The relation between dependency distance and frequency

Xinying Chen
University of Ostrava, Czech Republic
Xi'an Jiaotong University, China
xy@yuyanxue.net

Kim Gerdes
LPP (CNRS)
Sorbonne Nouvelle, France
kim@gerdes.fr

Abstract

This present pilot study investigates the relationship between dependency distance and frequency based on the analysis of an English dependency treebank. The preliminary result shows that there is a non-linear relation between dependency distance and frequency. This relation between them can be further formalized as a power law function which can be used to predict the distribution of dependency distance in a treebank.

1 Introduction

As a well-discussed norm (Hudson 1995; Temperley 2007; Futrell et al. 2015; Liu et al. 2017), dependency distance shows several attractive features for quantitative studies. First, its definition is rather clear. It is the linear distance between a word and its head.[1] Second, it is very easy to quantify. We can simply compute dependency distance as the difference of the word ID and its head's ID in a CoNLL style treebank (Buchholz & Marsi 2006). These features together with the emergence of large-scale dependency treebanks made dependency distance one of the popular topics in quantitative syntactic studies.

Among various interesting discussions, the most striking finding is probably the dependency distance minimization phenomena. After empirically examining the dependency distance distributions of different human languages and comparing the results with different random baselines, Liu (2008, 2010) found that there is a universal trend of minimizing the dependency distance in human languages. Futrell et al. (2015) conducted a similar study which widened the language range and added one more random baseline. Their results are coherent with Liu's finding. Both Liu (2008) and Futrell et al. (2015) connect this phenomenon with the short-term memory (or working memory) storage of human beings and the least effort principle (Zipf 1949). Since long dependencies, which have longer distance, occupy more short-term memory storage, they are more difficult or inefficient to process. Therefore, for lowering the processing difficulty and boosting the efficiency of communications, short dependencies are preferable according to the least effort principle.

Initially, the least effort principle was brought up by Zipf for explaining the observed power-law distributions of word frequencies. Later on, similar power-law frequency distributions have been repeatedly observed in various linguistic units, such as letters, phoneme, word length, and etc. (Altmann & Gerlach 2016). The power law distribution, therefore, has been considered as a universal linguistic law. After investigating the relationships between different word features (such as length vs frequency, frequency vs polysemy, and etc.), people found out an interesting phenomenon. The relations between two highly correlated word features are usually non-linear and can be formulated as a power law function (Köhler 2002). Kohler (1993) further proposed a word synergetic framework to model the interactions between different word features. This model has proved quite successful also then adapted to syntax features. The first studies mainly focused on the analysis of phrase structure treebanks (Köhler 2012), which naturally are limited in language types since phrase structure grammar is less suitable for describing free word order languages (Mel'čuk 1988). As the dependency treebanks are getting dominant, studies based on dependency grammar start to take lead. We can find recent studies discussing the relations between sentence lengths, tree heights, tree widths, and mean dependency distances (Jing & Liu 2017, Zhang & Liu 2018, Jiang & Liu 2015).

[1] Hudson's original measures takes two adjacent words to have distance zero. We prefer the alternative definition where $x=y \Leftrightarrow d(x,y)=0$, i.e. a word has distance zero with itself, making the measure a metric in the mathematical sense.

Knowing that short dependencies are preferable by languages due to the least effort principle and that syntax features behaviour similar to word features, we can easily draw our hypotheses:

- *The relation between dependency distance and frequency can be formulated as a non-linear function (probably also a power law function).*

Contrary to above-mentioned studies, our study here is not focusing on mean dependency distances but the distribution of the distance of every single dependency. In the dependency minimization studies or synergetic syntax studies, the observed feature is mean dependency distance per sentence. In a way, these observed dependency distances are treated as a dependent feature of dependency trees. This is a very reasonable choice since the dependency distance is defined as the linear distance between two words in the same sentence. In particular, when the studies discuss other tree-related features such as tree heights and widths, mean dependency distance is a more easily comparable feature than a group of individual dependency distances. However, we believe the value of individual dependency distances is neglected. Individual dependency distances (Liu 2010, Chen & Gerdes 2017, 2018) provide more details of the fluctuation than the average which would level-up differences of dependencies in a sentence and it should be given the same attention as the mean dependency distance. Therefore, our study here is trying to pick up the missing detail of previous studies by investigating the relations between individual dependency distances and their frequencies.

The paper is structured as follows. Section 2 describes the data set, the Parallel Universal Dependencies (PUD) English treebank of Universal Dependencies treebanks, and introduces our computing method for dependency distance and frequency. Section 3 presents the empirical results and discussions. Finally, Section 4 presents our conclusions.

2 Material and Methods

Universal Dependencies (UD) is a project of developing a cross-linguistically consistent treebank annotation scheme for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of Stanford dependencies (De Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages while allowing language-specific extensions when necessary. UD is also an open resource which allows for easy replication and validation of the experiments (all Treebank data on its page is fully open and accessible to everyone). For the present paper, we used the PUD English Treebank from the UD 2.3 dataset for our study since English is a rather reasonable choice for a pilot study. Furthermore, PUD is a parallel treebank with a wide range of languages, namely Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish. This makes PUD a good choice for future studies which would further test whether our finding here can be generalized into different human languages. We use the Surface-syntactic UD version of the treebank (Gerdes et al. 2018), which is more suitable for studies in distributional dependency syntax as it corrects the artificially long dependency distances of UD into a more standard syntactic analysis based on distributional criteria (Osborne & Gerdes 2019).

We first compute the dependency distance for every single dependency in the treebank except the root relation. The dependency distance is computed as the absolute difference between the word ID and its head's word ID. For instance, in Figure 1, there are 4 dependencies. We would take three of them into account except the root dependency. The dependency distances of these three dependencies are: *abs (1-2) =1* (for *subj*), *abs (4-2) =2* (for *comp*), and *abs (3-4) =1* (for *det*).

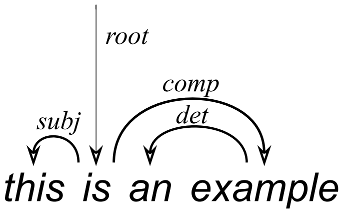


Figure 1: Example dependency tree in SUD analysis.

After computing all the dependency distances of the treebank, we then count the frequencies of each dependency distance, i.e. we count how many dependencies with dependency distance 1 occurred in the treebank, how many dependencies with distance 2 occurred, and so on. We then try to formulate the relation into a non-linear function. We will test different non-linear functions to see which one can predict the empirical data best. In other words, we try to see whether our data can be fitted by the power law function. This result can then either confirm or reject our hypothesis.

We also introduce two random baselines to see whether we can observe similar phenomenon in random dependency trees. Based on the PUD English treebank, we generate two random tree-banks. For the random treebank RT, we just randomly reorder the words of each sentence. For the random treebank PRT, we randomly reorder the words in a way that keeps the sentence’s dependencies projective (non-crossing).

3 Results and Discussion

The PUD English treebank is part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on multilingual parsing (Zeman et al. 2017). There are 1000 sentences in each language. The sentences are taken from the news domain and from Wikipedia. The PUD English treebank contains 21,176 tokens. See Appendix for the frequencies of dependency distances in the treebank.

The scatter plot Figure 2 shows that the relationship between dependency distance and frequency is indeed non-linear.

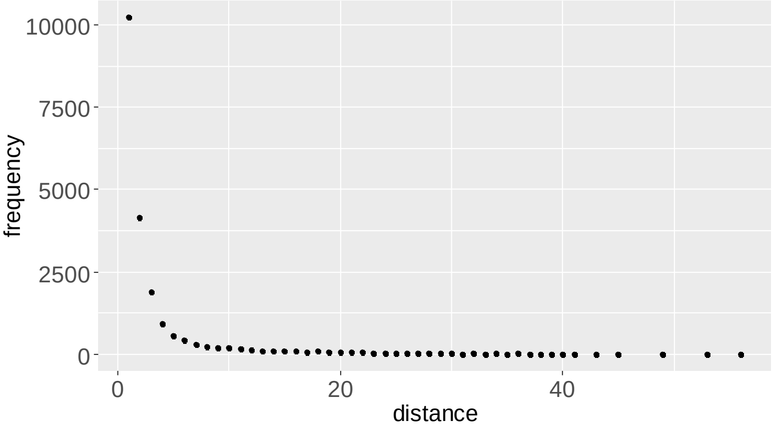


Figure 2: Scatter plot of dependency distance and frequency of PUD English treebank.

Since the observed data points scatter as a L-ish shape, we tried to fit the data to four non-linear functions, namely quadratic, exponent, logarithm, and power law functions. Although there are different ways of measuring the goodness-of-fit (Mačutek & Wimmer 2013), we choose to use the most common Pearson chi-square goodness-of-fit test to evaluate the fitting results in this study. The formula of the test is defined as

$$R^2 = \sum_{i=1}^{n} \frac{(f_i - NP_i)^2}{NP_i} \tag{1}$$

with f_i being the observed frequency of the value i, P_i being the expected probability of the value i, n being the number of different data values and N being the sample size. The obtained results of R-squared is presented in Table 2.[2]

Non-linear Model	Function	R^2
Quadratic	$y=2963.44-206x+3.1x^2$	0.34
Exponent	$log(y)=7.11-0.16x$	0.92
Logarithm	$y=4100.8-1262log(x)$	0.49
Power Law	$log(y)=10.71-2.56log(x)$	0.91

Table 1: R^2 of four non-linear models.

The results show that the observed data can indeed be formulated as a power law function. However, it seems that the data also fits an exponent regression very well. This is a very common issue in quantitative linguistic studies (Baixeries et al. 2013). In many situations, both exponent and power-law models can describe the data fluctuation reasonably well. One way to decide which model is better is by adding more observations from other languages. However, this is out of the scope of this pilot study. Another solution can be introducing baselines for comparison, which is our choice in this paper. By comparing the results in Table 1 with the results of two different random treebanks, we try to deliver the answer for this question, which model is better to represent the relation between dependency distance and frequency, exponent or power law?

For the two random English PUD treebank variations, RT and PRT, we replicate the same computation for the frequency and dependency distance, see Appendix. The scatter plots Figure 3 and 4 show that the relations between dependency distance and frequency in RT and PRT are both non-linear.

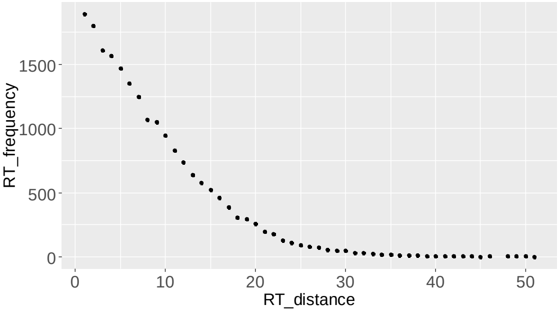


Figure 3: Scatter plot of RT.

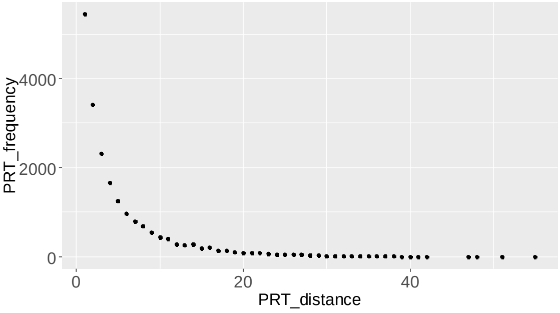


Figure 4: Scatter plot of PRT.

Similarly, we fit the data points to four non-linear models, see Tables 2 and 3 for results. We can see from Table 2 that RT fits to all non-linear models very well except to the power law function, which is very different from the PUD English treebank who fits to power law very well but does not fit to quadratic and exponent models. When we add the projectivity restriction, the fitting results of PRT seems more ‘human language’ like. Similar to the results of PUD in Table 1, PRT fits to exponent and power-law models better. However, the power law fitting result of PUD is clearly more satisfying than the result of PRT.

Non-linear Model	Function	R^2
Quadratic	$y=1883.88-106.28x+1.43x^2$	0.98
Exponent	$log(y)=8.42-0.17x$	0.98
Logarithm	$y=2220.88-611.66log(x)$	0.96
Power Law	$log(y)=11.23-2.37log(x)$	0.74

Table 2: R^2 results of RT.

[2] All parameter values in the models were obtained by *R* software. The same below.

Non-linear Model	Function	R^2
Quadratic	*$y=2551.07-168.63x+2.49x^2$*	0.62
Exponent	*$log(y)=7.99-0.17x$*	0.97
Logarithm	*$y=3258.25-972.05log(x)$*	0.75
Power Law	*$log(y)=11.28-2.55log(x)$*	0.84

Table 3: R^2 results of PRT.

Beyond considering the projectivity feature of dependency trees that deals with the crossing problem, we would also like to have a closer look at the role of syntax in this question. Our way of addressing this is to exclude less syntactic dependencies from the analysis. The UD/SUD annotation scheme includes predefined dependency structures for some constructions, in particular for MWE and punctuation. The distance of relations such as *fixed*, *compound*, *flat*, and *punct* are not based on distributional criteria of the tokens involved. Therefore, we also tested the results when these dependencies are excluded from our analysis, taking into account only syntactic dependencies (*subj*, *aux*, *cop*, *case*, *mark*, *cc*, *dislocated*, *vocative*, *expl*, *discourse*, *det*, *clf*). See the Appendix for the details. We first tested these three data sets with a linear regression model, and the results are similar to the previous analysis (PUD R2=0.21, RT R2=0.77, PRT R2=0.34). We then repeated the same non-linear regression analysis on these three selected data sets and the results are presented in Table 4.

Syntactic Data Set	Non-linear Model	Function	R^2
PUD English	Quadratic	*y=1216.36-148.07x+3.82x2*	0.44
	Exponent	*log(y)=5.84-0.25x*	0.81
	Logarithm	*y=1380.2-523.3log(x)*	0.56
	Power Law	*log(y)=8.45-2.53log(x)*	0.97
RT	Quadratic	*y=434.78-25.62x+0.36x2*	0.98
	Exponent	*log(y)=6.78-0.16x*	0.97
	Logarithm	*y=510.91-142.85log(x)*	0.95
	Power Law	*log(y)=9.1-2.07log(x)*	0.74
PRT	Quadratic	*y=656.17-50.02x+0.86x2*	0.6
	Exponent	*log(y)=6.27-0.16x*	0.95
	Logarithm	*y=810.18-251.99log(x)*	0.73
	Power Law	*log(y)=8.89-2.13log(x)*	0.89

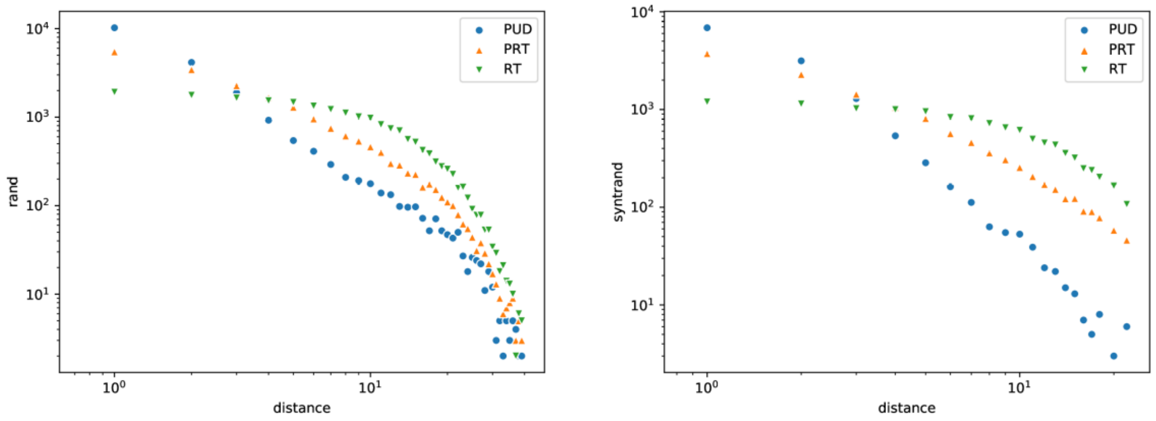
Table 4: R2 results for syntactic dependencies.

Very similar to the results of the previous analysis, PRT is closer to the PUD English results. However, the results with syntactic dependencies demonstrate more clearly that a power law model is the better choice for representing the relation between dependency distance and frequency. First, the original PUD data fits to the power law function best, whereas in the previous analysis we could not easily draw such a conclusion due to the very similar R2 values for both power law and exponent models. Secondly, the goodness of the power law model fitting somehow can distinguish the natural PUD data from random baselines.

4 Conclusion

Our results are coherent with our hypothesis that there is indeed a non-linear relation between dependency distance and frequency. Furthermore, this relation can be formulated as a power law function.

However, the results in Table 1 show that the power-law model is not the only candidate for formulating the relation, and we could also apply an exponential model to it. For figuring out which model is better for representing the relation, we introduce two random baselines. By randomly reordering the words in a sentence, while preserving the words' dependencies, we generate random treebanks: PRT with and RT without the projectivity restriction, in which PRT possesses a more 'natural' structure reproducing more closely the rarity of non-projective relations. We replicate the same analysis on these two random treebanks and compare the results with the PUD results. We found that we can distinguish the PUD from RT and PRT by looking at the results of power-law fitting. Therefore, we would like to cautiously draw our conclusion here that the power law model is probably a better choice for representing the relation between dependency distance and frequency, a hypothesis that is further strengthened by the results on purely syntactic dependency relations.



5a: All functions.

5b: Syntactic functions only.

Figure 5: Joint plot of the frequency of dependency distance on a logarithmic scale showing the greater linearity of PUD compared to the random treebanks.

Another interesting phenomenon we can observe from our data is that the projective random data-set has almost as good a fit to a power law function as the syntactically parsed true treebank. Although we need more samples to conduct a statistical significance testing for the difference, it seems that if we compare the natural PUD and the control PRT on the most relevant “syntactic functions only”, for example in the logarithmic presentation of Figure 5b., there is practically no difference between the linearity of PRT and PUD. This shows that projectivity has a major role as the responsible factor for the power-law function of dependency distance. Of course, our conclusion based on this pilot study needs to be tested with more languages in the future. This leads to the open question to actually pinpoint the additional syntactic constraint of PUD, compared to random treebanks, that results in the power law distribution.

We believe the result presented here has several potential applications. We can use the power law model to predict the distribution of dependency distance in a treebank. Since natural language treebanks fit to power law model betters than random treebanks, we might even use it as an index for assessing the quality of parse results.

Acknowledgements

This work is supported by the European Union & Ministry of Education of the Czech Republic (No. CZ.02.2.69/0.0/0.0/16_027/0008472) and the National Social Science Fund of China (2018CYY031).

Reference

Altmann, Eduardo Gabriel, and Martin Gerlach. 2016. Statistical laws in linguistics. In *Creativity and Universality in Language* (pp.7-26). Springer, Cham.

Baixeries, Jaume, Brita Elvevåg, and Ramon Ferrer-i-Cancho. 2013. The evolution of the exponent of Zipf's law in language ontogeny. *PloS one*, 8(3): e53227.

Buchholz, Sabine and Erwin Marsi. 2006, June. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning* (pp. 149-164). Association for Computational Linguistics.

Chen, Xinying and Kim Gerdes. 2017. Classifying languages by dependency structure: Typologies of delexicalized universal dependency treebanks. In *Proceedings of the Fourth International Conference on Dependency Linguistics* (Depling 2017), Pisa, September. Linköping University Electronic Press.

Chen, Xinying and Kim Gerdes. 2018. How Do Universal Dependencies Distinguish Language Groups? In *Quantitative Analysis of Dependency Structures*, 72: 277-294.

De Marneffe, Marie-Catherine and Christopher D. Manning. 2008. The Stanford typed dependency representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

Futrell, Richard, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112 (33): 10336-10341.

Gerdes, Kim, et al. 2018, November. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*.

Hudson, Richard. 1995. Measuring syntactic difficulty. Draft of manuscript, available at *http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf.*

Jiang, Jingyang. and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency treebank. *Language Sciences*, 50: 93−104.

Jing, Yingqi and Haitao Liu. 2017. Dependency distance motifs in 21 Indoeuropean languages. In *Motifs in Language and Text* (pp.133-150).

Köhler, Reinhard. 1993. Synergetic linguistics. In *Contributions to Quantitative Linguistics*, pp.41-51. Springer, Dordrecht.

Köhler, Reinhard. 2002. Power law models in linguistics: Hungarian. *Glottometrics*, 5: 51-61.

Köhler, Reinhard. 2012. *Quantitative Syntax Analysis*, 65. Walter de Gruyter.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9 (2): 159-191.

Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120 (6): 1567-1578.

Liu, Haitao, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21: 171–193.

Mačutek, Jan and Wimmer, Gejza. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20 (3): 227-240.

Mel'čuk, Igor Aleksandrovic. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

Osborne, Tim and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a Journal of General Linguistics*, 4 (1): 17. 1-28.

Petrov, Slav, Dipon Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LRE*C.

Temperley David. 2007. Minimization of dependency length in written English. *Cognition*, 105: 300–333.

Zeman, Daniel. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of LREC*.

Zeman, Daniel, et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. *CoNLL 2017*.

Zhang, Hongxin and Haitao Liu. 2018. Interrelations among Dependency Tree Widths, Heights and Sentence Lengths. In *Quantitative Analysis of Dependency Structures*, 72: 31-52.

Zipf George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Appendix

The table shows the complete dependency distance frequency data from the SUD version of the English PUD treebank. The first three frequency columns take into account all dependency relations of the treebank. The last three frequency columns only count syntactic relations that correspond to actual head-daughter relations, which are the following relations in SUD: *appos*, *clf*, *comp*, *det*, *discourse*, *dislocated*, *expl*, *mod*, *subj*, *vocative*.

Distance	PUD_all	PRT_all	RT_all	PUD_syntactic	PRT_syntactic	RT_syntactic
1	10,236	5,473	1,912	6,866	3,742	1,194
2	4,157	3,438	1,768	3,148	2,285	1,140
3	1,887	2,270	1,646	1,295	1,434	1,021
4	924	1,662	1,532	538	1,050	997
5	544	1,292	1,468	285	809	951
6	412	955	1,335	162	566	829
7	292	747	1,222	112	459	807
8	209	613	1,113	63	360	719
9	192	536	1,009	55	306	650
10	177	462	975	53	255	613
11	139	400	824	39	206	497
12	133	299	741	24	171	454
13	98	287	701	22	152	433
14	96	233	561	15	122	356
15	97	225	521	13	123	319
16	72	162	422	7	91	248
17	52	175	386	5	90	238
18	71	152	312	8	78	203
19	52	124	276	0	66	168
20	47	110	258	3	58	165
21	43	100	226	1	51	147
22	50	79	159	6	46	107
23	27	62	162	1	35	91
24	18	55	122	0	36	66
25	26	44	91	1	21	53
26	24	31	78	1	16	50
27	22	38	78	0	20	48
28	11	29	53	1	15	35
29	18	22	53	0	9	32
30	12	17	34	0	10	20
31	3	13	29	0	7	15
32	5	9	18	0	2	14
33	2	6	21	0	3	9
34	5	7	14	0	3	6
35	3	8	13	0	7	5
36	5	9	10	0	6	4
37	4	3	2	0	0	2
38	2	5	6	0	1	4
39	2	3	5	0	1	3
40	1	2	2	0	2	1
41	1	3	1	0	2	1
42	0	0	3	0	0	2
43	1	3	2	0	2	2
44	0	3	2	0	1	1
45	1	1	3	0	1	1
46	0	0	2	0	0	1
47	0	3	2	0	2	1
48	0	2	0	0	1	0
49	1	0	2	0	0	1
50	0	1	0	0	0	0
53	1	0	1	0	0	0
55	0	1	0	0	0	0
56	1	1	0	0	1	0
57	0	1	0	0	0	0