "*_ART(def) *_ADJ(ge,*) *_N(sing)". This means:
* = wildcard (any word or tag)
*_ART(def) = any word that takes an ART(def) tag or is a definite article
_ADJ(ge,) = any word that takes an ADJ(ge,pos), ADJ(ge,sup), or ADJ(ge,comp) tag or is a general positive, superlative, or comparative adjective
*_N(sing) = any word that takes an N(sing) tag or is a singular noun

The result is:

Concordance for "*_ART(def) ADJ(ge,*) *_N(sing) (15/1613 lines)
Text: Corpus A – Academic Writing section (approx. 200,000 tokens)

the_ART(def) international_ADJ(ge,pos) proletariat_N(sing) and_CO
the_ART(def) American-style_ADJ(ge,pos) boss_N(sing) who_PRON(r
the_ART(def) present_ADJ(ge,pos) situation_N(sing) be_VB(lex,cop,
the_ART(def) English_ADJ(ge,pos) system_N(sing) ._PUNC(per
the_ART(def) Trade_ADJ(ge,pos) Unions_N(sing) hope_VB(lex,montr
the_ART(def) large_ADJ(ge,pos) number_N(sing) of_PREP(ge) Petite_
the_ART(def) economic_ADJ(ge,pos) crisis_N(sing) of_PREP(ge) the_
the_ART(def) Soviet_ADJ(ge,pos) Union_N(sing) ._PUNC(per
the_ART(def) Trade_ADJ(ge,pos) Unions_N(sing) redundant_ADJ(ge,
the_ART(def) new_ADJ(ge,pos) change_N(sing) in_PREP(ge) negotiati
the_ART(def) Cold_ADJ(ge,pos) War_N(sing) ,_PUNC(comma)
the_ART(def) great_ADJ(ge,pos) majority_N(sing) of_PREP(ge) firm_
the_ART(def) only_ADJ(ge,pos) answer_N(sing) if_CONJUNC(subord) i
the_ART(def) higher_ADJ(ge,comp) education_N(sing) level_ADV(ge,pos
the_ART(def) greater_ADJ(ge,comp) importance_N(sing) of_PREP(ge) pr

If we add one more "N(sing)" to the pattern to check it as "ART(def) ADJ(ge,*) N(sing) N(sing)" in the corpus, the result shrinks to 69 examples:

Concordance for "ART(def) ADJ(ge,*) N(sing) N(sing)" (7/69 lines)
Text: Corpus A – Academic Writing section (approx. 200,000 tokens)

the_ART(def) after_ADJ(ge,pos) war_N(sing) period_N(sing)
the_ART(def) outside_ADJ(ge,pos) business_N(sing) world_N(sing)
the_ART(def) chief_ADJ(ge,pos) policy_N(sing) maker_N(sing)
the_ART(def) right_ADJ(ge,pos) wing_N(sing) coalition_N(sing)
the_ART(def) prime_ADJ(ge,pos) minister_N(sing) role_N(sing)
the_ART(def) fundamental_ADJ(ge,pos) attribution_N(sing) theory_N(sing)
the_ART(def) individual_ADJ(ge,pos) constituent_N(sing) nation_N(sing)

This shows that the pattern "ART ADJ N" is much more common than "ART ADJ N N" in the corpora. In addition, the teacher may call the attention of the students about one important aspect: both patterns belong to noun phrases, even though some of them may have some other patterns embedded inside the noun phrases (this, of course, requires more context on both sides of the pattern).

3.2. Frequency listing

Unlike a frequency listing of words in plain texts, which present only the frequency of words, the frequency listing of grammatical elements is an exercise to let the students understand the number of occurrences of the wordtags (part of speech of a word) in a text. It answers questions like:

(a) Which are the most frequent parts of speech in the texts?
(b) What is the main difference in the frequency listing of parts of speech between two different categories of texts?

The following are frequency listings of grammatical items of the Science and Spoken sections of Corpus A:

Table 1. Frequency of POS tags across different text types

Science: 208,822 tokens/ 19,175 types		**Spoken: 204,753 tokens/ 10,958 types**	
N	63.328	N	43.469
VB	36.086	VB	42.726
PREP	28.202	PUNC	29.012
PUNC	25.649	PRON	27.197
ART	21.393	PREP	19.251
ADJ	20.902	ADV	15.821
PRON	14.848	ART	15.173
ADV	13.213	ADJ	13.465
CONJUNC	10.127	CONJUNC	10.292
NUM	5.242	PRTCL	3.022
PRTCL	2.909	NUM	2.551
TAG	1.790	EXTHERE	763
GENM	1.096	GENM	692
EXTHERE	368	MISC	532

From these data we answer the first question about the most often occurring parts of speech in the texts among the 14 most frequent grammatical elements.

There are some differences between the lists for the two text types. For instance, in the scientific genre there are far more nouns (N) than in the spoken genre. One plausible reason for such difference may be that written texts are, on average, more lexically elaborated than spoken texts. It is intuitively clear that many written texts, such as published scientific articles, are densely packed with information, whereas much spoken language, such as casual conversation, is not. There is a simple functional interpretation for such finding. On average, a written text is longer and has fewer repetitions than a comparable spoken text. It is permanent, highly edited, redrafted and rehearsed, rather than being unplanned and spontaneous as most casual conversation is.

The pedagogical use of text listings such as the ones presented above is that they may be used by students to find out stylistic characteristics across different text types, as well as to have a handle on how the quantity of different parts of speech may differ across a range of text types.

3.3. Gap filling

Gap filling is a simpler exercise to design. The teacher can extract some tagged sentences from the corpora, save them, and then edit them on a word processor. The teacher may delete the tags and replace them with a number. The students are required to recover the grammatical tag in the position of the number. Below is an example of a gap-filling exercise:

Aim: fill in the numbered blanks with an appropriate part of speech tag:

The_ART(**#1**) group_N(sing,collect) was_VB(aux,pass,past) impressed_VB(lex,**#2**,edp) with_PREP(ge) the_ART(def) proposal_(**#3**) 's_GENM low_ADJ(ge,pos) cost_N(sing) and_**#4**(coord) the_ART(def) technical_ADJ(ge,pos) merit_N(sing) of_PREP(ge) the_ART(def) Russian_**#5**(ge,pos) proposal_N(sing).

Students would not necessarily be expected to reconstitute the gaps in their technical format, but rather in a format closer to natural language, e.g. "singular noun" rather then the part of speech tag "N(sing)". The answer for the exercise above is #1 = def (definite article), #2 = montr (past participle monotransitive lexical verb), #3 = N(sing) (singular noun), #4 = CONJUNC (coordinate conjunction), #5 = ADJ (common positive adjective).

Students can either discuss the task among themselves or work independently. They can look up certain words and/or tags that are related to the target so as to find the best answer by using *WordSmith* on a corpus. For example, the students can look up the search string "impressed_VB(lex,*,edp)" in the corpus to find out what kind of verb "to impress" can be, e.g. *intransitive, monotransitive, ditransitive, dimonotransitive*, etc. For obvious reasons teachers should not allow students access to the corpus from which the test sentences were derived.

3.4. Additional exploratory searches

Two types of searches are discussed in this section: nearness searching and grammatical pattern searching. They can be expanded according to students' specific individual needs, and students can generate search methods to explore the corpora.

3.4.1 Nearness searching

WordSmith allows us to configure a search for a character string to both right and left of the keyword in the concordance. The configuration allows searches for close or far ranges from about two to eight words, and with or without wildcards. With this searching function, we are able to restrict the search to a more specific scale. In English, verbs can take a preposition to modify the original meaning of that verb to a certain extent: for instance the verb "turn" can be followed by "against" to mean "to stop supporting someone and oppose to them", and by "down" to mean "to refuse". Students can be asked to answer the following questions in the search exercise:

(1) What are all of the prepositions that can occur immediately after the verb "turn"?
(2) What do the combinations mean in context?
(3) What are the high frequency verb + preposition combinations?

To find out the pattern "turn + preposition", I present a text demonstration. We can first enter "turn" in a corpus tagged containing only lemmas and part of speech tags. All variants of "turn" will then be retrieved (e.g., *turn, turning, turned*). After that, in the *Getting started menu,* we specify the search string as "turn_VB(*) *_PREP(*)". These steps instruct *WordSmith* to search to the right of the verb "turn" in any form in a range of one word and the pattern of any word joined with "PREP" by an underscore. The result looks like this:

Concordance for "turn_VB(*) *_PREP(*)" (13 lines)
Text: Corpus A: Business (approx. 200,000 tokens)

modal,past) be_VB(aux,pass,infin)	turn_VB(lex,montr,edp) around_PREP(ge) a_ART(indef)
NUM(card,sing) PowerGen_N(sing)	turn_VB(lex,intr,pres) down_PREP(ge) cut_N(sing) price_N(sin
irline_N(plu) have_VB(aux,perf,pres)	turn_VB(lex,cxtr,edp) into_PREP(ge) the_ART(def) country_N(
100_NUM(card,sing) index_N(sing)	turn_VB(lex,cxtr,past) into_PREP(ge) a_ART(indef) 28.
VB(aux,modal,pres) not_ADV(neg)	turn_VB(lex,intr,infin) into_PREP(phras) another_ADJ(ord) Fox
REP(ge) the_ART(def) star_N(sing)	turn_VB(lex,intr,pres) in_PREP(ge) a_ART(indef) unhappy_AD
rike_N(sing) and_CONJUNC(coord)	turn_VB(lex,montr,edp) in_PREP(ge) a_ART(indef) fall_N(sing)
mma) the_ART(def) market_N(sing)	turn_VB(lex,montr,edp) in_PREP(ge) a_ART(indef) resilient_A
n) government_N(plu) to_PRTCL(to)	turn_VB(lex,intr,infin) over_PREP(ge) all_PRON(univ) service_
sion_N(sing) have_VB(aux,perf,past)	turn_VB(lex,intr,edp) to_PREP(ge) the_ART(def) treaty_N(sing)
PRON(ass) may_VB(aux,modal,pres)	turn_VB(lex,intr,infin) to_PREP(ge) spending_N(sing) rather_C
ertiser_N(plu) be_VB(aux,prog,pres)	turn_VB(lex,intr,ingp) to_PREP(ge) medium_N(plu) buy_VB(le
,modal,past) otherwise_ADV(ge,pos)	turn_VB(lex,intr,infin) to_PREP(phras) the_ART(def) internatio

Answering the three questions proposed at the outset of the exercise we have:

(1) The prepositions that can occur with "turn" and its variants in the corpus analyzed are *around, down, into, in, over*, and *to*.
(2) The meanings expressed in the combinations are:
   (a) *around* in "the group could be turned around" means to change to an opposite situation;
   (b) *down* in "PowerGen turns down cut prices" means to refuse;
   (c) *into* in "but it will not turn into another Fox" means to become something different;
   (d) *in* in "the market turned in a resilient performance" means to produce as a result of work;
   (e) *over* in "it will oblige government to turn over all services" means to deliver in the possession or control of somebody else;
   (f) *to* in "the commission had turned to the treaty" means to go for help or advice.
(3) The highest frequency combinations are as follows:

turn to	4
turn in	3
turn into	3
turn around	1
turn down	1
turn over	1

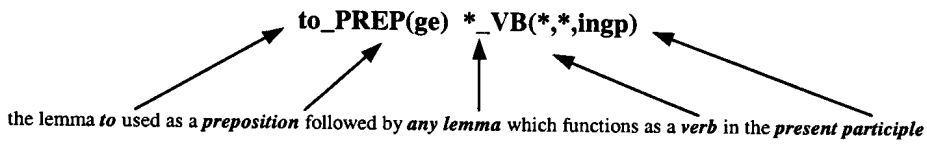
The Business section of Corpus A is too small to provide statistically reliable generalizable results beyond the current search, but *turn to*, *turn in*, and *turn into* did turn out to be above the mode. By running a search on the whole corpus it was observed that the highest frequency combinations continued to be the ones found in the Business section. Overall the pattern "turn + preposition" occurred 130 times in the whole corpus with *turn to* being the most common pattern, appearing 47 times, *turn into* appearing 20 times, and the third most common pattern being *turn towards* appearing 11 times.

In exercises like the one above, students are encouraged to infer the meaning of verb-preposition patterns from the context. However, this does not prevent students from using a dictionary when they are not sure about their guesses or when they fail to guess the meaning. Multiple references and information sources can provide practical input, depending on the task and the student's strategy.

3.4.2. Grammatical pattern searching

Along their learning horizon, students may come across certain ambiguities and grammatical rules which are difficult to grasp. For example, "to" acts most often as an infinitive marker followed by a verb in the infinitive. Nonetheless, sometimes "to" plays the role of a preposition followed by a gerund verb. For instance, we would normally ask a potential employee at a job interview "Do you object to working on Sunday?". In the sentence "to working" is an example of "to" as a preposition as part of the "object to" idiomatic pattern. So what other similar patterns are represented in the corpora? In

order to search for all such patterns of "to+verb-ing", we fill in the input box in the *Concord* tool with the following search string:



The first two asterisks which follow the tag VB, indicate respectively that the verb may be lexical, copular, auxiliary, etc, and monotransitive, ditransitive, complex transitive, etc. The *Concord* delivers the result like this:

Concordance for "to_PREP(ge) *_VB(*,*,ingp)" (9/141 lines)
Text: Corpus A – approx. 1,000,000 tokens

(1) in_ADV(connec):1/2 addition_ADV(connec):2/2	to_PREP(ge) do_VB(lex,montr,**ingp**) what
(2) moral_ADJ(ge,pos) approach_N(sing)	to_PREP(ge) achieve_VB(lex,montr,**ingp**)
(3) be_VB(aux,pass,edp) commit_VB(lex,montr,edp	to_PREP(ge) buy_VB(lex,montr,**ingp**)
(4) neuron_N(plu) devote_VB(lex,montr,edp)	to_PREP(ge) receive_VB(lex,montr,**ingp**) signal
(5) as_PREP(ge) the_ART(def) route_N(sing)	to_PREP(ge) perform_VB(lex,montr,**ingp**) a
(6) be_VB(lex,cop,past) sympathetic_ADJ(ge,pos)	to_PREP(ge) join_VB(lex,montr,**ingp**) the_ART
(7) the_ART(def) only_ADJ(ge,pos) route_N(sing)	to_PREP(ge) save_VB(lex,montr,**ingp**) the_ART
(8) ass,past) not_ADV(neg) suit_VB(lex,montr,edp)	to_PREP(ge) carry_VB(lex,montr,**ingp**) out_PRE
(9) doctor_N(plu) switch_VB(lex,montr,edp)	to_PREP(ge) use_VB(lex,montr,**ingp**) streptokina

The examples above show some of the patterns of "to+verb-ing" which were found in the corpus, e.g. (7) the *route to saving* something; or (9) to *switch to doing* something. The most common patterns found in the whole corpus were (no. of occurrences between parenthesis):

in addition to+ v-ing	(8)	key to + v-ing	(6)
admit to+ v-ing	(7)	lead to + v-ing	(4)
approach to + v-ing	(5)	oppose to + v-ing	(7)
close to +v-ing	(10)	view to +v-ing	(7)
commitment to +v-ing	(6)	way to + v-ing	(4)
commit to + v-ing	(8)		

Through exercises like this, students can explore the corpora and answer many questions by themselves, thus developing a more autonomous approach by accessing the data – the corpora – via *WordSmith* tools. The advantage of searching for a specific kind of "to *ing" in a tagged text is that we can specify "to" as a preposition, a feature which is not available when working with plain text corpus. If one wants to find "to *ing" in a plain text corpus, the result of searching for "*ing" can also include nouns and adjectives.

4. PEDAGOGICAL ASPECTS

Unlike using other computer-assisted language learning (CALL) software, using *WordSmith* on tagged corpora is open-ended, student-centered and heuristic. The students are motivated to create a question or hypothesis first, and then find the answer in the corpora, developing and exploiting strategies for exploring the language under study. This approach can help students think, judge, compare, and learn in a more autonomous way than in traditional teacher-centered, theory-guided grammar learning. By using *WordSmith* tools on tagged corpora students may acquire a more creative, heuristic and dynamic learning and cognitive style.

The overall cognitive framework is one of exploration. The learning activities have a threefold outcome: learning of language through exposure to authentic materials in a rich text environment, using computing tools to access this material in a way which is difficult to implement in other environments; learning of the grammar of the language through direct manipulation of grammatical information in a structured context; and learning of computer-based techniques for working with language, grammar and text. With appropriate guidance, students can be encouraged to develop both skills and modes of exploring language and asking language-direct questions, which would not arise so directly or accessibly in more conventional learning environments.

The teachers play the role of supervisors, advisers and additional knowledge resources while the students are the learners. Teachers are also trainers in the use of the software, in the nature of corpora, and in the relationship between corpora and language learning. It goes without saying that the correct and effective operation to achieve better learning is crucial in making the students feel confident and willing to use a software tool like *WordSmith*. Teachers can also decide the main topics of the grammar to be learned according to the syllabus and the teaching plan. The *real* teacher and knowledge providers are *WordSmith* and the corpora. The students can work individually, in pairs, or in groups. Tagged corpora can be part of the whole corpora used in learning and teaching a language. Students are recommended to become concordance-literate as soon as computers are introduced in the learning process.

Another advantage to be highlighted is that the format and formalism like those of tagged corpora and *WordSmith* tools provide a framework for exploring and hypothesizing about language. By running searches, say, for verb frameworks specified in terms of subject or object types, the student can investigate not only notions of transitivity but also the semantics of such verbs and their associated arguments. Or by searching for certain syntactic patterns like embedded clauses, students can explore the range of verbs which are able to introduce such clauses. Furthermore, there are also more than mere by-products in terms of students' reading and writing expertise, because one direct result of working with corpora and *WordSmith* tools is a heightened awareness of syntagmatic information available in the linear structure of the sentence for the resolution of problems of comprehension.

5. CONCLUSION

By using *WordSmith Tools* on tagged corpora the learner explores the resources of the English grammar in a free and unrestricted way. No pedagogical considerations were originally involved in the design and compilation of the corpus used for this study, and *WordSmith Tools* itself is largely pedagogically neutral as regards methodology. One potential advantage of using this pedagogical framework may be the creation of a more autonomous learning style, which leads students to generate hypothesis and to develop problem-solving abilities in grammar learning. Working on the corpora with *WordSmith Tools* is not a programmed task. With good preparation, including the study of the operations of *WordSmith* and a good understanding of the structures of the tagged corpus, in addition to familiarity with their tagging schemes, students can surely find it useful in learning grammar. By and large, using *WordSmith Tools* on tagged corpora is good practice for the learning and teaching of grammar. Not only can students benefit from and be interested in it, but teachers can also take full advantage of it as a support to their effective and creative teaching.

Acknowledgements: I am very grateful to two anonymous referees for their pertinent comments on an earlier draft of this paper and for suggesting me to rethink the title of the article. I am also thankful to my adviser, Dr. Viviane Heberle, and to Dr. Marco Rocha for helping me to improve this paper.

REFERENCES

AARTS, Jan; BARKEMA, Henk & OOSTJICK, Nelleke. (1997). *The TOSCA-ICLE Tagset – Tagging Manual.* TOSCA Research Group for Corpus Linguistics, University of Nijmegen.

BOURNARD, Lou; MCENERY, Tony. (1999). (eds.). *Rethinking Language Pedagogy from a Corpus Perspective.* Lodz Studies in Language, 2.

DAVIES, Mark. (1999). Using multi-million word corpora of historical and dialectal Spanish texts to teach advanced courses in Spanish linguistics. In: BOURNARD, Lou; MCENERY, Tony (eds.). *Rethinking Language Pedagogy from a Corpus Perspective*: 173-185.

DODD, Bill. (1998). Exploiting a Corpus of Written German for Advanced Language Learning. In: WICHMANN, Anne; FLIGELSTONE, Steven; MCENERY, Tony; KNOWLES, Gerry (eds.). *Teaching and Language Corpora.* Longman, London: 131-56.

GAVIOLI, Laura. (1996). Corpus di testi e concordanze: un nuovo strumento nella didattica delle lingue straniere, *Rassegna Italiana di Linguistica Applicata* 2.

_____. (1998). Exploring Texts through the Concordancer: Guiding the Learner. In: WICHMANN, Anne; FLIGELSTONE, Steven; MCENERY, Tony; KNOWLES, Gerry (eds.). *Teaching and Language Corpora.* Longman, London: 83-99.

JOHNS, Tim. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. Birmingham University, *English Language Research Journal* 4:25-45.

_____. (1993). Data-driven learning: an Update, *TELL&CALL* 2: 4-10.

_____. (1998). Contexts, the Background, Development and Trailing of a Concordance-based CALL Program. In: WICHMANN, Anne; FLIGELSTONE, Steven; MCENERY, Tony; KNOWLES, Gerry (eds.). *Teaching and Language Corpora*. Longman, London: 100-15.

JOHNS, Tim; KING, Phillip. (1991). (eds.). *Classroom Concordancing*. Birmingham University.

LEECH, Geophrey. (1993). Corpus Annotation Schemes, *Literary and Linguistic Computing* 8/4:275-81.

LOUW, Bill. (1991). Classroom Concordancing of Delexical Forms and the Case for Integrating Language and Literature. In: JOHNS, Tim; KING, Phillip (eds.). *Classroom Concordancing*. Birmingham University.

_____. (1998). The Role of Corpora in Critical Literary Appreciation. In: WICHMANN, Anne; FLIGELSTONE, Steven; MCENERY, Tony; KNOWLES, Gerry (eds.). *Teaching and Language Corpora*. Longman, London: 240-51.

SCOTT, Mike. (1996). *WordSmith Tools*. Oxford University Press.

STEVENS, Vance. (1991a). Classroom concordancing: vocabulary materials derived from relevant, authentic text, *English for Specific Purposes* 10:10-15.

_____. (1991). Concordance-based Vocabulary exercises: a Viable Alternative to Gap-filling. 1991(b). In: JOHNS, Tim; KING, Phillip (eds.). *Classroom Concordancing*. Birmingham University: 47-61.

_____. (1995). Concordancing with Language Learners: Why? When? What?. *CAELL Journal* 6/2:2-10.

TRIBBLE, Chris; JONES Glvn. (1990). *Concordances in the Classroom*. Longman, London.

WICHMANN, Anne; FLIGELSTONE, Steven; MCENERY, Tony & KNOWLES, Gerry. (1998). (eds.). *Teaching and Language Corpora*. Longman, London.

WITTON, Nic. (1993). Using the Mini-Concordancer with Languages other than English, *ON-CALL 7/2*:19-20.

APPENDIX
TOSCA-ICLE ALPHABETICAL LIST OF GRAMMATICAL WORDTAGS

Major wordclasses (16)

ADJ	adjective	NADJ	nominal adjective
ADV	adverb	NUM	numeral
ART	article	PREP	preposition
CONJUNC	conjunction	PROFM	proform
EXTHERE	existential *there*	PRON	pronoun
GENM	genitive marker	PRTCL	particle
MISC	miscellaneous	PUNC	punctuation
N	noun	VB	verb

Major features (92)

antit	anticipatory *it*	PRON	ingp	*-ing* participle	ADJ; NADJ; VB
ass	assertive	PRON	inter	interrogative	PRON
aux	auxiliary	VB	interjec	interjection	MISC
card	cardinal	NUM	intr	intransitive	VB
cbrack	closing bracket	PUNC	lex	lexical	VB
clause	clause	PROFM	modal	modal	VB
cleft	cleft *it*	PRON	montr	monotransitive	VB
collect	collective	N	mult	multiplicative	NUM
colon	colon	PUNC	neg	negative	ADV; PRON; VB
comma	comma	PUNC	nomplu	plural nominal	ADJ
comp	comparative	ADJ; ADV; NADJ	nomposs	nominal possessive	PRON
conj	conjoin	PROFM	nonass	non-assertive	PRON
connec	connective	ADV	number	number	N; PRON
coord	coordinating	CONJUNC	obrack	opening bracket	PUNC
cop	copula	VB	one	*one*	PROFM; PRON
cquo	closing quote	PUNC	oquo	opening quote	PUNC
cxtr	complex transitive	VB	ord	ordinal	ADJ NUM
dash	dash	PUNC	other	other	PUNC
def	definitive	ART	pass	passive voice	VB
dem	demonstrative	PRON	past	past tense	VB
dimontr	dimono-transitive	VB	per	period	PUNC
discourse	discourse	MISC	perf	perfective aspect	VB
ditr	ditransitive	VB	pers	personal	PRON
do	*do*	VB	phras	phrasal	ADV; PREP
edp	*-ed* participle	ADJ; NADJ; VB	phrase	phrase	PROFM
ellip	ellipsis	PUNC	plu	plural	N; NUM; PROFM; PRON
ellipt	elliptical	VB	pos	positive	ADJ; ADV; NADJ
encl	enclitic	PRON; VB	poss	possessive	PRON
exclam	exclamatory	PRON	prefix	prefix	MISC
exm	exclamation mark	PUNC	pres	present tense	VB
for	particle *for*	PRTCL	procl	proclitic	PRON; VB
foreign	foreign	MISC	prog	progressive aspect	VB
frac	fractional	NUM	qm	question mark	PUNC
ge	general	ADJ; ADV; PREP	quant	quantifying	PRON
hyph	hyphenated	NUM	recip	reciprocal	PRON
imp	imperative	VB	rel	relative	PRON
indef	indefinite	ART	scolon	semi colon	PUNC
infin	infinitive	VB	self	*-self / -selves*	PRON

semip	semi followed by *-ing* participle	VB	suffix	suffix	MISC
			sup	superlative	ADJ; ADV; NADJ
sing	singular	N; NUM; PROFM	to	*to*	PRTCL
so	*so*	PROFM	univ	universal	PRON
subjun	subjunctive	VB	wh	*wh-*	ADV
subord	subordinating	CONJUNC	with	*with*	PRTCL
such	*such*	PRON			