A Probabilisitic Approach to Linguistic Analysis in Machine Translation Output Evaluation

Olivier Gouirand
Laboratoire BABEL, Faculté de Lettres et Sciences Humaines
Université du Sud Toulon-Var
BP 20132, F-83957 La Garde (France)
olivier.gouirand@univ-tln.fr

Abstract

In order to overcome some shortcomings of previous machine translation (MT) evaluation methods, mainly pertaining to the difficulty of measuring actual linguistic quality (if possible in an automated way), a metric based on the distributional features of grammatical categories within the scope of the *Law on Anomalous Numbers* was devised : the text is tagged for parts-of-speech and these are computed into a positional distribution. The deviation of the latter from said law indicates translation errors, which alone is useful in diagnostic evaluations but is also of importance to theoretical linguists.

Natural language generated by means of computer software indeed disrupts an equilibrium that lies in the syntax-semantics relations, because a text may seem well-formed while being nonsense all the same. The reason is grammatical categories are primitive constructs controlled by semantic categories and structures into which they fit. Studying MT output is about identifying software issues but also raising interesting questions for linguists to address : why is it that parts-of-speech are organized (of course not necessarily in a uniform way across languages) along the lines of other human and natural constructs, as demonstrated herein? What is their ontological status? etc.

1 Introduction

The MT evaluation community has contributed a great number of approaches over the years but there are two problems which hamper progress in the field: first, generic evaluation techniques have always been considered difficult to achieve due to the specificity of user needs (King & Falkedal, 1990:1 ; Spark & Galliers, 1996:70) ; second, measuring the essential feature of MT i.e. text quality is a very complex task (Hovy et al., 2002:5) due to the many possible planes of analysis that it implies. The response to the former has been to design adaptative frameworks like EAGLES (EWG, 1996) or ISLE (described in Hovy et. al., 2002:4), the risk being to minimize the performance of core technologies factor as set forth by DARPA (White, O'Connell & O'Mara, 1994:introduction). As to the latter, the quest for objective metrics - preferably automatable ones - that would grasp text quality has lead to mainly statistical criteria, the risk being to abstract away from linguistic quality. For example, in the BLEU algorithm (Papinemi, 2001:2-3), the distance between any translation and a set of reference translations stems from a calculus involving *n*-gram similarity (modified unigram precision) deemed to mirror translation quality criteria adequacy and fluency. It only rests on word counts though, so the linguistic level cannot be directly invoked.

The present paper seeks to cope with both problems by asserting the priority of linguistic quality in MT output assessment over any other criterion because that is fundamentally what the end user needs. To pursue this aim, a method which would incorporate linguistic content into a generalizable framework is needed, other than statistical equalizations of human judgements *à la* ALPAC (1966).

What follows is basically the result of an investigation into the very nature of traditional grammatical categories in the wake of Aristotle, viz. nouns as "substance", determiners as "quantity", adjectives as "quality", verbs as "action / affection", just to name a few. We aim at demonstrating that the distribution of categories may be generalizable - although categories are not strictly mappable -, at least to a pair of natural languages, namely English and French ; that could be useful to better understand the ties between natural languages out of a linguistic standpoint (which is of paramount importance in translation) and shedding light upon phenomena best interpreted at the higher levels of cognition, or rather, semantics following Morris's classification (1938:6).

We shall also see that they clearly fit into the more complex scheme of the syntax-semantics interface, in the sense of words' distributional and semantic properties tallying. Indeed, one of the hurdles of MT, thus its evaluation, has been the gap between syntax and semantics, the former having historically prevailed over the latter, making full-blown linguistic evaluations troublesome, not to mention the other major stumbling block, i.e. the lack of expected results for any one translation, in other words a "correct" or "yardstick" translation.

2 Setup and experimenting with categorial distributions

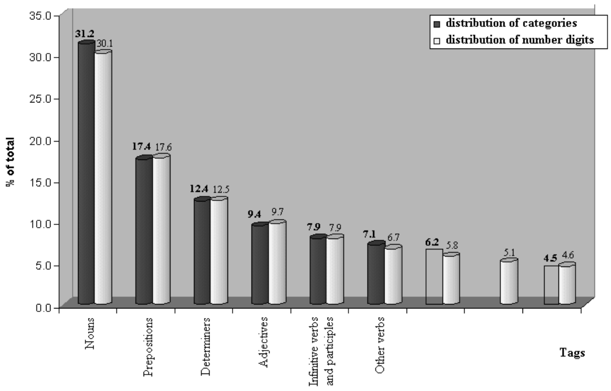
Our experiment started with an in-depth experiment conducted on four English-French parallel manually translated corpora, each about 180,000 words, drawn from various subject areas and contained in the well-documented MCI/ECI corpus (ACL, 1994) [subcorpora "ILO" and "CCITT"], as well the *Hansard* (Parliament of Canada, 1999) and the *Annual Reports of the UN Secretary General* (UN, 1993 ; UN, 1995-2000). Technically, the first stage of it was devoted to cleaning up the texts from unnecessary information (empty lines, doublets and the like) as well marking sentence boundaries, a requirement of the piece of software carrying out the actual linguistic work. A robust tagger previously trained on the *Penn Treebank* and the *Trésor de la Langue Française* corpora (WinBrill© by ATILF, formerly InaLF, 1998) was used to assign grammatical tags to parts-of-speech. After stripping the texts of its lexical information, the tags alone were fed into a standard spreadsheet application, each sentence forming a row. Then the columns were scanned for specific tags and these were totalled (see Figure I). The resulting positional "grammatical n-grams" formed the data, the interpretation of which follows. It should be emphasized that these are linguistic in the sense that every part-of-speech mirrors some syntactic relation : for example, a determiner is only such with respect to the "determined", an adjective (*adjectivum nomen*) relates to a noun, an adverb to a verb and so forth.

The starting point in the interpretation of the results could be the well-known fact that word distributions comply with Zipf's law (Zipf, 1965:38), that is : the rank of a word is inversely proportional to its frequency and their product is constant. But what is less known is that, above a threshold of about 2,000 tokens (although we first noticed this on a manually counted 500-word text, spurring our initial hypothesis), the distribution of parts-of-speech in English and French matches another statistical discovery, the *Law on Anomalous Numbers* (Benford, 1938). This law, which states that the distribution D of number digits equals $log(1+1/D)$, applies to all non-random (thus not to lotteries and the like) and non-constrained numbers (thus not to supermarket prices ending with 9 for marketing reasons etc.), ranging from land areas to stock-market prices, regardless of unit used. It is readily and successfully used in accounting to trace frauds but otherwise limited to a narrow audience despite its mathematical demonstration by Terence Hill in 1996, probably because it encompasses other laws in a disturbing fashion ; the history of science has been an array of laws where some paradigms seem to be both the product of man's analysis of his world and the world's laws, which some scientists as Thom (1990:344) acknowledge in the biological nature of language e.g. noun as *salience* (Thom, 1988:23).

In our experiment, the average distribution of grammatical categories in the English set of texts from the corpora is : nouns 31.23 %, prepositions 17.39 %, determiners 12.44 %, adjectives 9.43 %, infinitive verbs and participles 7.92 %, other verbs (very weakly correlated to the others positionally) 7.15 % and presumably two groups made up of some wh- words and conjunctions as well as one comprising pronouns and adverbs, account for 6.18 and 4.48 % respectively, see Table I below.

Table I

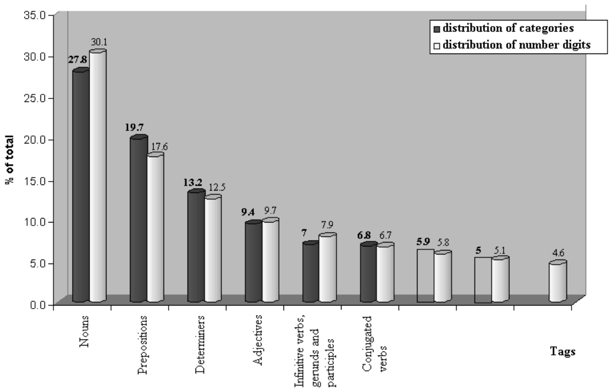
Compared distribution of grammatical categories and number digits (English set)



In the French texts, the distribution is : nouns 27.83 %, prepositions 19.73 %, determiners 13.19 %, adjectives 9.41 %, infinitive verbs and participles 6.96 %, conjugated verbs 6.80 % and presumably two groups made up of pronouns, as well as one comprising conjunctions, account for 5.91 and 4.96 % respectively, see Table II below.

Table II

Compared distribution of grammatical categories and number digits (French set)



The lower part of the spectrum is nonetheless kept truncated because intervals in such a distribution are too close to discriminate between aggregate "minor" categories with certainty, unless finer-grained groupings emerge from future research.

These results were checked for reliability : deviations were within acceptable limits and no data relative to the categories presented above that were below the confidence level of 1,000 due to the large number of tokens for each type. Also, only the most strongly correlated groupings were made. When it comes to informativity, the set underwent a subsequent series of processes consisting in a machine translation of all source texts, then the post-processing as described above and the comparison with the manually translated target texts. Interestingly, with a first-generation MT system (Systran), the distribution was distorted significantly : 30 % less proper nouns in English and 22 % in French, more than a fifth less present tense verbs and participles in English, a quarter and a sixth more determiners and prepositions respectively in English, roughly the same loss for French pronouns and other verbs (see Table III and IV). To take just one example for the sake of brevity, Systran has a persistent bug : proper nouns are treated as common nouns when there's a match, as in "Mr Pat O'Brien" translated as *M. tapotent O'Brien* ; that explains the aforesaid deflated proper nouns.

Table III

Compared distribution of grammatical categories (English)

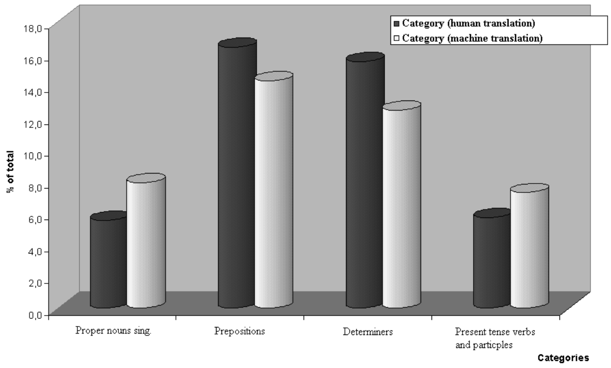
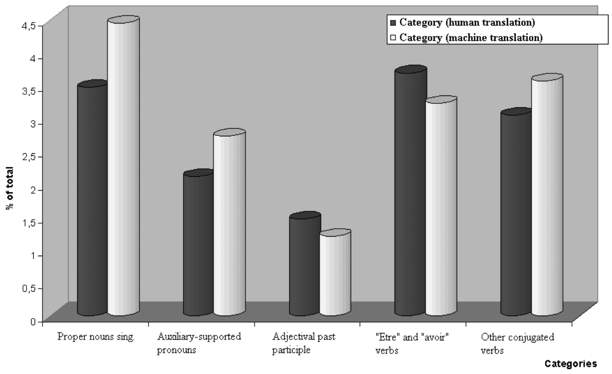


Table IV

Compared distribution of grammatical categories (French)



Since the various subsets of the corpus were coherent in our initial experiment, the distortion in the distribution of categories can only mean that it is sensitive to Natural Language Generation (NLG). What's more, the syntactic correctness was preserved despite altered relative values for each category and obviously with an altered sense, as usual with MT, owing to missing or incomplete semantic interpretation (Gouirand, 2002:75).

3 Discussion

Going back to Benford's law, the correlation with Zipf's law is that they are both based upon the principle of invariance (Zipf's law is insensitive to tagging while discriminating meaning, see Wilks 1996:6) and that there must be an underlying law which can account for the distribution. It has been claimed that lexical distributions relate to Zipf's law by virtue of word length ; however, it is easy to show that inflectional grammars such as German or Swedish have a different bearing on word lengths from those of, say, English or French. So the underlying law remains implicit and our duty as linguists is to make it explicit.

Yet the analogy is not complete, for one would not expect Benford's law to be relevant to linguistics as it conveys plain numerical rather than lexical information. The numbering of tags can be posited as arbitrary : the fact that some plot of land measures *n* acres doesn't tell much about what *n* actually *means*, it is simply the outcome of a composite geometrical law expressing surface areas. The numbers are merely arbitrary because whatever their name is, they always retain the underlying law's features, which is precisely why linguistic tags also apply : as Saussure (1916:98-100) showed at length, linguistic signs are arbitrarily chosen, they are signs pointing to "something else", at another level ; thus it is an underlying law which steers the said distribution.

Then, it is our contention that such an underlying law must be of semantic nature since the ancillary experiment set forth above clearly demonstrated that data-processing the corpora with machine translation software having a very limited semantic competence distorted the distribution. In other words, human processing of grammatical categories bears on ontological categories in the realm of semantics (Jackendoff, 1990:23) and therefore keeps the distribution within the boundaries of the law, whereas MT systems don't, insofar as there is no real semantic component attached to them (when they have one, they may to some extent).

Another proof of the validity of our model is given by feeding the text tags into a concordancer. Instead of using words as queries, we used our tags and then we reordered left-hand and right-hand side tags in decreasing order for each corpora.

As seen from Figure II and III (which only display the three most probable tags because as one moves down, volatility rises sharply), nouns (NN*x* in English/SB*x* in French), prepositions (IN/PREP) and determiners (DT/DTN) are once again found to be the most frequent top-level *n*-grams (ca. 70 percent of total) and roughly distributed in a way similar to categories, which makes sense : let us assume that tags were evenly scattered over all categories ; their count alone would create a vertical hierarchy within each category in building *n*-grams but this simply does not happen because the distribution takes place horizontally instead, meaning that the distribution is linked to relations among categories, which was the point to be made here.

Stating that syntactic constructs are driven by semantics means that ontological and/or biological nature of grammatical categories and their interplay can be investigated within the scope of an integrated theory of meaning where pre-verbal features of language are the driving force. In the domain of conceptual semantics (Jackendoff, 1983:57), syntactic structures parallel conceptual structures in that ontological categories (such as THING, EVENT, STATE, ACTION, PLACE, PATH, PROPERTY, AMOUNT) are the primitives associated with syntactic *constituents* ; by the way, some of the former are quite similar to Aristotle's findings, suggesting a continuum in the approach. However, in the internal workings of a MT system, invoking a concept from the lexical stratum is not straightforward, because lexemes may refer to several concepts, so clear-cut one-to-one correspondences are out of reach, at least for a computer program, so future research in lexicology should focus on a better mapping of lexical units into semantic categories.

For the French school of linguistics (Culioli, 1999:63), there is also a pre-verbal (and definitively pre-lexical) level of primitive relations where a distinction between syntax semantics is pointless. Hence the *notion* is a set of physical and cultural features that are inserted in a given context (Culioli, 1990:50). For Thom (1988:206), a concept is workable only through categorial operations which constrain the extensional perimeter centered around a prototypical core.

Another assumption is that the number of workable categories cannot be large or the system's output would result in unlimited semiosis and cognitive burdening. This is in line with Aristotle's initial formalization of categories but also with most speakers phenomenological experience.

The current state of this research allows an evaluation based on the distribution of grammatical *n*-grams that is useful to identify major translation failures but for a finer grained pinpointing, the definition of meaningful relations between syntax and semantics, parts-of-speech and meaning needs to be addressed. These relationships are to be grasped not only at the level of lexical entries analysable as semantic categories (with several impediments that are not yet solved) but also as ontological constructs or notional domains which are modelled as networks or semantic spaces. The overall sense then depends on such features but also on information conveyed by the discursive context as argumentative underpinning, demonstrated by Anscombre and Ducrot (1983:86) as well as Anscombre (1995:32), which provides directions for future research in the field.

Also, the question of whether the distributional features of grammatical categories in English and French are valid in other languages is still open. If they were to be ported, the metric would then be generic and so would the linguistic implications. There are obstacles, however : quality parallel corpora and robust parallel taggers for "exotic" languages (and even less exotic ones for that matter) are scarce, so the task of investigating the matter would be immensely time-consuming and it would require extensive linguistic expertise, let alone the "toolbox" problem. This is because some languages are often presented as resistant to some theories, for example Basque, which lacks a verbal category : the verbal function is still to be found though, in some other category from where it can be extracted ; In a language such as Swedish, determiners are embedded in nouns and can thus be backtracked. So applying our model to other languages wouldn't necessarily weaken it, only such an endeavour would require substantial efforts.

References

ALPAC. 1966. *Language and Machines. Computers in Translation and Linguistics.* A Report by the Automatic Language Processing Advisory Committee, Division of Behavioural Sciences, National Academy of Sciences. Publication 1416. National Research Council, Washington, DC.

Association for Computational Linguistics. 1994. *European Corpus Initiative (ECI)*. Multilingual Corpus 1(MCI) [CD-ROM]. ACL, Geneva/Edinburgh.

Jean-Claude Anscombre and Oswald Ducrot. 1983. *L'argumentation dans la langue*. Mardaga, Brussels, Belgium.

Jean-Claude Anscombre. 1995. *Théorie des topoï.* Kimé, Paris, France.

Frank Benford. 1938. The Law of Anomalous Numbers. *Proc. Amer. Phil. Soc.,* 78:551-554.

Antoine Culioli. 1990. *Pour une linguistique de l'énonciation. Opérations et représentations,* volume 1. Ophrys, Paris, France.

Antoine Culioli. 1999. *Pour une linguistique de l'énonciation. Formalisation et opérations de repérage,* volume 2. Ophrys, Paris, France.

EWG. 1996. *EAGLES Evaluation Group*. Final Report. Octobre 1996 version. Center foer Sprogteknologi, Copenhagen, Denmark.

Olivier Gouirand. 2002. Quelques réflexions sur la traduction automatique comme médiation. *Terminologie et Traduction (T&T)*. La revue des services linguistiques des institutions européennes, 2/2002:72-83. Office des Publications officielles des Communautés européennes, Luxembourg.

E. Hovy, M. King and A. Popescu-Belis. 2002. In Introduction to MT Evaluation. *Workbook of the LREC 02 Workshop on Machine Translation : Human Evaluators Meet Automated Metrics*, Las Palmas, Spain, 2002, pp. 1-7.

Institut National de la Langue Française (InaLF), Massachusetts Institute of Technlogy (MIT) and University of Pennsylvania. 1993-98. *WinBrill tagger and lemmatizer*. ATILF, Nancy, France.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press, Cambridge, Mass.

Ray Jackendoff. 1990. *Semantic Structures.* MIT Press, Cambridge, Mass.

Margaret King and Kirsten Falkedal. 1990. Using Test Suites in Evaluation of Machine Translation Systems. *Proceedings of COLING '90*, Helsinki, Finland, 1990.

Charles W. Morris. 1938. *Foundations of The Theory of Signs*. University of Chicago Press, Chicago, IL.

K. Papineni, S. Roukos, T. Ward et al. 2001. *BLEU : a Method for Automatic Evaluation of Machine Translation.* Computer Science IBM Research Report RC22176 (W0109-022). IBM Research Division, Yorktown, NY.

Parliament of Canada. 1999. *Debates of the House of Commons of Canada (Hansard), Dec. 1,2,3 & 10,* *1999*. Parliament of Canada, Ottawa. <http://www.parl.gc.ca/cgi-bin/36/pb_chb_hou_deb.pl?e> [page viewed April 28, 2000].

Ferdinand de Saussure. 1916. *Cours de linguistique générale.* Payot, Paris, France.

Karen Sparck-Jones and Julia Galliers. 1996. *Evaluating Natural Language Processing*. Springer, Berlin, Germany.

René Thom. 1988. *Esquisse d'une sémiophysique*. Interéditions, Paris, France.

René Thom. 1990. *Apologie du logos*. Hachette, Paris, France.

United Nations (UN). 1993. *Annual Report of the Secretary-General on the Work of the Organization*. UN, New York, NY. <http://www-rali.iro.umontreal.ca/arc-a2/BAF/> [page viewed on Aug 15, 2000].

UN. 1995. *Annual Report of the Secretary-General on the Work of the Organization.* UN, New York, NY. <http://www.un.org/Docs/SG/SG-Rpt/> [page viewed on Aug 15, 2000].

UN. 1996. *Annual Report of the Secretary-General on the Work of the Organization*. UN, New York, NY. <http://www.un.org/Docs/SG/sgrepor.htm> [page viewed on Aug 15, 2000].

UN. 1997. *Annual Report of the Secretary-General on the Work of the Organization.* UN, New York, NY. <http://www.un.org/Docs/SG/Report97/97con.htm> [page viewed on Jun 27, 2000].

UN. 1998. *Annual Report of the Secretary-General on the Work of the Organization.* UN, New York, NY. <http://www.un.org/Docs/SG/Report98/con98.htm> [page viewed on Jun 27, 2000].

UN. 1999. *Annual Report of the Secretary-General on the Work of the Organization.* UN, New York, NY. <http://www.un.org/Docs/SG/Report99/toc.htm> [page viewed on Aug 9, 2000].

UN. 2000. *Millennium Report of the Secretary-General*. United Nations, New York, NY. http://www.un.org/millennium/sg/report/ [page viewed on Aug 9, 2000].

J. White, T. O'Connell, F. O'Mara. 1994. *ARPA Machine Translation Program : 3Q94 Evaluation.* Georgetown University and Advanced Research Projects Agency, Washington, DC.

Yorick Wilks. 1996. *The grammar of sense: Is word-sense tagging much more than part-of-speech tagging?* Research Memoranda CS-96-05. University of Sheffield, Sheffield, UK..

George K. Zipf. 1965. *Psycho-Biology of Languages*. MIT Press, Cambridge, Mass.

APPENDIX

Figure I : illustration of spreadsheet processing

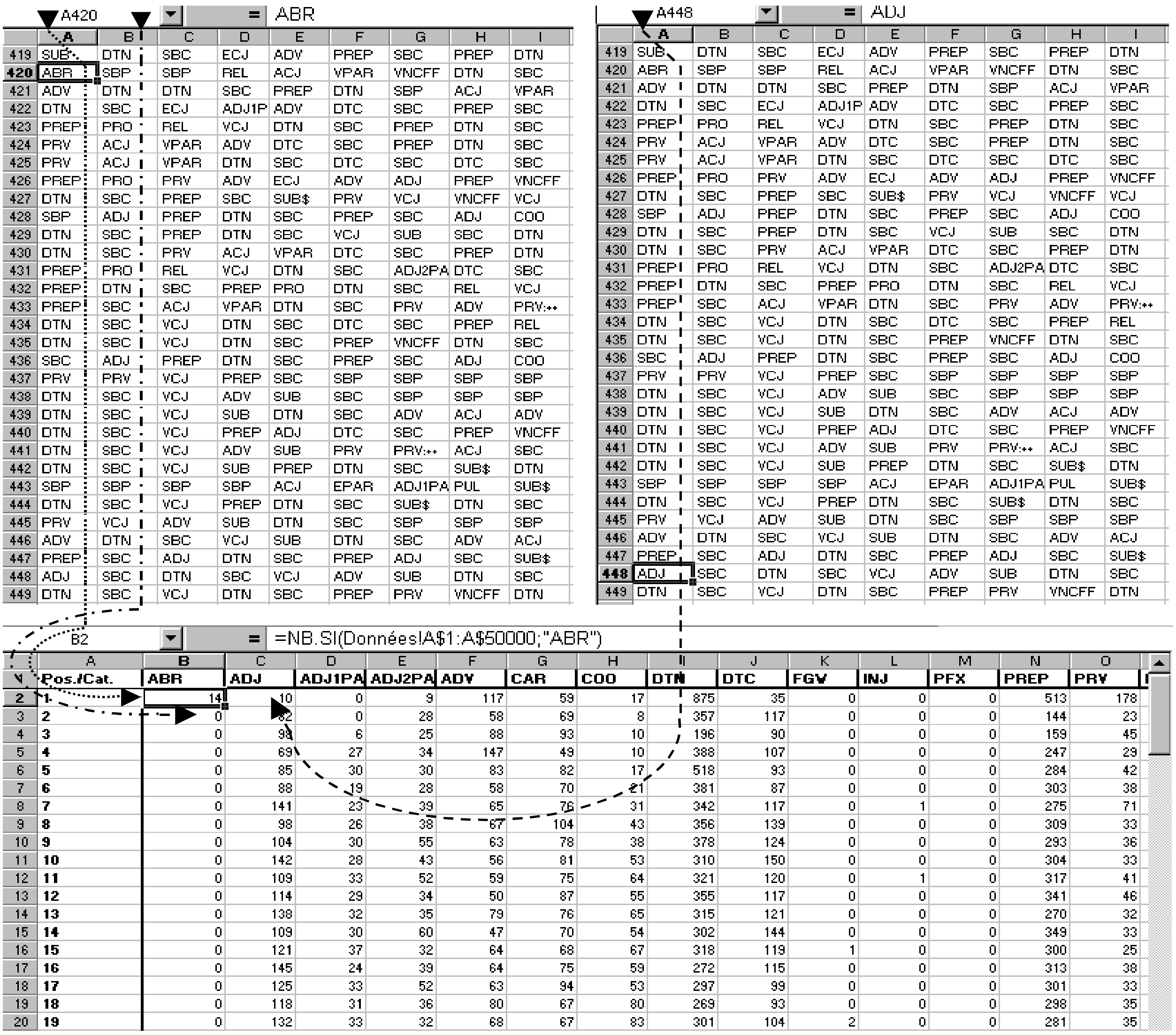


Figure II : Category n-grams for English

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
1174	2070	1657	1696	DT	jj	NN	1475	2858	2119	2596
690	1127	669	1237	IN	jj	NNS	869	1521	1023	1851
174	377	376	573	CC	jj	IN	217	388	359	573
152				JJ	jj	TO	171			
152				VB	jj	JJ	152			
131				VBD	jj	CC	146			
97				RB	jj	DT	79			
93				NN	jj	NNP	48			
91				VBN	jj	CD	14			
82				PRP$	jj	RB	13			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
12	24	43	54	DT	jjr	IN	23	73	73	89
9	23	20	36	IN	jjr	NN	16	45	57	83
9	21	20	30	RB	jjr	CC	4	21	31	38
4				PRP$	jjr	NNS	4			
3				CC	jjr	NNP	2			
3				NNS	jjr	CD	1			
3				VB	jjr	VBD	1			
2				NN	jjr	VBG	1			
2				VBD	jjr	DT	0			
2				VBG	jjr	PRP$	0			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
266	232	399	222	VBD	rb	VBN	330	336	486	358
200	196	346	216	VBZ	rb	IN	313	306	434	281
136	158	315	207	NN	rb	VB	229	248	376	245
136				RB	rb	DT	154			
118				NNP	rb	RB	136			
104				CC	rb	VBZ	124			
100				MD	rb	TO	99			
98				NNS	rb	JJ	97			
90				VBN	rb	VBD	85			
86				IN	rb	VBG	73			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
34	110	80	21	NN	wrb	DT	47	170	118	29
15	52	38	14	NNS	wrb	PRP	18	57	94	17
10	27	30	11	NNP	wrb	NN	10	52	29	13
9				CD	wrb	JJ	8			
7				IN	wrb	NNS	4			
5				RB	wrb	VBN	4			
4				VB	wrb	EX	3			
4				VBN	wrb	IN	3			
2				JJ	wrb	PRP$	3			
1				CC	wrb	RB	2			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
654	624	343	452	NNP	cd	NNP	496	342	277	227
521	474	247	231	NN	cd	IN	313	272	263	204
469	253	196	145	IN	cd	DT	189	217	252	179
140				NNS	cd	SYM	189			
132				CC	cd	NNS	184			
94				CD	cd	NN	182			
90				DT	cd	CC	165			
82				VBN	cd	CD	94			
33				TO	cd	VBD	61			
30				RB	cd	VBN	53			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
604	871	817	892	NN	cc	DT	275	490	304	549
466	527	404	683	NNS	cc	IN	265	315	268	497
323	221	282	449	NNP	cc	NN	243	299	221	350
165				CD	cc	NNP	202			
146				JJ	cc	JJ	174			
109				VBN	cc	NNS	171			
69				SYM	cc	CD	132			
43				VB	cc	RB	104			
20				RB	cc	VBN	73			
16				VBG	cc	VBD	68			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
161	149	213	106	NN	wdt	VBD	170	127	229	78
136	129	124	94	NNS	wdt	VBZ	67	103	95	60
80	129	39	66	IN	wdt	DT	65	97	90	49
37				NNP	wdt	VBP	55			
21				CC	wdt	PRP	52			
16				CD	wdt	MD	29			
15				TO	wdt	IN	12			
12				SYM	wdt	RB	11			
2				DT	wdt	NNS	9			
2				RB	wdt	NN	6			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
4796	5010	4058	3467	IN	**dt**	**NN**	**3975**	**4708**	**4559**	**2617**
621	745	972	522	VB	**dt**	**JJ**	**2620**	**2048**	**1657**	**1696**
474	614	556	350	TO	dt	NNP	1174	1098	1305	1572
425				VBZ	dt	NNS	1135			
332				VBG	dt	VBG	178			
303				VBD	dt	VBN	113			
293				NN	dt	CD	90			
275				CC	dt	RB	75			
211				NNP	dt	IN	36			
196				VBN	dt	SYM	33			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
438	126	585	196	NNP	vbd	VBN	604	108	337	219
435	86	377	186	NN	vbd	IN	387	65	321	199
313	66	157	163	NNS	vbd	DT	303	53	210	158
287				PRP	vbd	RB	266			
170				WDT	vbd	JJ	131			
85				RB	vbd	TO	113			
75				WP	vbd	NN	75			
68				CC	vbd	NNP	40			
61				CD	vbd	NNS	32			
38				EX	vbd	VBG	32			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
146	464	1219	403	NNS	vbp	VBN	122	328	346	294
55	127	375	235	PRP	vbp	RB	84	126	329	167
55	99	107	114	WDT	vbp	IN	50	105	306	127
36				NN	vbp	DT	42			
27				WP	vbp	JJ	30			
20				NNP	vbp	TO	18			
12				RB	vbp	VBG	12			
11				CC	vbp	NN	11			
8				IN	vbp	NNS	9			
7				CD	vbp	CC	6			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
423	836	717	501	NN	vbz	VBN	425	531	556	337
257	198	546	184	PRP	vbz	DT	325	395	399	239
237	158	229	136	NNP	vbz	RB	200	203	324	222
124				RB	vbz	IN	194			
99				IN	vbz	JJ	69			
67				WDT	vbz	TO	67			
54				CC	vbz	NN	43			
46				NNS	vbz	PRP$	22			
29				EX	vbz	NNS	21			
27				DT	vbz	VBG	19			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
149	675	579	156	NN	md	VB	392	1620	1069	517
121	531	272	152	NNS	md	RB	100	232	195	95
76	176	144	131	PRP	md	IN	16	31	47	9
65				NNP	md	CC	4			
29				WDT	md	TO	2			
17				CC	md	CD	1			
14				EX	md	DT	1			
9				VBN	md	JJ	1			
8				CD	md	MD	1			
7				DT	md	PRP	1			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
1151	1620	1785	1226	TO	vb	DT	474	1060	972	522
392	1063	1069	517	MD	vb	IN	312	754	663	251
229	306	434	151	RB	vb	VBN	266	334	324	244
56				CC	vb	NN	171			
27				NN	vb	JJ	152			
17				NNP	vb	NNS	112			
11				VBN	vb	TO	71			
8				NNS	vb	PRP	68			
6				CD	vb	RB	66			
6				VB	vb	PRP$	59			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
270	380	372	362	IN	vbg	DT	332	369	286	276
178	319	210	139	DT	vbg	NN	160	365	282	225
162	289	140	137	NNS	vbg	IN	129	254	203	196
145				NN	vbg	TO	116			
73				RB	vbg	NNS	112			
48				NNP	vbg	VBN	97			
38				CC	vbg	NNP	90			
36				CD	vbg	JJ	52			
32				VBD	vbg	PRP$	22			
26				TO	vbg	RB	18			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
604	1060	329	337	VBD	vbn	IN	1204	1816	786	847
330	531	324	294	RB	vbn	TO	300	406	249	244
298	425	324	281	NN	vbn	VBN	268	343	162	165
283				NNS	vbn	DT	196			
268				VBN	vbn	NN	156			
266				VB	vbn	CC	109			
194				VBZ	vbn	JJ	91			
122				VBP	vbn	RB	90			
113				DT	vbn	CD	82			
97				VBG	vbn	NNS	77			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
647	**588**	**790**	**548**	**NN**	**to**	**VB**	**1151**	**1063**	**1785**	**1226**
300	**343**	**262**	**301**	**VBN**	**to**	**DT**	**621**	**619**	**488**	**350**
189	316	249	244	NNP	to	NN	73	103	99	96
171				JJ	to	NNP	66			
160				VBG	to	JJ	61			
156				NNS	to	CD	33			
113				VBD	to	PRP$	28			
99				RB	to	NNS	27			
71				VB	to	VBG	26			
67				VBZ	to	IN	24			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
3276	3725	3263	3383	NN	in	DT	4501	4424	3850	3467
1492	1604	1274	1742	NNS	in	NN	1357	1877	1244	1390
1151	1594	748	847	VBN	in	NNP	875	1022	1101	1237
904				NNP	in	JJ	661			
369				VBD	in	NNS	502			
302				RB	in	CD	451			
294				CD	in	PRP$	347			
292				VBZ	in	IN	286			
286				IN	in	VBG	253			
281				VB	in	PRP	231			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
250	155	727	154	IN	prp	VBD	287	198	1219	235
68	113	322	102	VB	prp	VBZ	237	145	585	147
63	50	246	63	NN	prp	MD	76	74	579	136
52				WDT	prp	IN	74			
50				NNS	prp	VBP	55			
36				NNP	prp	RB	47			
21				TO	prp	TO	18			
19				VBD	prp	DT	15			
18				CC	prp	JJ	12			
18				WRB	prp	CC	10			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
58	7	77	16	NNS	wp	VBP	75	8	96	26
22	6	74	12	NN	wp	VBZ	27	5	91	19
17	6	59	11	NNP	wp	VBD	10	5	73	8
12				DT	wp	DT	8			
12				IN	wp	IN	8			
8				CC	wp	PRP	7			
6				VB	wp	MD	3			
3				VBG	wp	NN	2			
2				TO	wp	NNS	2			
1				CD	wp	JJ	1			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
375	111	424	299	IN	prp$	NN	313	106	466	264
59	35	108	95	VB	prp$	NNS	174	48	266	164
39	21	49	40	CC	prp$	JJ	82	45	149	160
31				VBD	prp$	NNP	25			
28				TO	prp$	CD	16			
22				VBG	prp$	VBN	9			
22				VBZ	prp$	JJR	4			
15				VBN	prp$	VBD	2			
6				RB	prp$	VBG	1			
5				NNS	prp$	CC	0			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
3927	2655	3770	2616	DT	**nn**	**IN**	**3442**	**2358**	**2809**	**3381**
1471	1979	1734	2594	JJ	**nn**	**NN**	**1066**	**1979**	**947**	**911**
1430	1755	1070	1389	IN	nn	NNS	711	996	685	891
1066				NN	nn	TO	640			
316				NNP	nn	CC	595			
310				PRP$	nn	CD	512			
240				CC	nn	VBD	435			
182				CD	nn	VBN	297			
169				VB	nn	DT	291			
155				VBN	nn	VBZ	253			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
1135	1529	1023	1851	DT	nns	IN	1595	1770	1338	1742
869	1521	918	784	JJ	nns	CC	466	531	404	683
713	993	485	681	NN	nns	VBD	313	527	375	403
536				IN	nns	VBN	283			
266				NNP	nns	DT	171			
184				CD	nns	VBG	162			
174				PRP$	nns	TO	156			
171				CC	nns	VBP	146			
112				VB	nns	CD	140			
112				VBG	nns	WDT	136			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
2620	1107	2916	1712	DT	nnp	NNP	2095	997	2916	1712
2095	997	1305	1572	NNP	nnp	IN	965	521	679	832
921	431	1123	1004	IN	nnp	CD	654	474	538	449
496				CD	nnp	VBD	438			
202				CC	nnp	VBZ	423			
114				NN	nnp	CC	323			
90				VBG	nnp	NN	322			
66				TO	nnp	NNS	266			
54				NNS	nnp	DT	211			
48				JJ	nnp	TO	189			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
43	16	42	387	IN	nnps	NNP	37	7	40	90
42	6	40	90	NNP	nnps	IN	19	3	33	76
12	3	32	24	JJ	nnps	CC	10	3	15	71
7				NN	nnps	NNS	7			
5				CD	nnps	VBD	7			
4				DT	nnps	NN	5			
4				TO	nnps	CD	4			
2				NNS	nnps	DT	2			
1				CC	nnps	JJ	2			
1				NNPS	nnps	TO	2			

Figure III : Category n-grams for French

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
3368	5927	2378	3773	SBC	adj	PREP	1181	2536	1302	1473
515	907	979	752	DTN	adj	SBC	908	1434	1038	1342
415	591	455	616	PREP	adj	DTC	634	738	459	893
340				SBP	adj	COO	513			
185				ADJ	adj	DTN	305			
174				ADV	adj	ADJ	185			
140				DTC	adj	VCJ	153			
116				COO	adj	ACJ	148			
99				ECJ	adj	.DJ2PAR	143			
61				CAR	adj	ADV	140			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
521	1311	461	375	EPAR	adj1par	PREP	670	1483	451	501
462	860	156	254	ECJ	adj1par	COO	98	191	63	57
111	224	132	152	ADV	adj1par	DTC	76	174	56	49
97				ENCFF	adj1par	DTN	64			
20				COO	adj1par	ADV	56			
3				DTC	adj1par	SUB	22			
3				PREP	adj1par	SUB$	22			
2				ADJ1PA	adj1par	SBC	19			
2				ENCNT	adj1par	ADJ	17			
1				ADJ	adj1par	.DJ2PAR	8			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
992	2480	615	808	SBC	adj2par	PREP	882	1721	389	685
143	389	95	167	ADJ	adj2par	DTC	180	345	120	133
127	212	85	150	ADV	adj2par	DTN	127	215	88	109
108				CAR	adj2par	ADV	71			
72				COO	adj2par	COO	71			
56				SBP	adj2par	SUB$	48			
41				VNCFF	adj2par	CAR	42			
38				ENCNT	adj2par	VCJ	36			
31				VCJ	adj2par	ACJ	29			
25				DTN	adj2par	SBC	23			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
535	830	937	630	SBC	adv	PREP	398	662	783	769
414	594	624	579	VCJ	adv	DTN	330	442	596	616
314	422	497	400	ACJ	adv	SUB	259	394	493	432
200				ECJ	adv	VCJ	242			
165				PRV	adv	ACJ	231			
158				ADV	adv	VPAR	198			
140				ADJ	adv	ADJ	174			
134				PREP	adv	ADV	158			
117				COO	adv	.DJ2PAR	127			
86				SBP	adv	ECJ	125			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
957	943	432	428	SBC	car	SBC	1029	1123	541	611
384	724	199	388	SBP	car	PREP	309	915	134	133
353	568	199	149	DTN	car	DTN	239	568	76	114
351				DTC	car	COO	214			
309				PREP	car	DTC	119			
108				COO	car	SBP	116			
91				CAR	car	.DJ2PAR	108			
90				ADJ	car	CAR	91			
42				ADJ2PA	car	ADJ	61			
35				VNCFF	car	ADV	60			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
1184	1884	1095	1534	SBC	coo	PREP	626	911	476	842
513	763	459	893	ADJ	coo	DTN	499	871	446	823
345	307	240	281	SBP	coo	SUB$	194	412	303	287
214				CAR	coo	DTC	172			
98				ADJ1PA	coo	SBC	160			
71				ADJ2PA	coo	SBP	134			
43				VNCFF	coo	ADV	117			
35				ADV	coo	ADJ	116			
29				VCJ	coo	CAR	108			
23				VPAR	coo	VCJ	98			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
2456	3484	1775	2180	SBC	dtc	SBC	3947	5845	2749	3535
634	744	379	491	ADJ	dtc	SBP	407	594	673	737
257	476	239	387	SBP	dtc	CAR	351	205	174	163
228				VNCFF	dtc	ADJ	140			
226				VCJ	dtc	ADV	16			
180				ADJ2PA	dtc	VNCFF	4			
172				COO	dtc	.DJ1PAR	3			
146				VPAR	dtc	.DJ2PAR	2			
119				CAR	dtc	PRO	1			
116				ADV	dtc	VCJ	1			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
4712	4692	3979	4744	PREP	dtn	SBC	8468	8673	7808	7826
609	830	838	824	SBC	dtn	SBP	761	493	898	1147
543	606	667	595	SUB	dtn	ADJ	346	434	731	544
524				VNCFF	dtn	CAR	214			
506				VCJ	dtn	DTN	94			
376				SUB$	dtn	ADV	45			
319				COO	dtn	PRO	25			
309				VPAR	dtn	.DJ2PAR	19			
234				ADJ	dtn	VCJ	8			
234				ADV	dtn	ACJ	4			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
686	262	642	388	SBC	acj	VPAR	987	143	919	811
351	108	361	349	PRV	acj	EPAR	504	143	350	288
231	75	213	207	ADV	acj	ADV	314	139	222	250
220				SBP	acj	DTN	118			
167				REL	acj	APAR	58			
148				ADJ	acj	PREP	53			
70				PRV:++	acj	SBC	18			
49				COO	acj	.DJ2PAR	17			
35				CAR	acj	PRO	6			
29				ADJ2PA	acj	COO	4			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
256	1009	860	408	PRV	ecj	.DJ1PAR	462	1311	497	375
249	370	430	349	SBC	ecj	ADV	200	422	461	337
125	293	195	118	ADV	ecj	ADJ	99	327	316	155
109				REL	ecj	PREP	85			
95				ADJ	ecj	DTN	70			
39				SBP	ecj	SBC	19			
21				CAR	ecj	SUB	10			
17				PRV:++	ecj	DTC	9			
14				ADJ2PA	ecj	PRO	7			
11				COO	ecj	.DJ2PAR	6			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
1271	1435	2002	954	SBC	vcj	DTN	697	979	937	743
604	1316	702	531	PRV	vcj	PREP	584	884	822	579
363	438	448	322	REL	vcj	SUB	584	862	730	510
242				ADV	vcj	ADV	414			
160				SBP	vcj	DTC	226			
153				ADJ	vcj	VNCFF	187			
98				COO	vcj	SBC	115			
45				CAR	vcj	ENCFF	67			
36				ADJ2PA	vcj	PRV:++	33			
28				PRO	vcj	.DJ2PAR	31			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
67	862	98	97	VCJ	encff	.DJ1PAR	97	860	97	114
28	86	42	37	PREP	encff	DTN	10	105	34	13
16	57	20	19	ADV	encff	SBC	10	35	13	13
10				VPAR	encff	ADJ	7			
4				SBC	encff	ADV	6			
3				ADJ	encff	.DJ2PAR	3			
3				COO	encff	DTC	3			
3				VNCNT	encff	VPAR	2			
2				PRV:++	encff	PREP	1			
1				ADJ1PA	encff	CAR	0			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
1135	1852	1485	1723	PREP	vncff	DTN	755	1569	1133	1196
187	766	513	281	VCJ	vncff	PREP	375	719	639	593
163	191	224	183	PRV	vncff	DTC	228	414	232	280
110				PRV:++	vncff	SUB	156			
53				VNCFF	vncff	ADV	73			
48				VPAR	vncff	SBC	53			
47				ADV	vncff	VNCFF	53			
35				COO	vncff	COO	43			
32				SBC	vncff	.DJ2PAR	41			
13				ADJ	vncff	CAR	35			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
178	397	130	151	SBC	vncnt	DTN	182	306	131	178
101	131	96	70	PREP	vncnt	PREP	108	230	72	104
60	91	28	61	ADJ	vncnt	DTC	63	89	41	44
34				PRV:++	vncnt	SUB	36			
21				PRV	vncnt	ADV	22			
19				SBP	vncnt	SBC	18			
17				CAR	vncnt	VNCFF	11			
14				COO	vncnt	.DJ2PAR	8			
8				ADJ2PA	vncnt	COO	7			
4				ADJ1PA	vncnt	SBP	6			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
504	125	175	288	ACJ	epar	.DJ1PAR	521	134	156	254
86	23	27	37	ADV	epar	ADV	31	8	12	31
19	4	2	4	ANCFF	epar	DTN	17	5	11	20
9				ANCNT	epar	VPAR	16			
3				SBC	epar	PREP	9			
1				ADJ	epar	SBC	8			
0				ADJ1PA	epar	VCJ	8			
0				COO	epar	ADJ	6			
0				DTN	epar	COO	2			
0				PREP	epar	PRO	2			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
987	139	919	811	ACJ	vpar	DTN	492	84	341	364
198	26	138	169	ADV	vpar	PREP	326	50	266	301
75	19	32	11	ANCFF	vpar	DTC	146	13	103	88
29				ANCNT	vpar	SUB$	138			
16				EPAR	vpar	VNCFF	48			
11				DTN	vpar	SBC	44			
7				PREP	vpar	ADV	38			
7				VPAR	vpar	COO	23			
5				PRO	vpar	CAR	16			
4				COO	vpar	ENCFF	10			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
4733	5459	4599	4538	SBC	prep	DTN	5008	4326	4324	4463
901	1026	878	1034	ADJ	prep	SBC	2402	3705	2351	2597
675	632	685	563	ADJ2PA	prep	VNCFF	833	855	1195	1146
489				ADJ1PA	prep	ADJ	313			
460				COO	prep	PRO	265			
425				VCJ	prep	SBP	244			
303				SBP	prep	CAR	194			
291				ADV	prep	REL	138			
283				VNCFF	prep	PRV	137			
227				VPAR	prep	ADV	91			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
270	383	591	317	SUB$	prv	VCJ	604	1316	2002	954
250	301	554	249	SUB	prv	ACJ	351	370	860	408
236	207	520	224	SBC	prv	ECJ	256	167	642	349
187				PREP	prv	ADV	165			
187				REL	prv	VNCFF	163			
113				PRV	prv	PRV	113			
75				COO	prv	PRV:++	49			
67				ADV	prv	VNCNT	21			
53				ADJ	prv	ADJ	2			
45				SBP	prv	ENCNT	2			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
60	102	214	111	PREP	prv:++	VNCFF	110	191	224	183
49	72	119	69	PRV	prv:++	ACJ	70	75	213	63
34	71	106	51	ADV	prv:++	VNCNT	34	39	73	52
33				VCJ	prv:++	VCJ	26			
28				SBC	prv:++	.DJ2PAR	22			
17				ADJ	prv:++	ECJ	17			
11				REL	prv:++	PRV:++	8			
10				COO	prv:++	SBP	3			
10				SBP	prv:++	ANCFF	2			
8				PRV:++	prv:++	ENCFF	2			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
360	288	370	249	PREP	pro	REL	226	144	210	142
38	166	161	109	DTN	pro	SUB$	74	133	185	97
33	54	84	37	SBC	pro	PREP	59	120	151	73
31				SUB	pro	DTC	57			
26				SUB$	pro	ADV	31			
22				ADJ	pro	VCJ	28			
21				COO	pro	DTN	23			
12				VCJ	pro	ACJ	19			
9				CAR	pro	SBC	17			
8				ADV	pro	COO	13			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
5392	4254	5436	5108	DTN	sbc	PREP	2830	3300	2832	3214
1778	2637	1441	1790	PREP	sbc	ADJ	1542	1336	1348	1771
1411	1500	1394	1683	DTC	sbc	DTC	1117	891	915	1065
411				ADJ	sbc	VCJ	548			
394				CAR	sbc	COO	490			
88				SBC	sbc	CAR	440			
64				SBP	sbc	.DJ2PAR	434			
62				COO	sbc	DTN	392			
52				VCJ	sbc	SBP	311			
26				VNCFF	sbc	ACJ	292			

N	N2	N3	N4	Left	NODE	Right	N	N2	N3	N4
1868	1962	1109	1411	SBP	sbp	SBP	1868	1162	997	668
1066	1014	997	737	DTN	sbp	PREP	424	1014	511	557
774	1000	673	557	SBC	sbp	CAR	384	724	335	428
407				DTC	sbp	COO	345			
385				PREP	sbp	ADJ	340			
134				COO	sbp	DTN	293			
123				ADJ	sbp	DTC	257			
116				CAR	sbp	ACJ	220			
21				ADJ2PA	sbp	SBC	191			
16				VCJ	sbp	VCJ	160			