Obedkova Maria, Statistical Methods in NLP II, HW2

1. Best Friends

==========

Text TEXTCZ1.txt

==========

Tokens count 222412

Unigram count 42827 sum 222412

Bigram count 147137 sum 222411

==========

Observed collocations for TEXTCZ1.txt using bigram distance 1 with PMI score.

Hamburger	SV	14.28895040192613
Los	Angeles	14.06244187211745
Johna	Newcomba	13.762881590258543
Č.	Budějovice	13.633598573313575
série	ATP	13.468967871540775
turnajové	Série	13.434410649504409
Tomáš	Ježek	13.428980853705104
Lidové	Noviny	13.329922182982436
Lidových	Novin	13.271028493928867
veřejného	Mínění	13.06244187211745
teplota	Minus	12.981521876733883
Ján	Čarnogurský	12.955526668200939
jaderné	Zbraně	12.955526668200939
Milan	Máčala	12.897811170344651
lidských	Práv	12.862877380235263
společném	Státě	12.708433806236165
akciových	společností	12.692492262367145
Pohár	UEFA	12.625378066508606
privatizačních	Projektů	12.615676665316313
George	Bushe	12.603010253480152

==========

Text TEXTEN1.txt

==========

Tokens count 221098

Unigram count 9607 sum 221098

Bigram count 73246 sum 221097

==========

Observed collocations for TEXTEN1.txt using bigram distance 1 with PMI score.

La	Plata	14.169370473805218
Asa	Gray	14.031866950055282

Fritz	Muller	13.362015551747612
worth	while	13.332869206088096
faced	tumbler	13.262479878196698
lowly	organised	13.216898843887803
Malay	Archipelago	13.110476784751649
shoulder	stripe	13.053893256385281
Great	Britain	12.914556574776391
United	States	12.847442378917854
English	carrier	12.525514284030493
specially	endowed	12.401816559805589
Sir	J	12.377363516081047
branched	off	12.377363516081047
de	Candolle	12.362015551747612
mental	qualities	12.362015551747612
Galapagos	Archipelago	12.344942038388671
red	clover	12.32388042286084
self	fertilisation	12.316927662219076
systematic	affinity	12.25183263399719

We can see here proper names, phrasal verbs and words that usually occur together because they have compound meaning.

We did not get negative PMI values because we are considering the top 20 results where joint probability of two words is bigger than multiplication of probabilities of independent occurrences of two words. If we consider the end of this list, we probably will see negative values because the number of independent occurrences of each of the words will be higher than the number of their joint occurrences.

==========

Text TEXTCZ1.txt

==========

Tokens count 222412

Unigram count 42827 sum 222412

Bigram count 5267775 sum 10896914

==========

Observed collocations for TEXTCZ1.txt using bigram distance 2-50 with PMI score.

výher	výher	9.855552171462799
žel	žel	9.13636127902457
Bělehrad	Benfica	8.95193670788714
h	teplota	8.90040640724706
13h	13h	8.8555521714628
ODÚ	VPN	8.826405825803283
Sandžaku	Sandžaku	8.826405825803283
Petrof	Petrof	8.785763841305936
Atény	Benfica	8.581987098136835
13h	zataženo	8.575444252270064
13h	skoro	8.547429876100468

CIA	CIA	8.504477730915921
vychází	h	8.485368907968216
výher	IV	8.478482522382976
13h	st	8.456456216052977
pořadí	výher	8.456456216052977
Bělehrad	Kyjev	8.439382702694035
IFS	IFS	8.437840537885629
km	žel	8.435145709807564
Benfica	Bělehrad	8.366974207165985

==========

Text TEXTEN1.txt

==========

Tokens count 221098

Unigram count 9607 sum 221098

Bigram count 2083895 sum 10832528

==========

Observed collocations for TEXTEN1.txt using bigram distance 2-50 with PMI score.

dried	floated	8.735396013752029
floated	dried	8.64793317250169
dried	germinated	8.358426555306703
avicularia	vibracula	8.313815668606413
dried	dried	8.303285001114244
eastern	Pacific	8.301843026940338
stripe	shoulder	8.216954129353823
floated	germinated	8.210869366892847
floated	floated	8.192253688725499
layer	hexagonal	8.165781477364309
survival	fittest	8.139786268831365
dried	days	8.097966093136737
heath	heath	8.060834927436542
dimorphic	trimorphic	8.060834927436542
geese	webbed	8.028277953614372
clover	clover	8.02644479575542
floated	days	8.02048734045902
germinated	dried	7.94338905602786
carrier	faced	7.8178581739440025
CHAPTER	THE	7.8178581739440025

We got word pairs with repeated words because it is highly likely that a word can be repeated later in a sentence. As we can see, there are no collocations in the list but some semantic relatedness can be traced.

2. Word Classes

==========

Obedkova Maria, Statistical Methods in NLP II, HW2

Text TEXTCZ1.ptg

==========

History of merges for Czech words for first 8 000 tokens.

#classes	MI	Loss	Merged words		
61	7.5580016277741535	0.0030828842705999185	listopadu	+	OKD
60	7.554918743503553	0.0033734316421669547	který	+	které
59	7.551545311861386	0.004024966367527465	J	+	státu
58	7.547520345493866	0.0044216001660482285	bude	+	musí
57	7.543098745327817	0.004603782679326721	aby	+	ale
56	7.538494962648488	0.004647048343347088	nás	+	bylo
55	7.533847914305141	0.004991409281117513	pouze	+	si
54	7.528856505024022	0.0050000000000000044	mezi	+	už
53	7.523856505024026	0.00556127812445914	NATO	+	&slash;
52	7.518295226899566	0.005564269906886257	jeho	+	ze
51	7.51273095699267	0.005742784599970026	být	+	však
50	7.506988172392698	0.005749999999999998	byl	+	jsou
49	7.501238172392688	0.005999999999999998	?	+	před
48	7.495238172392681	0.006249999999999992	budou	+	jako
47	7.488988172392683	0.006391442323723712	Na	+	listopadu
46	7.482596730068964	0.006899419316170927	zákona	+	při
45	7.4756973107527935	0.0075558125338524434	za	+	u
44	7.468141498218931	0.00760310398655982	J	+	ČSFR
43	7.4605383942323655	0.0077407546362640856	který	+	aby
42	7.4527976395961035	0.0077677040973247805	včera	+	nás
41	7.445029935498771	0.008789719718317365	od	+	mezi
40	7.436240215780454	0.010210213421119205	bude	+	být
39	7.426030002359341	0.010213905859793385	byl	+	pouze
38	7.415816096499543	0.01025000000000005	po	+	pro
37	7.405566096499555	0.010649206557307167	V	+	Na
36	7.3949168899422455	0.011670205302727568	jeho	+	to
35	7.383246684639511	0.011811278124459104	ve	+	NATO
34	7.371435406515043	0.012485569199940066	budou	+	?
33	7.35894983731511	0.013123668537318306	který	+	že
32	7.345826168777775	0.013298794940695385	k	+	zákona
31	7.332527373837088	0.014886997925414425	včera	+	by
30	7.317640375911702	0.015907961185191027	do	+	za
29	7.301732414726547	0.016596309274685833	)	+	J
28	7.285136105451854	0.01681233428201584	i	+	je
27	7.2683237711698325	0.018072324377885474	:	+	po
26	7.250251446791958	0.018164396525671284	od	+	bude
25	7.2320870502662675	0.02080778873941319	ve	+	z
24	7.211279261526833	0.02123254501589912	(	+	byl
23	7.1900467165109285	0.02413792427036343	s	+	budou
22	7.165908792240571	0.024390009805841953	k	+	jeho
21	7.1415187824347575	0.026095823503601456	V	+	
20	7.1154229589311555	0.02894319910029089	včera	+	i
19	7.086479759830832	0.03294986076568068	do	+	od
18	7.053529899065105	0.03467795337934593	:	+	-

17	7.018851945685794	0.03761236390971759	)	+	ve
16	6.9812395817760855	0.04326932588696272	(	+	o
15	6.937970255889177	0.047543977145028116	k	+	s
14	6.890426278744138	0.050139737986712535	včera	+	který
13	6.840286540757499	0.0536432796701477	se	+	na
12	6.786643261087394	0.058668339011072046	V	+	)
11	6.727974922076265	0.06915184188756723	:	+	do
10	6.658823080188756	0.07330603764174076	(	+	v
9	6.585517042546945	0.08678617733435695	k	+	a
8	6.4987308652126465	0.09655450347632993	včera	+	se
7	6.402176361736232	0.12667284415642238	V	+	:
6	6.275503517579718	0.1499044684127691	(	+	k
5	6.125599049166728	0.1744488937327484	,	+	včera
4	5.951150155434103	0.21513105882098993	V	+	.
3	5.736019096613375	0.294743215882453	(	+	,
2	5.44127588073123	0.45987061788252226	V	+	(

Members of 15 classes.

1) V Na listopadu OKD "
2) .
3) : po pro –
4) (byl jsou pouze si o
5) ,
6)) J státu ČSFR ve NATO &slash; z
7) včera nás bylo by i je
8) k zákona při jeho ze to
9) se
10) a
11) do za u od mezi už bude musí být však
12) na
13) s budou jako ? před
14) který které aby ale že
15) v

==========

Text TEXTEN1.ptg

==========

History of merges for words for first 8 000 tokens.

#classes	MI	Loss	Merged words		
112	4.997263261625145	0.0021965665335756945	subject	+	case
111	4.9950666950915705	0.002669139511099379	may	+	cannot
110	4.992397555580471	0.0026748091526128774	structure	+	individuals
109	4.989722746427855	0.0034794003704524297	It	+	there
108	4.986243346057402	0.0036556390622295605	even	+	less
107	4.982587706995175	0.003690950620214791	variation	+	nature
106	4.978896756374957	0.0038977958561413548	shall	+	see
105	4.974998960518816	0.0039056390622295677	short	+	slight

104	4.971093321456585	0.003992156793911787	certain	+	distinct
103	4.967101164662676	0.004241409281117516	such	+	manner
102	4.962859755381556	0.004262927799019989	must	+	can
101	4.958596827582534	0.004276663948769327	subject	+	state
100	4.954320163633762	0.004298794940695395	what	+	differ
99	4.950021368693065	0.004456435556800396	)	+	cases
98	4.945564933136266	0.004478026866846278	me	+	only
97	4.94108690626942	0.004540204221812914	nearly	+	how
96	4.936546702047606	0.004563547053784804	domesticated	+	domestic
95	4.931983154993819	0.004613365915588918	varieties	+	races
94	4.927369789078232	0.004762832413873118	if	+	when
93	4.922606956664359	0.004872805225352821	(	+	than
92	4.9177341514390065	0.004949203505429163	do	+	believe
91	4.912784947933577	0.005125548031345378	my	+	great
90	4.907659399902228	0.005600648386150979	will	+	could
89	4.902058751516075	0.0056162670739184	structure	+	variation
88	4.89644248444216	0.005689639960195725	facts	+	plants
87	4.890752844481965	0.005755267783733753	:	+	between
86	4.88499757669823	0.005811278124459127	its	+	different
85	4.879186298573773	0.005975681770144445	In	+	The
84	4.873210616803624	0.0060700564650807	conditions	+	breeds
83	4.867140560338544	0.006271686891698415	short	+	long
82	4.860868873446843	0.006340717981448564	these	+	each
81	4.854528155465395	0.006573989711723821	may	+	would
80	4.847954165753669	0.006827542130200634	any	+	very
79	4.84112662362347	0.0068922354584405945	many	+	most
78	4.834286267852441	0.00705157544418946	often	+	so
77	4.827234692408247	0.007132892772609464	we	+	they
76	4.820101799635638	0.007267364206776171	much	+	what
75	4.812834435428863	0.0077607636136982905	do	+	shall
74	4.805073671815167	0.007963306419960459	more	+	even
73	4.797110365395207	0.008140838062060646	it	+	It
72	4.788969527333148	0.008021486604234428	)	+	animals
71	4.780948040728919	0.008262320542435385	certain	+	an
70	4.772685720186485	0.008613791297558793	may	+	must
69	4.7640719288889235	0.008613981720057576	such	+	wild
68	4.755457947168874	0.009403894369706636	all	+	nearly
67	4.746054052799175	0.009566699054152673	(	+	but
66	4.736487353745029	0.009645128745293885	subject	+	conditions
65	4.726842224999732	0.00971936038246713	one	+	other
64	4.717122864617261	0.009771757892383	our	+	my
63	4.707351106724876	0.009939401578158993	me	+	at
62	4.697411705146716	0.010202515793020624	some	+	these
61	4.687209189353697	0.010269270619067447	several	+	its
60	4.676939918734621	0.010783272624463762	species	+	varieties
59	4.66615664611016	0.011352568841686311	:	+	under
58	4.654804077268472	0.011381899303950685	for	+	if
57	4.643422177964519	0.011689745227658754	any	+	their
56	4.631732432736847	0.011892908154906491	facts	+	)
55	4.619839524581943	0.01178550974034076	such	+	domesticated

54	4.608054014841602	0.01271421911097162	same	+	many
53	4.595391675418038	0.01271875176254414	will	+	may
52	4.582672923655502	0.014081256250479762	with	+	by
51	4.5685916674050295	0.014296521573941304	not	+	often
50	4.554295145831087	0.01515438594956188	subject	+	structure
49	4.539140759881527	0.015230507292150683	I	+	we
48	4.523910252589384	0.015314995035446463	short	+	more
47	4.508595257553931	0.01605976960660145	much	+	all
46	4.492535487947327	0.016531788868154693	certain	+	several
45	4.4760036990791825	0.01674273120342816	has	+	is
44	4.459260967875757	0.016743740756604744	this	+	which
43	4.442517227119157	0.01734735788363606	as	+	(
42	4.425169869235511	0.01762608134527688	on	+	me
41	4.407543787890236	0.018288141427282342	facts	+	species
40	4.389255646462947	0.01816467774864086	have	+	do
39	4.371090968714305	0.01845061772228135	been	+	not
38	4.352640350992023	0.01970986727465046	some	+	any
37	4.332930483717359	0.02020325299887775	one	+	same
36	4.312779110405903	0.02126695847118454	:	+	from
35	4.2915121519347235	0.021519518313568187	I	+	it
34	4.2699926336211576	0.021834911514086563	for	+	or
33	4.248157722107066	0.023487676836598886	certain	+	our
32	4.22467004527047	0.02394320993616625	In	+	are
31	4.200726835334311	0.028497520212267924	short	+	such
30	4.172229315122055	0.028739567510663644	facts	+	subject
29	4.1434897476113965	0.028900814156270216	has	+	have
28	4.114588933455133	0.029286088071844962	some	+	a
27	4.085354725070692	0.02932014166350655	as	+	for
26	4.056034583407185	0.029476881178438857	much	+	this
25	4.026557702228736	0.029410657291278286	with	+	that
24	3.9971470449374595	0.03401557225887736	be	+	been
23	3.9631314726785773	0.034440093724066784	on	+	:
22	3.9286913789545137	0.035106133434460174	will	+	has
21	3.893585245520054	0.037698251654943976	certain	+	one
20	3.855938873552507	0.03840787951742555	;	+	In
19	3.8175309940350823	0.04821396965251373	on	+	in
18	3.7693170243825858	0.05021157716667879	much	+	with
17	3.7191054472159077	0.05227430565052835	as	+	and
16	3.666831141565386	0.05648931257478906	certain	+	short
15	3.6103937086779903	0.06492074006430248	to	+	be
14	3.5454729686136934	0.0674327519980753	the	+	some
13	3.478092096303017	0.07693586877459832	I	+	;
12	3.401156227528414	0.09276780591575623	on	+	of
11	3.3083884216126584	0.08952305957156542	certain	+	facts
10	3.218917241728507	0.1024620163002668	to	+	will
9	3.116455225428239	0.10269885621583974	as	+	much
8	3.013756369212419	0.11957091447329193	.	+	,
7	2.894185454739137	0.1139706860528677	as	+	I
6	2.780214768686225	0.18946397400202647	certain	+	the
5	2.5908545540590495	0.1935109255462631	as	+	to

4	2.3973436285128114	0.22608907966631003	on	+	.
3	2.171254548846486	0.3003960329048642	as	+	certain
2	1.8709622753165487	0.43820115197646387	on	+	as

Members of 15 classes.

1) on me only at : between under from in
2) .
3) as (than but for if when or and
4) ,
5) I we they it It there
6) much what differ all nearly how this which with by that
7) certain distinct an several its different our my great one other same many most short slight long more even less such manner wild domesticated domestic
8) facts plants) cases animals species varieties races subject case state conditions breeds structure individuals variation nature
9) the
10) of
11) to
12) will could may cannot would must can has is have do believe shall see
13) be been not often so
14) some these each any very their a
15) ; In The are

We can see that some words are organized in meaningful groups. 1) consists mostly of prepositions, 2) includes conjunctions, 5) describes pronouns, 7) gathers mostly adjectives and prenominal words, 8) includes nouns, 12) consists of modal verbs and some meaning verbs, 6) and 14) mixes quantifiers, pronouns and determiners. Interestingly, all initial merges occur between semantically similar words.

3. Tag Classes

==========

Text TEXTCZ1.ptg

==========

History of merges for Czech tags for first 40 000 tokens.

#classes	MI	Loss	Merged tags		
343	1.8800186052587906	6.887218755408618e-05	PDFS6----------	+	CrFS6----------
342	1.8799497330712365	0.0001708473905425601	P4FS1----------	+	PE--1----------
341	1.8797791173924043	0.00017728748056790962	PP-P2--1-------	+	P5XP2--3-------
340	1.8796018299118369	0.00018118574037376094	PJYS1----------	+	P4XP3----------
339	1.8794207020993907	0.00018134024509867692	ClFS6----------	+	PDFS6----------
338	1.8792393618542924	0.00018834180623305587	ClXP6----------	+	P8XP6----------
337	1.8790510200480597	0.00020000965465464095	A2--------A----	+	Vf--------A---1
336	1.87885102004806	0.00020824844434264803	Vi-P---1--A----	+	Vi-P---2--A----
335	1.8786428198769909	0.0002145629458345285	P6-X4----------	+	P5XP4--3-------
334	1.8784282569311561	0.00021894942479121606	P4NS1----------	+	P4IS4----------
333	1.8782093847436023	0.00022395889711063232	AAFS1----1N----	+	NNFS1-----A---1
332	1.8779854548104553	0.00022758507619400617	AAMP1----1N----	+	ACMP------A----

331	1.8777578697342618	0.0002305800922819591 8	CrFS1----------	+	AAFS1----3A----
330	1.877527337915253	0.000243529708980847 7	P5FS6--3-------	+	P5XP6--3-------
329	1.8772838082062724	0.000251806336825494 4	AANS1----1N----	+	AAIS1----1N----
328	1.8770320018694464	0.000264757088971894 1	PSZS1FS3-------	+	AAMS1----1N----
327	1.876767249607802	0.000266448304927224 4	NNFSX-----A----	+	NNFS5-----A----
326	1.87650086405813	0.000276134538940419 13	P5ZS7--3-------	+	PDXP7----------
325	1.876224729519189	0.000277003759557061 4	P8XP2----------	+	AAFP2----1N----
324	1.8759477257596313	0.000277585076194005 86	P8FP4---------1	+	PZFP4----------
323	1.8756701406834377	0.000288414300468849 3	PZMP1----------	+	ClZS7----------
322	1.8753817505196053	0.000291826340531257 1	Vi-S---2--A----	+	AAFS4----1N----
321	1.8750899724523469	0.000292185295940646 9	CrIS1----------	+	AAIS1----3A----
320	1.8747978161203702	0.000293960626995345 63	P5ZS6--3-------	+	PDXP6----------
319	1.8745038554933748	0.000296726270966967 4	P8IS4----------	+	PZ--4----------
318	1.874207138877062	0.000308701716553241 86	AUFS6M---------	+	ClFS6----------
317	1.8738984419878364	0.000308981895858794 1	AANS4----1N----	+	P8NS4---------1
316	1.8735894697466322	0.00031552563949398	PDIP1----------	+	AGIP1-----A----
315	1.8732739634164475	0.000323405656068827	PJYS2----------	+	AGIS1-----A----
314	1.8729506060336523	0.000326914601029490 26	AGNS1-----A----	+	AAXXX----1A---8
313	1.8726237059146045	0.000327849256085560 16	Dg-------3A---1	+	PH-S4--1-------
312	1.8722958614858465	0.000327858910740201 45	AAIP1----1N----	+	PZFP1----------
311	1.8719680170570885	0.000331127812445913 57	AAIS1----2A----	+	AAFP1----1N----
310	1.8716368892446422	0.000336167545980189 5	ClFS1----------	+	AAFS1----2A----
309	1.871300731353316	0.000349218080392393 8	Vf--------N----	+	P8YP4---------1
308	1.8709515181002514	0.000350004827327320 5	PLFP4----------	+	PDMP4----------
307	1.8706015181002518	0.000350805971967631 4	PDFS1----------	+	AUFS1M---------
306	1.8702507555742303	0.000352673515635265 35	P4ZS6----------	+	P8ZS6----------
305	1.869898082058595	0.000358972241204152 94	VB-P---3P-NA--1	+	VB-P---3F-NA---
304	1.869539109817392	0.000368881842208727 83	PP-P3--2-------	+	ClFS7----------
303	1.8691702376298376	0.000375665629098865 3	P8FS2---------1	+	AGFS2-----A----
302	1.8687945720007386	0.000381127812445913 7	PPFS3--3-------	+	PLNS1---------1
301	1.8684134441882934	0.000394101952107618	PSZS2-P1-------	+	P8ZS2----------
300	1.8680193422361864	0.000394942954755360 4	ClFS4----------	+	PSFS4-P1-------
299	1.8676244041087584	0.000396070827588254 1	PP-P2--1-------	+	P9XP2----------
298	1.8672283332811692	0.000400019309309282 55	P1XXXXP3-------	+	NNIS2-----A---1
297	1.8668283332811695	0.000426823344672300 47	PJYS1----------	+	P4IP1----------
296	1.8664016209650256	0.000429656212059326 74	P4FP4----------	+	P4ZS7----------
295	1.865972017853567	0.000434967218963374 85	AUIS1M---------	+	PDXP3----------
294	1.8655370506346038	0.000435049432686053 5	AANP1----1A----	+	CrNP1----------
293	1.8651020687844995	0.000439456262524911 9	AAMS4----1A----	+	PDFP4----------
292	1.8646626463132656	0.000440583442347308 1	VpMP---XR-AA--1	+	PPXP2--3-------
291	1.8642220821802276	0.000445629510005856 46	A2--------A----	+	AGNS1-----A----
290	1.8637764768068588	0.000447597113625826 73	Cn-P2----------	+	Ca--2----------
289	1.8633288845205602	0.000450655050255050 45	P5FS6--3-------	+	NNMS6-----A----
288	1.8628782342976322	0.000455517980704265 23	ClYP4----------	+	Vf--------N----
287	1.8624230641452426	0.000481127812445913 5	PLXP3----------	+	AANP4----1A----
286	1.861941936332797	0.000488504661207823 2	ClXP1----------	+	PSHS1-P1-------
285	1.861453455808226	0.000497593723986227 5	Vc-P---1-------	+	PPFS4--3-------
284	1.8609558620842392	0.000498100422125338 2	VB-S---1P-AA--1	+	VB-S---2P-AA---
283	1.8604578147627138	0.000517244771417981 6	P5ZS6--3-------	+	P9ZS6----------
282	1.8599405699912959	0.000518569439809389 8	PZMP1----------	+	PWM-1----------

281	1.8594220295154502	0.0005195731060065774	AAFS7----1A----	+	P8FS7---------1
280	1.858902490200733	0.0005240751147493642	PPYS1--3-------	+	PQ--1----------
279	1.8583785164598576	0.0005265324767435172	VB-S---3F-NA---	+	AAMS3----1A----
278	1.8578520081197503	0.0005349545979805503	NNIXX-----A---8	+	NNMXX-----A---8
277	1.8573171066223706	0.0005494186833630561	P8FS4---------1	+	Vc-S---1-------
276	1.8567676879390071	0.0005497296593615394	VB-S---1P-NA---	+	VB-P---1P-NA---
275	1.8562180258622274	0.0005521990411266509	VpTP---XR-NA---	+	AAMP1----1N----
274	1.8556658268210993	0.0005578289795561946	ClXP7----------	+	AAMP7----1A----
273	1.8551080026688704	0.0005598541762053748	PZ--1----------	+	PW--1----------
272	1.8545481533199923	0.00056824022942094	P4NS1----------	+	P4FP4----------
271	1.8539800434284095	0.0005699478423760035	CrFS1----------	+	AAFS1----1N----
270	1.8534101728232704	0.0005731154820221339	ClHP1----------	+	PDFP1----------
269	1.8528370959598672	0.0005794038534137937	VB-P---3P-AA--1	+	VB-P---3P-NA--1
268	1.8522577162430898	0.0005896596952239758	AUIS2M---------	+	PP-S4--1-------
267	1.851668056547866	0.0005963213733298552	VpTP---XR-AA--1	+	AGMP2-----A----
266	1.8510717544838444	0.000599352035763564	CrIS1----------	+	PSZS1FS3-------
265	1.850472436239372	0.0006064333625330764	P4FS4----------	+	PKM-1----------
264	1.849866162178641	0.0006121107625459907	AAFS6----1A----	+	AUFS6M---------
263	1.849254061070749	0.0006205689031046407	P8FP4---------1	+	PLFP4----------
262	1.8486334969949718	0.0006212263104807404	NNFSX-----A----	+	NNMS5-----A----
261	1.8480123624037104	0.0006245257317656155	PSFS2-P1-------	+	PSFSXFS3-------
260	1.8473878511539261	0.0006264662506490406	PLXP2----------	+	PZXP2----------
259	1.8467613849032778	0.0006347783932303206	AANP6----1A----	+	ClXP6----------
258	1.8461266065100477	0.0006391831387904594	PW--4----------	+	PLNS1----------
257	1.8454874378532382	0.0006444601696923655	Ca--1----------	+	Ca--4----------
256	1.844842992165528	0.0006526210591368165	Cv----------	+	PP-S1--1-------
255	1.8441904338616464	0.0006540316557507261	AGFS4-----A----	+	P1XXXXP3-------
254	1.8435364263425318	0.0006597687552659433	AANS1----1N----	+	AUIS1M---------
253	1.8428766575872657	0.0006663057996081842	Db----------8	+	A2--------A----
252	1.8422104048882582	0.0006680263486673458	ACYS------A----	+	Vi-S---2--A----
251	1.8415424412948453	0.000671651883513538	ClYP1----------	+	PDIP1----------
250	1.8408708376846044	0.0006737952261212724	PDXP2----------	+	P8XP2----------
249	1.8401970424584837	0.000675828042312622	AANS4----1A----	+	AANS4----1N----
248	1.83952123372548	0.0006813782991247231	NNMP7-----A----	+	NNNP7-----A----
247	1.838839860253683	0.0006843894573027873	PJYS2----------	+	ClYS1----------
246	1.8381555528609441	0.000684891705694031	P4FP1----------	+	PJYS1----------
245	1.8374708976942877	0.0006868518431526905	VpQW---XR-AA--1	+	VpQW---XR-NA---
244	1.8367840555057897	0.0006901413089863513	Dg-------3A---1	+	PP-P3--2-------
243	1.8360939286787854	0.0007222387452714583	PLYS1----------	+	PZM-1----------
242	1.8353716947608403	0.0007363162661308837	VpMP---XR-AA--1	+	VpMP---XR-NA---
241	1.8346354074586735	0.0007482866835508244	PPFS3--3-------	+	PPXP4--3-------
240	1.8338871256024494	0.0007490545431533685	PDFS4----------	+	ClFS4----------
239	1.8331380903686048	0.0007794306878340573	PDFS2----------	+	P8FS2---------1
238	1.8323586596807706	0.0007957064345421871	Vi-P---1--A----	+	VB-S---1P-AA--1
237	1.8315630546201023	0.0007958116614359169	AAMP1----1A----	+	PDMP1----------
236	1.8307673105412485	0.0007967069246659275	PP-P4--1-------	+	P6-X4----------
235	1.8299706036165817	0.000815683209067216	NNNP3-----A----	+	VsMP---XX-AP---
234	1.8291549204075137	0.0008459443766842523	AAIS1----2A----	+	P8IS4----------
233	1.8283089856854844	0.0008560274060358279	AGFS1-----A----	+	VB-S---3F-NA---
232	1.8274530162073765	0.000858215991384019	ClYP4----------	+	ClXP4----------

231	1.826594814697975	0.0008745706997252065	AAMP3----1A----	+	PPZS4--3------2
230	1.8257202536529031	0.0008747484392689995	ClXP2----------	+	Cn-P2----------
229	1.8248455100409615	0.0008984978637497547	ACNS------A----	+	Dg-------3A----
228	1.8239470363138481	0.0009047052676888099	P4ZS6----------	+	P5FS6--3-------
227	1.8230423358734875	0.0009376958856789786	PDFS1----------	+	ClFS1----------
226	1.8221046930884097	0.0009395856245930072	PLXP3----------	+	AAIP3----1A----
225	1.8211651171184728	0.0009399222295559933	P5ZS7--3-------	+	ClXP7----------
224	1.8202251997162442	0.0009420309817949968	NNFPX-----A----	+	AGFS4-----A----
223	1.8192832218350492	0.000949123739551915	ClXP1----------	+	AAIP1----1N----
222	1.8183341367141153	0.0009605501728933812	C}----------	+	NNIXX-----A---8
221	1.8173738037709501	0.0009797979549441473	PHZS4--3-------	+	PP-P3--1-------
220	1.816394015470661	0.0009847515933470979	AAMS4----1A----	+	P8FP4---------1
219	1.8154093024959315	0.0010106818837558578	P4FS4----------	+	PQ--4----------
218	1.8143989150791415	0.0010218341487074562	Ca--1----------	+	Cn-S1----------
217	1.8133771388583624	0.0010254938807318552	VpNS---XR-NA---	+	VpTP---XR-NA---
216	1.812351649804958	0.0010293132535654764	P4FS1----------	+	P4YS1----------
215	1.8113228772120533	0.0010311278124459124	P8FS4---------1	+	PHZS3--3-------
214	1.8102917493996071	0.0010336034203751836	PDYS1----------	+	Db----------1
213	1.8092582183891417	0.0010583104362930865	NNMS3-----A----	+	NNMS3-----A---1
212	1.808199946571466	0.0010619383779946591	NNNP4-----A----	+	AUIS2M---------
211	1.807138008193472	0.001063984781625036	AAIS4----1A----	+	PDIS4----------
210	1.806074066857793	0.001072801171338838	PLXP2----------	+	ClXS2----------
209	1.8050012705137817	0.001072832771563807	P4MP1----------	+	P4FP1----------
208	1.8039288191010747	0.0010731373426196861	PLMP1----------	+	PZMP1----------
207	1.8028557541683647	0.0010787635438642366	NNFXX-----A---8	+	NNIPX-----A----
206	1.8017772706094843	0.001094441968927746	Vc-P---1-------	+	PPXP3--3-------
205	1.800682833467884	0.0011202560478317195	PDZS3----------	+	AAIS3----1A----
204	1.7995625919020342	0.0011212412327846998	VB-S---1P-NA---	+	VB-P---2P-AA---
203	1.798441427906487	0.0011285353329259324	PZ--1----------	+	PW--4----------
202	1.7973129118828715	0.001136201230159905	AANS1----1A----	+	CrNS1----------
201	1.7961770147743332	0.001141377497959199	P4NS1----------	+	PPYS1--3-------
200	1.795035868988084	0.0011836156345775258	ACYS------A----	+	PDZS7----------
199	1.7938523499000534	0.0012469232300987825	PDZS6----------	+	P4ZS6----------
198	1.792605436324608	0.0012692369427504523	VpTP---XR-AA--1	+	VpMP---XR-AA--1
197	1.7913362476551316	0.0012751656766783904	ClHP1----------	+	ClXP1----------
196	1.79006115921569	0.0012797108497816945	PJYS2----------	+	CrIS1----------
195	1.7887815642217637	0.001299186777384373	AAIP1----1A----	+	ClYP1----------
194	1.7874825319188552	0.0013002187815841348	VB-P---3P-AA--1	+	VB-P---3F-AA---
193	1.7861823565832171	0.0013139141866273258	Db----------8	+	NNFSX-----A----
192	1.7848685872164094	0.001320066027477639	Dg-------1N----	+	Dg-------2A---1
191	1.7835485742895303	0.0013246556825491848	ClYP4----------	+	AAIP4----1A----
190	1.7822239379162903	0.0013252502517210104	PLXP3----------	+	AAFP3----1A----
189	1.78089870214655	0.0010460719435235449	NNFP3-----A----	+	NNIP3-----A----
188	1.7798526350303547	0.0013324372520155045	PDFS2----------	+	PSFS2-P1-------
187	1.7785202122603212	0.0013414028288144635	P5ZS6--3-------	+	AANP6----1A----
186	1.777178809431507	0.0013638881527391622	AAIP7----1A----	+	P5ZS7--3-------
185	1.775814926106095	0.0013659443659754036	AAMS7----1A----	+	AANS7----1A----
184	1.774449020358738	0.0013872042331228053	NNMP1-----A---1	+	Vi-P---1--A----
183	1.773061965772761	0.001390571168445506	PP-P2--1-------	+	PDZS2----------
182	1.7716713994316426	0.001403124379193113	AAIS1----2A----	+	RV--4----------

181	1.7702682895344322	0.0014467858355483318	AAMP3----1A----	+	PLYS1----------
180	1.768821518180867	0.0014513955849143221	Cv----------	+	Dg-------3A---1
179	1.7673701998331892	0.0014809383145902803	AANS3----1A----	+	AAFS3----1A----
178	1.765889276000581	0.001511828276807764	RR--3----------	+	RV--3----------
177	1.7643775828889356	0.0015139468935023558	VsTP---XX-AP---	+	NNNP3-----A----
176	1.762863650477416	0.0015212928869562364	AAMS2----1A----	+	PSZS2-P1-------
175	1.7613423962090777	0.001530592385099394	PDFS1----------	+	CrFS1----------
174	1.7598119341618164	0.0015586772395284765	NNIP7-----A----	+	NNMP7-----A----
173	1.7582532617496134	0.0015694620992283206	NNIS6-----A----	+	NNIS6-----A---1
172	1.7566838286143471	0.001573075478895937	PHZS4--3-------	+	PPFS3--3-------
171	1.755110767617433	0.001587326136242303	VpYS---XR-NA---	+	VpYS---XR-AA--1
170	1.7535234897544638	0.00161970332339691	Vc-P---1-------	+	PH-S3--1-------
169	1.7519037960857207	0.001670324607780757	P4FS1----------	+	P4FS4----------
168	1.7502343066055668	0.001700734756141277	VsNS---XX-AP---	+	ACNS------A----
167	1.7485336442593356	0.001712595941120514	ClXP2----------	+	PDXP2----------
166	1.7468210531455417	0.001767768322490133	NNFPX-----A----	+	AAXXX----1A----
165	1.745053381369598	0.0017844278095804286	VpQW---XR-AA---	+	VpQW---XR-AA--1
164	1.7432691369984565	0.001816463354457891	AAFP4----1A----	+	AAMS4----1A----
163	1.741452712262618	0.0018440271390413546	PSXXXZS3-------	+	PSXXXXP3-------
162	1.7396087575334864	0.0018454670001829092	NNNS3-----A----	+	NNFS3-----A----
161	1.7377633194972661	0.0019013589218913381	AGFS1-----A----	+	AANS1----1N----
160	1.735862018503303	0.0019218920235430405	VB-P---3P-NA---	+	VB-S---1P-NA---
159	1.7339402375082886	0.001942510701752909	PP-P4--1-------	+	AAMP4----1A----
158	1.7319977461158462	0.0019519825673428096	ClHP1----------	+	AANP1----1A----
157	1.7300459083683226	0.0019576314666495433	Cn-S4----------	+	Ca--1----------
156	1.7280883396569275	0.001967919813830047	AAFP6----1A----	+	P5ZS6--3-------
155	1.726120419843097	0.0015898220432833052	NNFP6-----A----	+	NNNP6-----A----
154	1.7245306074544668	0.0019725095409741655	J,-X---3-------	+	P4NS1----------
153	1.722558643401481	0.0019859180857281743	PJYS2----------	+	PDYS1----------
152	1.720572913581518	0.002047568061680872	NNNXX-----A----	+	NNIXX-----A----
151	1.7185255530949115	0.0020696853912790594	AAFS4----1A----	+	PDFS4----------
150	1.7164559932141432	0.0020869781555440167	NNMS4-----A----	+	NNNP4-----A----
149	1.7143690150585988	0.002135608599911399	C}----------	+	NNFXX-----A---8
148	1.7122339036734018	0.0021602966144427943	AAMP1----1A----	+	PLMP1----------
147	1.7100737470514502	0.0021620065678038	P8FS4---------1	+	PZ--1----------
146	1.7079117597929552	0.0021784030789456336	NNFP3-----A----	+	NNMP3-----A----
145	1.7057333760233182	0.002202876034087864	AAFP7----1A----	+	AAIP7----1A----
144	1.7035305096438866	0.0019423999959907907	NNIP7-----A----	+	NNFP7-----A----
143	1.7015881241298774	0.0022512339406305718	PLXP2----------	+	ClXP2----------
142	1.6993368998439016	0.0022836090714663153	Dg-------2A----	+	Dg-------1N----
141	1.6970534018009635	0.0023039732061943125	VpTP---XR-AA--1	+	VpNS---XR-NA---
140	1.6947494816953705	0.0023768944767737846	RR--7----------	+	RV--7----------
139	1.6923727899663463	0.002501520356456651	VsQW---XX-AP---	+	AAIS1----2A----
138	1.6898712937465268	0.002505946452544502	AAMP3----1A----	+	P7-X3----------
137	1.6873653714306198	0.0025424432531505024	ACYS------A----	+	Cv----------
136	1.684823101961253	0.0025589441835573955	C}----------	+	Db----------8
135	1.6822647998122289	0.0026051455107124477	AAIS6----1A----	+	PDZS6----------
134	1.6796596880928087	0.002634067650987157	PHZS4--3-------	+	Vc-P---1-------
133	1.6770256445784588	0.002591552896038553	J,-X---3-------	+	P4MP1----------
132	1.6744350185292656	0.0026905801178533084	AAFP6----1A----	+	AAIP6----1A----

131	1.67174444323874	0.002214740981192051	NNFP6-----A----	+	NNIP6-----A----
130	1.6695297360488393	0.00272420042707219 3	VpTP---XR-AA---	+	VB-P---3P-AA--1
129	1.6668056659596044	0.00273964371298239 2	VB-P---1P-AA---	+	NNMP1-----A---1
128	1.6640663070589352	0.0027566540628638572	AAFS1----1A----	+	PDFS1----------
127	1.6613103964044793	0.0027765359990247138	VB-S---3P-NA---	+	VB-S---3F-AA---
126	1.6585340293619135	0.002777775068768676	VB-P---3P-NA---	+	VpYS---XR-NA---
125	1.6557564135949474	0.00275079162502766 9	VsYS---XX-AP---	+	AGFS1-----A----
124	1.6530056992071551	0.0028643162464957947	PDFS2----------	+	AANP2----1A----
123	1.6501414070972962	0.00295373834458882 9	PDZS3----------	+	NNIS3-----A----
122	1.647187808745198	0.00297285644162105 5	NNMP4-----A----	+	NNMS4-----A----
121	1.644214966785561	0.00302618872293944 7	NNFPX-----A----	+	NNMS3-----A----
120	1.6411889132277875	0.00303480850732486 6	TT----------	+	VsNS---XX-AP---
119	1.638154423324065	0.00307118669678493 9	RR--6----------	+	RV--6----------
118	1.6350845158690217	0.00307879111842202 5	AAMS7----1A----	+	PLXP3----------
117	1.6320057778511992	0.00309382510770561 5	PDNS4----------	+	PP-P4--1-------
116	1.6289119768801308	0.00327665641184676 8	P4FS1----------	+	J,-X---3-------
115	1.6256370824427564	0.00333406724668742 53	Xx----------	+	XX----------
114	1.6223035896480194	0.00340565839366249 73	RR--2----------	+	RV--2----------
113	1.61889862638949	0.00343610007358844 53	NNMS2-----A----	+	AAMS2----1A----
112	1.6154628111282139	0.00345776844768469 83	PJYS2----------	+	PDNS1----------
111	1.6120053902480964	0.0035032700036596 29	P8FS4---------1	+	Dg-------2A----
110	1.6085022505822741	0.0036422692022053 29	NNMS7-----A----	+	NNNS7-----A----
109	1.6048600924085972	0.00369790345629373 43	NNNS3-----A----	+	AANS3----1A----
108	1.6011622323982486	0.00371392795330999 57	VsTP---XX-AP---	+	VpTP---XR-AA--1
107	1.5974483720275199	0.00372832201639166 15	NNFP1-----A----	+	NNNP1-----A----
106	1.5937206099810979	0.0025919170386081 99	ClHP1----------	+	AAFP1----1A----
105	1.5911290453373859	0.00364327837426598 6	VpTP---XR-AA---	+	VB-P---3P-AA---
104	1.5874860566027618	0.00380437825361640 97	PHZS4--3-------	+	AAMP3----1A----
103	1.5836817266224186	0.0039138852939277 25	PLXP2----------	+	AAMP2----1A----
102	1.5797678558104726	0.00395664389701422 3	VB-S---3P-NA---	+	VpNS---XR-AA---
101	1.5758114870711148	0.00406213646411825 5	ACYS------A----	+	AAIS7----1A----
100	1.5717495823187078	0.00406718039300030 25	AAFP4----1A----	+	AANS4----1A----
99	1.567682459853636	0.0028891377416565	NNFP4-----A----	+	NNNS4-----A----
98	1.5647933655579251	0.00401620300783761 5	VB-P---3P-NA---	+	VB-S---1P-AA---
97	1.5607774473623985	0.00449125635686243 8	VsYS---XX-AP---	+	VsQW---XX-AP---
96	1.5562862923794107	0.00456886380847395 7	X@----------	+	Xx----------
95	1.5517192533006667	0.00465870679972235 6	C}----------	+	NNNXX-----A----
94	1.5470613961105524	0.00474189061575032 4	VB-P---1P-AA---	+	VpMP---XR-AA---
93	1.5423199399542593	0.00476864312739796 7	AAIP1----1A----	+	AAMP1----1A----
92	1.5375515912938285	0.00339770362610428 5	NNIP1-----A----	+	NNMP1-----A----
91	1.5341542931632204	0.00483379303036771 8	NNMS1-----A----	+	AAMS1----1A----
90	1.5293225662289456	0.0048478941219204 1	AANS2----1A----	+	PP-P2--1-------
89	1.5244747155529716	0.00486398139613396 6	NNFP3-----A----	+	PDZS3----------
88	1.5196108934586385	0.00486774931776950 3	NNIP7-----A----	+	AAFP7----1A----
87	1.5147431682775059	0.00492817769989464 8	AAIS6----1A----	+	AANS6----1A----
86	1.5098150388508838	0.00275856468356919 24	NNIS6-----A----	+	NNNS6-----A----
85	1.5070565079586065	0.00496016368812022 75	PSXXXZS3-------	+	PDFS2----------
84	1.5020964408170314	0.0048685857351567 88	NNNP2-----A----	+	NNMP2-----A----
83	1.4972279612830752	0.00453519735190924 95	PLXP2----------	+	Cn-S4----------
82	1.4926928411684028	0.0049329661401238 75	NNMP4-----A----	+	NNIP4-----A----

81	1.487759899164915	0.004377992908588431	PDNS4----------	+	ClYP4----------
80	1.4833819497022727	0.004834031407856657	P8FS4---------1	+	TT----------
79	1.478548367235858	0.005142442273384715	P7-X4----------	+	PHZS4--3-------
78	1.4734060890915999	0.005279975826538379	AAMS7----1A----	+	AAFS7----1A----
77	1.468126200156955	0.0039822512596741355	NNMS7-----A----	+	NNFS7-----A----
76	1.4641441323357194	0.005479137346465509	NNFXX-----A----	+	NNFPX-----A----
75	1.4586652508376023	0.005582160517104216	VB-S---3P-AA---	+	VB-S---3P-NA---
74	1.453084277843022	0.005641604707049941	Db----------	+	P8FS4---------1
73	1.4474449130158473	0.005679612212897391	VsTP---XX-AP---	+	VB-P---1P-AA---
72	1.4417658028449891	0.005617979518555263	NNFP3-----A----	+	NNNS3-----A----
71	1.4361480260741837	0.0059602419769104464	P7-X4----------	+	Vc-X---3-------
70	1.430187957881057	0.006373298425912818	VpTP---XR-AA---	+	VsTP---XX-AP---
69	1.4238154511368277	0.006478915371039826	VsYS---XX-AP---	+	Vf--------A----
68	1.4173368205780994	0.0065983895034497225	VB-S---3P-AA---	+	VpQW---XR-AA---
67	1.410739802035609	0.006572977322223193	C}----------	+	X@----------
66	1.4041694990527207	0.006574857923346951	NNMS7-----A----	+	NNIS7-----A----
65	1.3975949452509944	0.004749225629979065	AAMS7----1A----	+	ACYS------A----
64	1.392846038224618	0.006561399007268334	Db----------	+	Dg-------1A----
63	1.3862872942473732	0.006317383973007275	AAFP4----1A----	+	PDNS4----------
62	1.3799700116482398	0.0034676969871722044	NNMP4-----A----	+	NNFP4-----A----
61	1.3765023822436522	0.006555442039224174	J,----------	+	P4FS1----------
60	1.3699507006924105	0.00660646496252415	NNFP2-----A----	+	NNNP2-----A----
59	1.3633443805497045	0.00432804424689464	AAFP2----1A----	+	PLXP2----------
58	1.3590164714679758	0.006801263252696421	Db----------	+	PJYS2----------
57	1.3522182108128689	0.006907185073768179	VB-P---3P-NA---	+	VpYS---XR-AA---
56	1.3453119863772376	0.0071140330574757035	C}----------	+	NNFXX-----A----
55	1.3382008835074455	0.007585931000244013	AAFP2----1A----	+	AAIP2----1A----
54	1.3306114786729673	0.0035063468861256386	NNFP2-----A----	+	NNIP2-----A----
53	1.3271053007433011	0.007744754262302998	AAFS4----1A----	+	AAIS4----1A----
52	1.3193607154374545	0.003088971308215338	NNFS4-----A----	+	NNIS4-----A----
51	1.316271917913026	0.00788255394192431	NNNS2-----A----	+	AANS2----1A----
50	1.3083894556903222	0.007353092035444861	NNMS2-----A----	+	NNNS2-----A----
49	1.3010367401864056	0.008160692449044132	NNFP6-----A----	+	AAFP6----1A----
48	1.292876086355979	0.008221567306238305	AAIP1----1A----	+	ClHP1----------
47	1.284655165911599	0.0039184694435178274	NNIP1-----A----	+	NNFP1-----A----
46	1.280737661933544	0.009856989354393185	AAIS1----1A----	+	AANS1----1A----
45	1.2708813966782482	0.00424423387219048	NNNS1-----A----	+	NNIS1-----A----
44	1.266639151664915	0.010674180421870197	AAFP2----1A----	+	PSXXXZS3-------
43	1.2559652560553574	0.011071535438487545	AAFS6----1A----	+	AAIS6----1A----
42	1.2448937785447987	0.0027516638743176614	NNFS6-----A----	+	NNIS6-----A----
41	1.2421422015623724	0.011232567874886842	NNMS7-----A----	+	NNIP7-----A----
40	1.2309099619457433	0.010361341297001232	RR--7----------	+	AAMS7----1A----
39	1.2205491420000936	0.011205276778184337	VB-S---3P-AA---	+	VB-P---3P-NA---
38	1.2093461968210073	0.013172739602339348	NNMS2-----A----	+	NNIS2-----A----
37	1.196173949606055	0.009934770511449822	NNMS2-----A----	+	AAIS2----1A----
36	1.1862397535465552	0.01305867887286033	RR--3----------	+	NNFP3-----A----
35	1.1731814125866067	0.013473159399134307	NNFS4-----A----	+	NNMP4-----A----
34	1.1597084945538347	0.005013963234175009	AAFS4----1A----	+	AAFP4----1A----
33	1.154694801649986	0.013701871078212038	C}----------	+	C=----------
32	1.1409985399261213	0.015163985405903593	NNNS1-----A----	+	NNFS1-----A----

31	1.125837977095291	0.0031301165714826412	AAFS1----1A----	+	AAIS1----1A----
30	1.1227093280313114	0.016689674201070315	AAIP1----1A----	+	NNIP1-----A----
29	1.106021266157567	0.017122269215961133	RR--4----------	+	AAFS4----1A----
28	1.0888996969040683	0.017423083927356184	VB-S---3P-AA---	+	VpTP---XR-AA---
27	1.0714797362574875	0.01732993713964559	AAFS6----1A----	+	NNFP6-----A----
26	1.0541498956643913	0.0185454834291316	NNMS1-----A----	+	AAIP1----1A----
25	1.0356080906586789	0.018378450571722618	NNFS2-----A----	+	NNFP2-----A----
24	1.0172298331800482	0.00807538351411527	AAFS2----1A----	+	AAFP2----1A----
23	1.0091547393055713	0.018427433329175416	Db----------	+	P7-X4----------
22	0.9907304823577782	0.019701372888266885	RR--3----------	+	NNMS7-----A----
21	0.971029775640681	0.021166201390149054	RR--7----------	+	RR--3----------
20	0.949864761773054	0.02126106129467771	Db----------	+	VsYS---XX-AP---
19	0.9286071616720639	0.02316635580540974	NNFS6-----A----	+	AAFS6----1A----
18	0.9054409893050921	0.024389112223965365	VB-S---3P-AA---	+	Db----------
17	0.8810584615555896	0.025786423823072904	J,----------	+	J^----------
16	0.8552772657280044	0.026865311092945862	AAFS2----1A----	+	RR--2----------
15	0.8284129394098321	0.023220211658212225	NNMS2-----A----	+	NNFS2-----A----
14	0.805193495296664	0.030009635728538007	NNNS1-----A----	+	AAFS1----1A----
13	0.7751887496507023	0.026788569710309698	NNMS1-----A----	+	NNNS1-----A----
12	0.7484087484463875	0.036093691027191545	J,----------	+	VB-S---3P-AA---
11	0.7123268698891522	0.04436521561449734	C}----------	+	NNMS1-----A----
10	0.6679758321349969	0.04498149285164461	NNFS4-----A----	+	RR--4----------
9	0.6229952806121796	0.04277232774602822	RR--7----------	+	NNFS4-----A----
8	0.580225081717501	0.049065326994688685	NNMS2-----A----	+	AAFS2----1A----
7	0.5311615070426297	0.06376194265107979	C}----------	+	Z:----------
6	0.46743563059583937	0.06897859430291521	J,----------	+	RR--7----------
5	0.3984674554186047	0.07441294404895171	C}----------	+	NNMS2-----A----
4	0.32409057787807316	0.09521848873970688	RR--6----------	+	NNFS6-----A----
3	0.2288734096393519	0.0998901939551032	C}----------	+	J,----------
2	0.1290192825559398	0.12232466676342124	C}----------	+	RR--6----------

==========

Text TEXTEN1.ptg

==========

History of merges for English tags for full data.

#classes	MI	Loss	Merged tags		
36	0.8833408451808273	0.00021893989181066603	RBR	+	WP$
35	0.8831219052890166	0.00034889006874716695	JJR	+	RBR
34	0.8827730219170419	0.0008195430637212581	SYM	+	NNPS
33	0.8819536596661836	0.0010847585942971028	PRP	+	EX
32	0.8808689496234892	0.0012250612982291423	NNP	+	FW
31	0.8796504126724939	0.0013178714035305544	.	+	(
30	0.8783334686290095	0.0013896729306041022	WP	+	“
29	0.8769438006224411	0.0017370170113054084	JJ	+	JJS
28	0.8752068902985749	0.002107615755988947	JJR	+	RBS
27	0.8730992843906575	0.0023104738200171553	WRB	+	WP
26	0.8707888237014015	0.00265398947475487	DT	+	PRP$

25	0.8681349130112173	0.0032881296488037615	JJ	+	CD
24	0.8648468998979237	0.0036557257836587545	NNP	+	SYM
23	0.8611976985322322	0.006011978838287406	WRB	+	WDT
22	0.8551857646899309	0.005948844014717597	,	+	:
21	0.8492386575202724	0.006199486842402527	VBD	+	VBN
20	0.8430393191646239	0.006975957392948817	VBZ	+	VBP
19	0.8360638492281897	0.008532429272879588	JJ	+	JJR
18	0.8275315489438029	0.009149970933491347	NN	+	NNS
17	0.8183823354427422	0.011352485408837536	WRB	+	VBG
16	0.8070299370261439	0.015370842495556702	.	+	,
15	0.7916617102972389	0.019135393939358022	WRB	+	CC
14	0.7725267663177406	0.026204319239137908	TO	+	MD
13	0.7463225700676309	0.030055641894816423	NNP	+	PRP
12	0.7162734530191119	0.030873869512718803	VBD	+	RB
11	0.6854001662082907	0.032931546022582275	VB	+	VBZ
10	0.6524705811769713	0.03689673674282974	WRB	+	NNP
9	0.6155803694726563	0.039308770298801005	WRB	+	TO
8	0.5762781242452113	0.04047371848518348	VBD	+	VB
7	0.5358068702035981	0.053964971034472696	WRB	+	VBD
6	0.4818484242927004	0.08059765316038091	WRB	+	.
5	0.4012572962665058	0.0799165766797299	JJ	+	DT
4	0.32134192409737555	0.0984602921047264	WRB	+	IN
3	0.2228881571352464	0.19134191363319272	WRB	+	JJ
2	0.03155276865176232	0.03155929380487748	WRB	+	NN

1) Class . + (

The full stop and opening bracket share similar distribution, namely what is following a full stop (the beginning of the next sentence) probably can appear after opening bracket (independent sentence within brackets). Thus, in theory these sentences can be indistinguishable.

The interesting thing is that finally most of the punctuation marks fall into one cluster except for quotation mark which quite early clusters with relative pronoun group WP. Quotation mark shares more distributional properties with these pronouns because both tend to appear in the beginning of the sentence. Other punctuation marks cluster together according to their relatedness to structural formation of the sentence.

2) Class VBD + VBN vs. VBZ + VBP

Surprisingly, the earliest merge for verbs was VBD (finite verb in past tense) + VBN (participle). Thus, the expected earliest merge VBZ (finite verb for 3rd person in present tense) + VBP (finite verb in present tense) which relies mostly on the subject of action was outweighed by the adverbs or adverbial phrases occurring near these verbs. This guess is due to the different grammatical usages of VBD and VBN, so they don`t share neither subject for VBD nor auxiliary verb for VBN. The only sufficient source of shared co-occurrences is adverbs of time which are quite frequent so do not allow clusterization of VBD with VBZ + VBP earlier and draw the line between verbal present and past tenses.

3) Class MD + TO

The clustering of MD and TO together occurs due to their common position before a main verb. However, their merge takes place quite late in the hierarchy. The main reason for that could be that the prepositional meaning of “to” was interfered.