

What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks

Sylvain Kahane, Chunxiao Yan, Marie-Amélie Botalla

▶ **To cite this version:**

Sylvain Kahane, Chunxiao Yan, Marie-Amélie Botalla. What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks. 4th international conference on Dependency Linguistics (Depling), Sep 2017, Pise, Italy. pp.73-82. hal-01675325

HAL Id: hal-01675325
https://hal.science/hal-01675325

Submitted on 4 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks

Sylvain Kahane
Modyco
Université Paris Nanterre
CNRS – France
sylvain@kahane.fr

Chunxiao Yan
Modyco
Université Paris Nanterre
CNRS – France
yanchunxiao@yahoo.fr

Marie-Amélie Botalla
Lattice
Université Sorbonne Nouvelle
CNRS – France
marie-amelie.botalla
@sorbonne-nouvelle.fr

Abstract

The aim of this paper is to study some characteristics of dependency flux, that is the set of dependencies linking a word on the left with a word on the right in a given position. Based on an exploration of the whole set of UD treebanks (12M word corpus), we show that what we have called the flux weight, which measures center embeddings, is less than 3 in 99.62 % of the inter-word positions and is bounded by 6, which could be due to short-term memory limitations.

1 Introduction

It is generally recognized that speaker performance is limited by several factors and especially by short-term memory. Yngve (1960) was one of the first to take these limitations into account in language modeling, on the grounds that "although all languages have a grammar based on constituent structure, the sentences actually used in the spoken language have a depth that does not exceed a certain number equal or nearly equal to the span of immediate memory (presently assumed to be 7 ± 2)." This 7 ± 2 bound refers to the famous paper by Miller (1956). Miller (1962) and Chomsky and Miller (1963) stated that center-embedded constructions are limited. Very few studies have been conducted, however, on limitations on the syntactic structure. Gibson (1998) stated that "memory cost is hypothesized to be quantified in terms of the number of syntactic categories that are necessary to complete the current input string as a grammatical sentence", as well as the length during which "a predicted category must be kept in memory before the prediction is satisfied". Muratu et al. (2001) verified on a 20K word corpus of Japanese that the number of words on the left of a position that can have a dependent on the right (which will be called the left span of flux here) was bounded by 10. Liu (2008), Liu et al. (2009), and Liu (2010) expressed Gibson's hypothesis in terms of dependency length and studied it on Chinese data and on treebanks of 20 different languages.

In this paper, we will study *dependency flux*, that is the set of dependencies linking a word on the left with a word on the right in a given inter-word position. The notion of dependency flux was introduced in Kahane (2001:67) and previously studied on corpora of written French (Jardonnet 2009) and spoken French (Botalla 2014). This new study (Yan 2017) was conducted on the whole series of dependency treebanks provided by the Universal Dependencies (UD) project (Nivre et al. 2016), comprising 12M words and 630K sentences distributed in 70 treebanks of 50 languages.[1] Several features of the flux were measured: size, left and right spans, weight and density. Weight, which measures center embeddings and nested constructions, has stable properties: it seems to be distributed quite similarly in each corpus and language, and it is less than 3 in the overwhelming majority of the inter-word positions (99.62 %) and it never exceeds 6.

Dependency flux and its main characteristics are defined in Section 2 and studied on the UD treebanks in Section 3. A closer look at weight is proposed in Section 4.

[1] Our experiments have been done on UD v2 available in May 2017.

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 73-82,
Pisa, Italy, September 18-20 2017

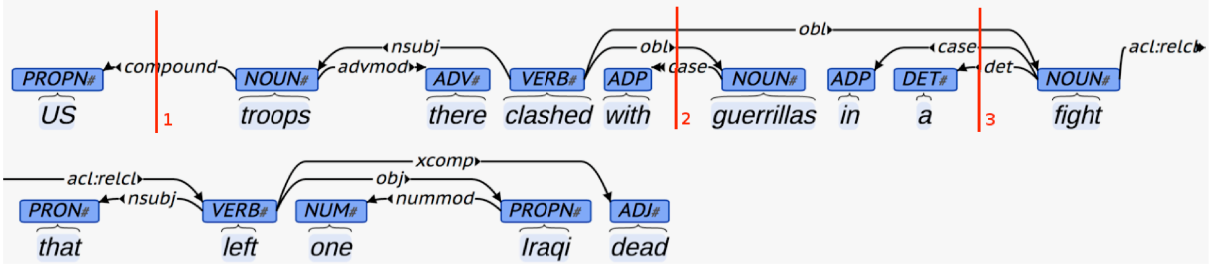


Figure 1. A UD dependency tree with three inter-word positions marked

2 Dependency flux and its characteristics

2.1 Definition and size

The *dependency flux* in a given inter-word position is the set of dependencies at this position, that is, linking a word on the left with a word on the right. In Fig. 1, the flux contains one dependency at position 1, three at positions 2 and 3.

The *size* of the flux is the number of dependencies belonging to it. The size of the flux is the most basic information about the flux. It is therefore a useful starting point for apprehending other concepts about flux.

Two dependencies are said to be *concomitant* if they belong to the same flux. The dependencies "*with* <case *guerillas*" and "*clashed* obl> *fight*" are concomitant at position 2.

The flux represents the set of pending syntactic relations that the speaker has to keep in mind after every word. One might expect it to be limited by the same boundary as that stated by Miller (1956) and not exceed 7 ± 2. We will see that this is not the case.

2.2 Spans and bouquets

Other characteristics of the flux can be considered. The *left span* (resp. *right span*) of the flux is the number of words on the left (resp. right) which are vertices of a dependency in the flux. For instance, the left span is 1 in position 1 (*US*), 2 in position 2 (*clashed, with*) and 3 in position 3 (*clashed, in, a*).

The left span in a given position corresponds to the number of words awaiting a governor or a dependent on the right of this position and the right span to the number of elements expected. In a transition-based parser (Bohnet & Nivre 2012, Dyer et al. 2015), it is the minimal number of words that must be stored in the stack.[2] Again, one might expect the left span to be bounded due to short-term memory limitations, but it is not really the case. This can be illustrated by looking at what happens at position 3: the left span is 3 but the right span is only 1; all the words of the left span are linked to the same word (*fight*) on the right. This means that the information can be factorized and that the three words in the left span count more or less as one, which is their common target.

The flux configuration in position 3 is called a *left-branching bouquet*. A *bouquet* is a set of dependencies sharing the same vertex. When the common vertex is on the left, the bouquet is *right-branching*, and *left-branching* when the common vertex is on the right (Fig. 2).

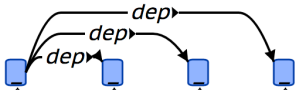


Figure 2. Right-branching bouquet

2.3 Disjoint dependencies and weight

We would like to measure the flux modulo the bouquets. This measure will be called the *weight* of the flux.

A set of dependencies is said to be *disjoint* if the dependencies do not share any vertex (Fig. 3). The *weight* is the size of the largest disjoint subset of dependencies in the flux.

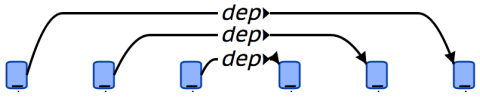


Figure 3. Disjoint dependencies

The weight of the flux is equal to 1 in position 3: it is not possible to find two disjoint dependencies. The weight is equal to 2 in position 2 because the subset { *with* <case *guerrillas, clashed* obl> *fight* } is disjoint but there is no disjoint subset with 3 elements.

[2] In practice all the nodes that are likely to have a dependent on the right are stored in the stack in arc-standard and arc-eager parsing strategies.

As we will see in the next section, the weight is clearly bounded. The weight measures more or less the center-embeddings: the fact that the dependency "*with* <`case` *guerrillas*" is disjoint from "*clashed* `obl`> *fight*" but concomitant means that the phrase *with guerrillas* headed by *guerrillas* is center-embedded in the phrase headed by *clashed*.[3] In other words, the weight is likely to measure the cognitive cost of parsing. This is noticeable if we compare the flux in positions 2 and 3. We saw earlier that the sizes of the flux at these two positions are equal: both have a value of 3. However, their weights are unequal: the weight at position 2 has a value of 2, whereas the weight at position 3 has a value of 1. Position 3 is simpler than position 2, because, as said before, the three dependencies at position 3 have a common target and requires less cognitive space than the disjoint dependencies at position 2.

We hypothesize that dependencies forming bouquets are cognitively less costly than dependencies forming disjoint subsets. This hypothesis is supported by the fact that the flux weight is clearly bounded while the size is not. We suppose that information can be factorized in case of dependencies sharing a same vertex. It is quite intuitive with right-branching bouquet, when the common vertex is on the left: only this vertex must be stored to analyze the bouquet. We postulate that the complexity is quite similar in case of left-branching bouquet, when two words are waiting for the same target word, but it remains to be proved by further studies.

Another advantage of the weight on the size is that it smooths out some idiosyncrasies of the UD scheme. For instance, coordination is analyzed in UD with every conjunct depending on the first conjunct, forming potentially very extended right-branching bouquets.

To calculate the weight, we have to find the biggest subset of disjoint dependencies in the flux. We can start with any dependency D in the flux with at least one vertex that is not shared with other dependencies in the flux (such a dependency exists because the structure is acyclic). Then we suppress all the dependencies that share a vertex with D and therefore cannot be disjoint from D. If the remaining flux is not empty we start over exactly the same process: choosing a dependency with at least one vertex that is not shared with other dependencies in the remaining flux and deleting all the dependencies sharing a vertex with it. At the end we obtain one of the biggest sets of disjoint dependencies in the flux. This simple algorithm is linear in time.

[3] In UD, prepositions are dependent on the noun they introduce. That is why the head of the PP *with guerrillas* is the noun in the dependency tree taken as an example here.

2.4 R/L ratio and density

Note that the size of the flux is higher than the left and right span, which are both higher than the weight. Some ratios can be interesting to study.

Head-initial languages, such as Standard Arabic or Welsh, have right-branching dependency trees, while head-final languages, such as Japanese, Korean, or Turkish, have left-branching dependency trees. In other words, head-initial languages should have an *R/L ratio* (where R is the right span and L is the left span) higher than 1 and head-final languages an R/L ratio less than 1. Unfortunately, UD is not a very good resource to measure that due to some idiosyncrasies of the UD scheme, such as the right-branching analysis of coordination, which is particularly irrelevant for head-final languages.

The *density* of the flux is the W/S ratio, where W is the weight and S is the size. This ratio measures the proportion of bouquets in the flux: a disjoint flux, that is, a flux without bouquets, has a density of 1. The more bouquets the flux has, the lower the density is. For instance, the density in position 1 is 1, in position 2, 2/3, and in position 3, 1/3.

3 Results on UD

3.1 The UD corpus

We studied the flux on the whole collection of UD treebanks. The 70 dependency treebanks distributed by the UD project have all been corrected manually and they follow a common annotation scheme. Nevertheless, these treebanks were developed by different teams, who may have interpreted the guidelines differently and the coherence and quality of the different treebanks have not yet been verified. And as mentioned above, some of the decisions made for the UD annotation scheme are not very suitable for a study of flux. Despite the defects of this resource, however, it is the only available resource of this scale allowing a cross-linguistic study of 50 different languages.

3.2 List of measures

Table 1 gives the following measures of the flux for each UD treebank. Average values were

calculated on the values in each inter-word position.

- S-max: maximum size
- S-av: average size
- W-max: maximum weight
- W-av: average weight
- L-max: maximum left span
- R-max: maximum right span
- L-av: average left span
- R-ax: average right span
- R/L-av: average R/L ratio
- D-av: density = average W/S ratio

3.3 Sizes

The maximum size varies from 8 for Kazakh, Sanskrit, Uyghur, and Vietnamese, to 97 for Ancient Greek. The average size ranges from 1.92 for Polish to 3.61 for Czech-CLTT. As said before, the highest sizes are due to the bouquet-wise annotation of some constructions, such as coordination (`conj`), apposition (`appos`), flat (sic!) constructions (`flat`), and multiword expressions (`fixed`). We converted the annotations to obtain a string-analysis of these constructions, giving a maximum size between 6 for Sanskrit and 77 for Arabic-NYUAD and an average size between 1.89 for Polish and 3.44 for Persian.[4] Further investigations are needed to understand what could cause excessive flux sizes.

3.4 Weights

Compared to the size, the weight is more stable. The maximum weight ranges from 3 (only for Sanskrit) to 6. In the whole UD database only one occurrence with a weight of 7 was found, for Czech-CLTT. Most of the fluxes with a maximum weight that we checked were due to erroneous analysis. The average weight varied from 1.18 for Polish and Slovak to 1.77 for Czech-CLTT. Weight is studied in greater detail in the next section.

3.5 Spans

The left span is more stable among the various treebanks than the right span with values between 7 and 17 against values between 5 and 97. As expected, treebanks with the highest R/L ratio are head-initial languages: 1.31 for Old Church Slavonic, 1.37 for Irish, 1.55 and 1.32 for Arabic, 1.22 for Indonesian and 1.23 for Gothic. The first exception is the value of 1.36 for Czech-CLTT, but this small corpus of Czech is atypical, the other two Czech treebanks having R/L ratios of 1.03 and 1.00. The second exception is that we have the value of 1.29 for Dutch-LassySmall, while the other Dutch treebank has an R/L ratio of 0.92 for Dutch.

The results for head-final languages are not relevant, as forecasted. Japanese has an R/L ratio of 1.17, Turkish, 1.04, and Korean 0.99, while the minimum ratio is 0.77 for Persian. The average ratio on the whole database is 1.05.

3.6 Densities

The density is quite stable with an average value between 57.00 % for Persian and 72.20 % for Polish, with 65.31 % for the whole database. This means that about 2/3 of dependencies in the flux form together disjoint sets and 1/3 are additional dependencies forming bouquets with the 2 other thirds. In fact, many fluxes have a density of 1 with only one element, as the flux at position 1 in Fig. 1, and form disjoint fluxes consequently.

4 A closer look at weight

4.1 Distribution of weight

Table 2 shows the distribution of the value of the weight of the flux for the 70 treebanks. For each treebank and each value between 1 and 6, we indicate the percentage of inter-word positions in the treebank with this value.

The first main result is that 99.62 % of inter-word positions in the whole UD database have a weight less than (or equal to) 3. Only 0.36 % have a weight of 4, 0.02 %, of 5, and 0.00 % of 6. For Polish, Sanskrit, Slovak and Vietnamese, 99.9 % of positions have a weight less than 3.

We have seen that some small corpora, such as Czech-CLLT, can have more exceptional values. If we put corpora with fewer than 1,000 sentences aside, Arabic, Chinese, and Korean are the three languages with more than 10% of positions with weight of 3.

Positions with a flux weight of 1 account for 62.15 % of positions in the whole database, and more than 80 % of positions in Finnish-FTB, Polish, and Slovak.

[4] The maximum size for Arabic is due to a sentence with 385 words and 77 nominal modifier (`nmod`) relations depending on the 5th word, which is likely to be a wrong analysis.

	Tokens	Trees	S-max	S-av	W-max	W-av	L-max	R-max	L-av	R-av	R/L-av	D-av
UD_Ancient_Greek	182030	12613	97	3,01	6	1,49	12	97	2,31	1,99	1,13	60,32%
UD_Ancient_Greek-PROIEL	198034	15865	31	2,89	6	1,49	12	29	2,19	1,99	1,14	61,96%
UD_Arabic	254120	6984	36	2,93	5	1,66	9	35	2,06	2,41	1,32	66,47%
UD_Arabic-NYUAD	738889	19738	78	3,12	6	1,66	12	78	1,95	2,74	1,55	64,65%
UD_Basque	97069	7194	13	2,25	5	1,36	9	11	1,86	1,63	1,05	70,68%
UD_Belarusian	6864	333	17	2,48	4	1,44	9	17	1,98	1,78	1,09	69,28%
UD_Bulgarian	140425	10022	14	2,24	5	1,28	9	14	1,90	1,50	0,97	67,67%
UD_Catalan	474069	14832	20	2,69	6	1,48	13	19	2,17	1,83	1,03	64,16%
UD_Chinese	111271	4497	27	3,24	6	1,65	14	25	2,77	1,86	0,84	61,28%
UD_Coptic	8519	320	9	2,74	4	1,43	8	8	2,23	1,74	1,00	60,08%
UD_Croatian	183816	8289	13	2,52	5	1,40	11	13	2,13	1,65	0,98	65,74%
UD_Czech	1332566	77765	56	2,43	6	1,37	17	56	2,03	1,63	1,00	67,23%
UD_Czech-CAC	483520	24081	47	2,50	6	1,39	11	47	2,04	1,71	1,03	66,49%
UD_Czech-CLTT	26781	814	28	3,61	7	1,77	10	24	2,36	2,83	1,36	62,24%
UD_Danish	90710	4947	16	2,61	4	1,34	16	12	2,20	1,61	0,97	63,32%
UD_Dutch	197925	13050	15	2,89	5	1,43	12	15	2,46	1,69	0,92	60,44%
UD_Dutch-LassySmall	91793	6841	29	2,74	4	1,33	10	29	2,06	1,87	1,21	61,32%
UD_English	229733	14545	18	2,58	6	1,35	13	17	2,19	1,58	0,92	63,01%
UD_English-LinES	67197	3650	25	2,54	5	1,35	10	24	2,13	1,61	0,95	63,89%
UD_English-ParTUT	38114	1590	15	2,63	5	1,39	9	14	2,26	1,60	0,89	62,79%
UD_Estonian	34628	3172	10	2,26	5	1,25	9	10	1,82	1,58	1,11	68,04%
UD_Finnish	180911	13581	33	2,31	6	1,31	10	33	1,89	1,63	1,06	68,53%
UD_Finnish-FTB	143326	16856	14	2,06	5	1,19	11	14	1,77	1,39	0,98	70,29%
UD_French	392230	16031	34	2,51	5	1,39	11	34	2,04	1,71	1,02	65,09%
UD_French-ParTUT	17927	620	11	2,70	5	1,44	11	10	2,31	1,68	0,91	62,86%
UD_French-Sequoia	60574	2643	31	2,63	5	1,44	12	31	2,15	1,75	1,00	64,58%
UD_Galician	109106	3139	15	2,56	5	1,41	11	15	2,04	1,80	1,08	64,54%
UD_Galician-TreeGal	15436	600	13	2,55	4	1,43	9	12	2,12	1,71	1,00	65,30%
UD_German	281974	14917	28	3,00	6	1,46	13	26	2,51	1,76	0,96	59,84%
UD_Gothic	45138	4372	21	2,53	4	1,38	10	20	1,87	1,91	1,23	65,96%
UD_Greek	51351	2065	13	2,51	5	1,41	10	9	2,12	1,65	0,95	65,57%
UD_Hebrew	149088	5725	62	2,56	5	1,48	11	61	2,01	1,86	1,11	66,99%
UD_Hindi	316274	14963	18	3,20	6	1,58	13	15	2,76	1,84	0,85	59,67%
UD_Hungarian	31584	1351	13	2,83	6	1,54	10	10	2,44	1,75	0,89	64,54%
UD_Indonesian	110143	5036	28	2,31	5	1,39	9	28	1,75	1,85	1,22	70,30%
UD_Irish	13826	566	18	2,88	5	1,56	7	18	1,94	2,34	1,37	64,95%
UD_Italian	282611	13402	35	2,50	5	1,39	10	34	2,10	1,65	0,96	65,69%
UD_Italian-ParTUT	42651	1590	14	2,59	5	1,43	9	14	2,20	1,66	0,93	64,46%
UD_Japanese	173458	7675	15	2,79	5	1,55	15	11	2,17	2,03	1,17	64,52%
UD_Kazakh	529	31	8	2,67	4	1,52	6	5	2,21	1,82	1,00	67,07%
UD_Korean	63426	5350	23	2,73	5	1,62	9	20	2,25	1,93	0,99	68,80%
UD_Latin	18184	1334	17	2,86	5	1,52	8	16	2,31	1,87	1,02	63,32%
UD_Latin-ITTB	280734	16508	11	2,67	6	1,46	10	10	2,30	1,65	0,89	64,10%
UD_Latin-PROIEL	159407	15324	28	2,77	6	1,47	14	28	2,15	1,91	1,12	64,14%
UD_Latvian	44795	3054	18	2,48	6	1,39	9	17	2,04	1,68	0,99	67,31%
UD_Lithuanian	5356	263	14	2,43	4	1,38	9	13	2,06	1,60	0,95	68,01%
UD_Norwegian-Bokmaal	280256	18106	38	2,44	5	1,30	11	38	2,08	1,54	0,96	64,18%
UD_Norwegian-Nynorsk	276580	16064	38	2,50	6	1,32	11	38	2,12	1,57	0,96	63,82%
UD_Old_Church_Slavonic	47532	5196	20	2,48	5	1,34	8	19	1,76	1,93	1,31	66,03%
UD_Persian	136896	5397	14	3,45	6	1,64	13	10	3,03	1,81	0,77	57,00%
UD_Polish	72763	7127	10	1,92	4	1,18	8	7	1,62	1,40	1,04	72,20%
UD_Portuguese	217591	8891	19	2,54	5	1,43	13	19	2,12	1,70	0,98	65,84%
UD_Portuguese-BR	287884	10874	38	2,54	5	1,45	10	38	2,05	1,77	1,04	66,46%
UD_Romanian	202187	8795	14	2,39	6	1,40	9	14	1,95	1,69	1,05	67,75%
UD_Russian	87841	4429	31	2,34	5	1,37	10	30	1,83	1,74	1,12	69,62%
UD_Russian-SynTagRus	988460	55398	18	2,34	6	1,37	10	17	1,94	1,63	1,01	68,64%
UD_Sanskrit	1206	190	8	2,23	3	1,29	6	5	2,05	1,39	0,82	68,76%
UD_Slovak	93015	9543	10	2,00	4	1,18	9	8	1,74	1,36	0,96	70,22%
UD_Slovenian	126593	7212	17	2,50	5	1,30	13	17	2,21	1,47	0,87	64,12%
UD_Slovenian-SST	19488	2137	14	2,77	4	1,33	12	8	2,34	1,62	0,94	60,51%
UD_Spanish	419587	15587	38	2,51	5	1,42	11	38	2,03	1,74	1,04	65,91%
UD_Spanish-AnCora	496953	15959	31	2,63	5	1,47	12	31	2,16	1,76	1,00	65,21%
UD_Swedish	76442	4807	31	2,58	5	1,32	10	31	2,07	1,68	1,04	62,95%
UD_Swedish-LinES	64787	3650	25	2,55	5	1,34	10	24	2,07	1,67	1,03	63,25%
UD_Tamil	9581	600	10	2,42	4	1,48	9	8	2,07	1,74	1,00	70,91%
UD_Turkish	48093	4660	13	2,44	6	1,48	9	13	2,00	1,77	1,04	71,20%
UD_Ukrainian	12846	863	11	2,19	4	1,27	8	9	1,85	1,49	0,98	69,22%
UD_Urdu	123271	4595	32	3,44	5	1,66	15	29	2,92	1,96	0,85	58,34%
UD_Uyghur	1662	100	8	2,93	5	1,73	7	6	2,75	1,80	0,77	67,31%
UD_Vietnamese	31799	2200	8	2,09	4	1,25	7	8	1,68	1,57	1,12	70,45%
Total	1E+007	630518	97	2,62	7	1,43	17	97	2,11	1,79	1,05	65,31%

Table 1: Size, weight, left and right spans, R/L ratio and density for the 70 UD treebanks available

	Tokens	Trees	1	2	3	4	5	6
UD_Ancient_Greek	182030	12613	57.77%	35.79%	5.96%	0.45%	0.02%	0.00%
UD_Ancient_Greek-PROIEL	198034	15865	57.81%	35.97%	5.74%	0.46%	0.02%	0.00%
UD_Arabic	254120	6984	47.15%	41.10%	10.60%	1.10%	0.05%	0.00%
UD_Arabic-NYUAD	738889	19738	47.16%	40.86%	10.67%	1.23%	0.08%	0.00%
UD_Basque	97069	7194	67.85%	28.28%	3.66%	0.21%	0.01%	0.00%
UD_Belarusian	6864	333	62.43%	31.70%	5.37%	0.50%	0.00%	0.00%
UD_Bulgarian	140425	10022	73.86%	24.20%	1.87%	0.06%	0.00%	0.00%
UD_Catalan	474069	14832	57.82%	36.58%	5.30%	0.30%	0.01%	0.00%
UD_Chinese	111271	4497	49.73%	37.35%	10.91%	1.78%	0.22%	0.01%
UD_Coptic	8519	320	61.76%	33.74%	4.37%	0.13%	0.00%	0.00%
UD_Croatian	183816	8289	64.46%	31.70%	3.66%	0.17%	0.01%	0.00%
UD_Czech	1332566	77765	66.78%	29.61%	3.42%	0.17%	0.01%	0.00%
UD_Czech-CAC	483520	24081	65.16%	30.82%	3.80%	0.22%	0.01%	0.00%
UD_Czech-CLTT	26781	814	42.78%	41.74%	12.19%	2.71%	0.53%	0.05%
UD_Danish	90710	4947	69.13%	27.85%	2.90%	0.12%	0.00%	0.00%
UD_Dutch	197925	13050	63.41%	31.12%	5.00%	0.44%	0.02%	0.00%
UD_Dutch-LassySmall	91793	6841	70.30%	27.04%	2.50%	0.15%	0.00%	0.00%
UD_English	229733	14545	68.45%	28.08%	3.29%	0.17%	0.00%	0.00%
UD_English-LinES	67197	3650	68.40%	28.16%	3.20%	0.23%	0.01%	0.00%
UD_English-ParTUT	38114	1590	64.99%	31.09%	3.77%	0.15%	0.00%	0.00%
UD_Estonian	34628	3172	77.26%	20.37%	2.21%	0.15%	0.01%	0.00%
UD_Finnish	180911	13581	72.60%	23.79%	3.28%	0.30%	0.03%	0.00%
UD_Finnish-FTB	143326	16856	82.77%	15.91%	1.22%	0.10%	0.00%	0.00%
UD_French	392230	16031	64.31%	32.23%	3.27%	0.17%	0.01%	0.00%
UD_French-ParTUT	17927	620	60.66%	34.99%	4.05%	0.26%	0.03%	0.00%
UD_French-Sequoia	60574	2643	61.25%	33.76%	4.69%	0.29%	0.01%	0.00%
UD_Galician	109106	3139	62.78%	33.73%	3.34%	0.14%	0.00%	0.00%
UD_Galician-TreeGal	15436	600	61.98%	33.75%	4.02%	0.25%	0.00%	0.00%
UD_German	281974	14917	59.60%	35.60%	4.46%	0.33%	0.01%	0.00%
UD_Gothic	45138	4372	65.83%	30.51%	3.46%	0.21%	0.00%	0.00%
UD_Greek	51351	2065	62.49%	33.73%	3.59%	0.19%	0.00%	0.00%
UD_Hebrew	149088	5725	58.04%	36.56%	5.17%	0.23%	0.00%	0.00%
UD_Hindi	316274	14963	49.02%	44.76%	5.65%	0.54%	0.03%	0.00%
UD_Hungarian	31584	1351	56.04%	35.24%	7.58%	0.99%	0.14%	0.02%
UD_Indonesian	110143	5036	64.82%	31.38%	3.61%	0.18%	0.01%	0.00%
UD_Irish	13826	566	53.11%	38.57%	7.47%	0.82%	0.03%	0.00%
UD_Italian	282611	13402	64.93%	31.55%	3.37%	0.16%	0.01%	0.00%
UD_Italian-ParTUT	42651	1590	61.56%	34.48%	3.78%	0.17%	0.00%	0.00%
UD_Japanese	173458	7675	50.98%	42.82%	6.03%	0.17%	0.00%	0.00%
UD_Kazakh	529	31	55.27%	37.42%	6.88%	0.43%	0.00%	0.00%
UD_Korean	63426	5350	51.30%	37.10%	10.24%	1.28%	0.09%	0.00%
UD_Latin	18184	1334	56.34%	35.90%	7.01%	0.69%	0.06%	0.00%
UD_Latin-ITTB	280734	16508	60.44%	33.25%	5.85%	0.45%	0.02%	0.00%
UD_Latin-PROIEL	159407	15324	61.30%	31.41%	6.27%	0.92%	0.10%	0.00%
UD_Latvian	44795	3054	67.21%	27.51%	4.75%	0.48%	0.05%	0.01%
UD_Lithuanian	5356	263	66.76%	28.97%	4.07%	0.21%	0.00%	0.00%
UD_Norwegian-Bokmaal	280256	18106	72.73%	24.95%	2.23%	0.08%	0.00%	0.00%
UD_Norwegian-Nynorsk	276580	16064	70.73%	26.67%	2.50%	0.10%	0.00%	0.00%
UD_Old_Church_Slavonic	47532	5196	69.31%	27.41%	3.15%	0.13%	0.00%	0.00%
UD_Persian	136896	5397	45.75%	45.14%	8.52%	0.59%	0.01%	0.00%
UD_Polish	72763	7127	82.46%	16.96%	0.57%	0.00%	0.00%	0.00%
UD_Portuguese	217591	8891	61.69%	33.94%	4.16%	0.21%	0.01%	0.00%
UD_Portuguese-BR	287884	10874	60.06%	35.50%	4.24%	0.20%	0.01%	0.00%
UD_Romanian	202187	8795	64.42%	31.62%	3.74%	0.22%	0.00%	0.00%
UD_Russian	87841	4429	66.76%	29.75%	3.28%	0.21%	0.00%	0.00%
UD_Russian-SynTagRus	988460	55398	67.30%	29.04%	3.42%	0.23%	0.01%	0.00%
UD_Sanskrit	1206	190	71.95%	27.07%	0.98%	0.00%	0.00%	0.00%
UD_Slovak	93015	9543	83.09%	16.24%	0.66%	0.01%	0.00%	0.00%
UD_Slovenian	126593	7212	72.12%	25.49%	2.31%	0.08%	0.00%	0.00%
UD_Slovenian-SST	19488	2137	70.09%	26.59%	3.17%	0.15%	0.00%	0.00%
UD_Spanish	419587	15587	61.54%	34.75%	3.56%	0.15%	0.01%	0.00%
UD_Spanish-AnCora	496953	15959	58.20%	36.45%	5.10%	0.25%	0.00%	0.00%
UD_Swedish	76442	4807	70.77%	26.62%	2.50%	0.10%	0.01%	0.00%
UD_Swedish-LinES	64787	3650	69.32%	27.51%	3.03%	0.13%	0.00%	0.00%
UD_Tamil	9581	600	58.14%	36.18%	5.23%	0.45%	0.00%	0.00%
UD_Turkish	48093	4660	60.71%	31.69%	6.69%	0.85%	0.06%	0.00%
UD_Ukrainian	12846	863	74.84%	23.60%	1.49%	0.06%	0.00%	0.00%
UD_Urdu	123271	4595	44.94%	45.09%	8.82%	1.04%	0.11%	0.00%
UD_Uyghur	1662	100	43.87%	42.16%	11.50%	2.12%	0.34%	0.00%
UD_Vietnamese	31799	2200	76.10%	22.71%	1.18%	0.01%	0.00%	0.00%
Total	1.2E+07	630518	62.15%	32.75%	4.71%	0.36%	0.02%	0.00%

Table 2: Percentage of inter-word positions for every possible value of the weight

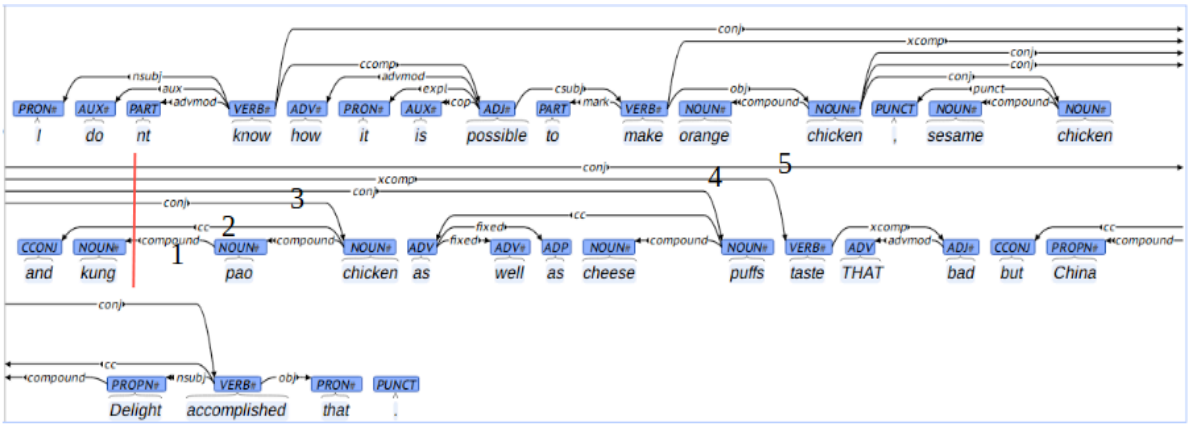


Figure 4. A dependency tree from UD-English with weight 5

We do not have enough metadata to know if the differences between treebanks are due to differences between languages or to differences between genres. It is highly likely that some kinds of texts (e.g. legal texts, specification sheets) are much more complicated than others. For 16 languages, there are two or three treebanks and noticeable divergences are observed in only three cases (Finnish, Dutch, and again Czech-CLTT). At first sight variations between languages appear to be greater than variations between corpora in the same language, but this point needs further investigation.

4.2 Examples

The only example of English with weight 6 was erroneously annotated. We give here two examples with weight 5. In all the examples, punctuation links (`punct`) have been removed and are not considered.

(1) *I dont know how it is possible to make orange chicken, sesame chicken and kung pao chicken as well as cheese puffs taste THAT bad but China Delight accomplished that.* (en-ud-train.conllu sent_id = reviews-235423-0012)

Sentence (1) has a weight of 5 between *kung* and *pao* (Fig. 4). A set of five disjoint dependencies at this point is:

1: kung <`compound` pao
2: and <`cc` chicken$_3$
3: chicken$_1$ `conj`> puffs
4: make `xcomp`> taste
5: know `conj`> accomplished

(2) *as an example they took payment for 5 out of 6 monthly plan premiums for a yearly policy and cancelled the contract for the remainder of the policy for reasons they stated was not receiving information on other licensed drivers in the household?* (en-ud-train.conllu sent_id = reviews-217359-0006)

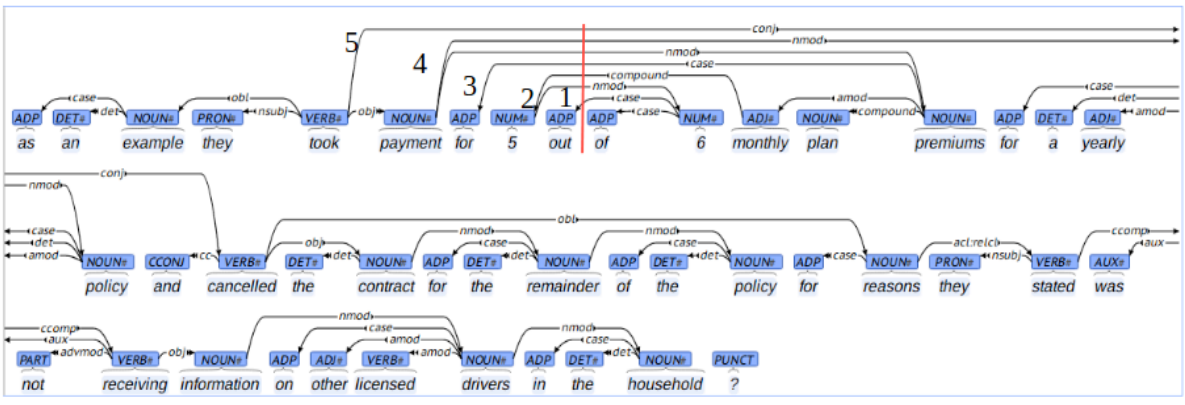


Figure 5. Another dependency tree from UD-English with weight 5

Sentence (2) has two positions with weight 5. We consider the flux between *out* and *of* (Fig. 5).

1: out `case`> 6
2: 5 <`compound` monthly
3: for <`case` premiums
4: payment `nmod`> policy
5: took `conj`> cancelled

If we except the two small corpora of Czech and Uyghur, Chinese appears to be the language with the largest number of positions with a weight higher than 5 (0.23 %). We will study an example with weight 6.

(3) 一 級 抗體 對於 檢測 如
one level antibody for detect such_as

癌症、 糖尿 病、 帕金森 氏 症
cancer, diabetes disease, Parkinson 's disease

和 阿爾茨海默 氏 病 等 疾病
and Alzheimer 's disease etc. disease

所 特有 的 生物 標記
that specifically_have de(PART) biology marker

是 非常 有用 的。
be very useful de(PART).

(zh-ud-train.conllu id=21)

'Primary antibodies are useful for detecting biomarkers that diseases such as cancer, diabetes, Parkinson's disease, Alzheimer's disease, etc. specifically contain.'

The weight 6 appears between the noun 阿爾茨海默 'Alzheimer' and the case particle 氏 ('s). This flux contains 9 dependencies and can be separated into 6 disjoint bouquets of dependencies:

1: 阿爾茨海默 'Alzheimer' <`case:suff` 氏
2: 和 'and' <`cc` 病 'disease'
3: 癌症 'cancer' `conj`> 病 'disease'
癌症 'cancer `acl`> 等 'etc.'
癌症 'cancer' `appos`> 疾病 'disease'
4:如 'such_as' <`csubj` 特有 'specifically_have'
5: 檢測 'detect' `obj`> 疾病 'disease'
檢測 'detect' `xcomp`> 有用'useful'
6: 抗體'antibody' <`nsubj` 有用'useful'

The complexity of this Chinese sentence, compared to its English translation, is in great part due to word order differences.

1. In Chinese, adverbs and adverbial modifiers are placed before the verb. As a result, 有用 'useful' is at the end of the sentence and the long adverbial modifier 'for detecting …' is between the subject and the verb.

2. Noun modifiers are placed before the noun and '[diseases [such as cancer, diabetes, Parkinson's disease, and Alzheimer's disease, etc.]' becomes '[[such as cancer, diabetes, Parkinson's disease, and Alzheimer's disease, etc.] diseases]').

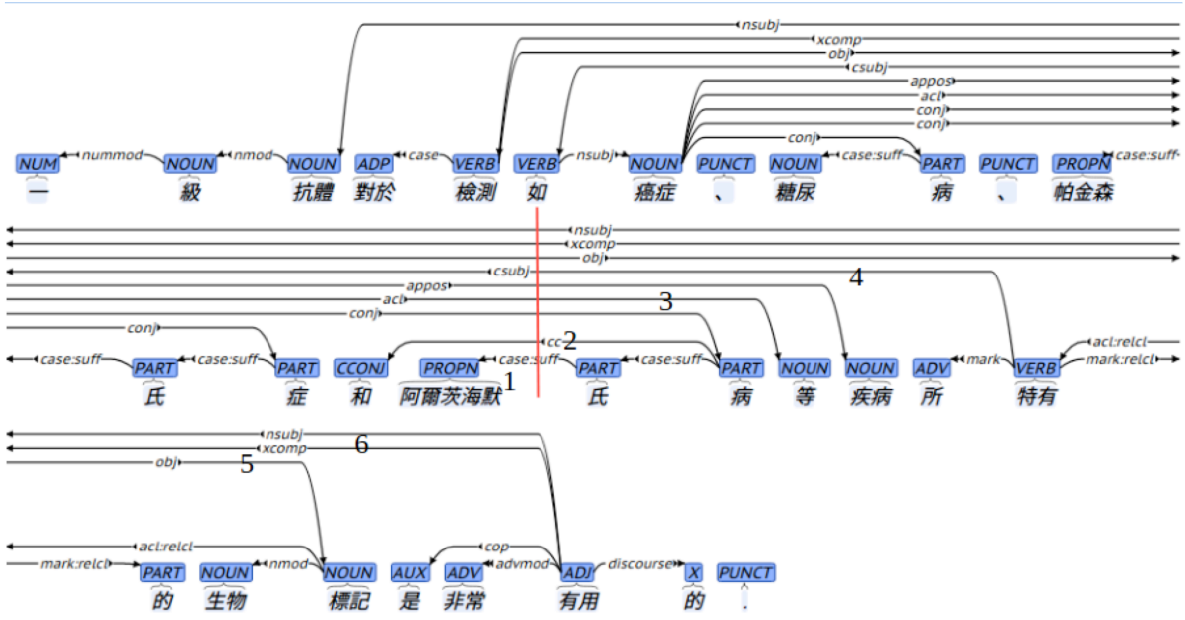


Figure 6. A dependency tree from UD-Chinese with weight 6

3. Relative clauses are also placed before the noun, which is a source of complexity discussed in Hsiao & Gibson (2003): "A key word-order difference between Chinese and other Subject-Verb-Object languages is that Chinese relative clauses precede their head nouns. Because of this word order difference, the results follow from a resource-based theory of sentence complexity, according to which there is a storage cost associated with predicting syntactic heads in order to form a grammatical sentence."
In any case, [biomarkers [(that are) specific to[diseases [such as cancer, diabetes, Parkinson's disease, and Alzheimer's disease etc.]]]] becomes [[[[such as cancer, diabetes, Parkinson's disease, and Alzheimer's disease etc.] disease] (that) specifically have] biomarkers].

5 Conclusion

We have studied different parameters concerning the dependency flux on a set of treebanks in 50 languages. We saw that the size, as well as the left and right spans, of the flux can vary considerably depending on the corpus and its language, and that they are not clearly bounded. Moreover, these values are quite heavily dependent on certain annotation choices. For instance the fact that UD proposes a bouquet-based analysis (rather than a string-based analysis) of coordination (and other similar constructions) significantly increases the size and the right span of the dependency flux.

Conversely, the dependency flux weight appears to be more homogeneous across languages and much less dependent on particular annotation choices (such as bouquet vs. string-based analysis of coordination). Weight measures what is traditionally called center embedding in constituency-based formalisms. We observe that weight is bounded by 5 except for very few positions (less than 1 position for 10,000 with weight of 6), which could be related to short-term memory limitations.

What now remains is to study all the data we have collected to determine, language after language, genre after genre, what are the most complex constructions and under which conditions they can appear. In particular, a comparison between weight and dependency distance (Liu 2010) is needed to determine how they are correlated and which one is the best predictor of the complexity.[5]

Acknowledgments
We acknowledge our three reviewers for their comments. We could not answer their numerous suggestions but we hope to do that in further works.

References

Maria Babyonyshev, Edward Gibson. 1999. The Complexity of Nested Structures in Japanese, *Language, 75*(3), 423-450.

Bernd Bohnet, Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. *Proceedings of EMNLP,* 1455-1465.

Marie-Amélie Botalla. 2014. Analyse du flux de dépendance dans un corpus de français oral annoté en microsyntaxe. Master thesis. Université Sorbonne Nouvelle.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax.* Cambridge MA: MIT Press.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *Proceedings of ACL*, Beijing.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1-76.

Franny Hsiao, Edward Gibson. 2003. Processing relative clauses in Chinese. *Cognition*, *90*(1), 3-27.

Ugo Jardonnet. 2009. Analyse du flux de dépendance. Master thesis. Université Paris Nanterre.

Sylvain Kahane. 2001. Grammaires de dépendance formelles et Théorie Sens-Texte. Tutorial. *Proceedings of TALN*, vol. 2, 17-76.

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

Haitao Liu, Richard Hudson, Zhiwei Feng. 2009. Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory*, *5*(2), 161-174.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua, 120,* 1567-78.

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

G. A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for pro-

[5] Fluxes with important weight or size tend to contain long dependencies and long dependencies to belong to large fluxes, but the two measures are quite different and remain partly independent.

cessing information. *Psychological review*, *63*(2), 81-97.

G. A. Miller. 1962. Some psychological studies of grammar. *The American Psychologist, 17,* 748-762.

G. A Miller, Noam Chomsky. 1963. Finitary models of language users. In D. Luce (ed.), *Handbook of Mathematical Psychology*. John Wiley & Sons. 2-419.

M. Murata, K. Uchimoto, Q. Ma, H. Isahara. 2001. Magical number seven plus or minus two: Syntactic structure recognition in Japanese and English sentences. *International Conference on Intelligent Text Processing and Computational Linguistics.* Lecture notes in computer science, Springer, 43-52.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of LREC*.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

Chunxiao Yan. 2017. Étude du flux de dépendance dans 70 corpus (50 langues) de UD. Master thesis. Université Sorbonne Nouvelle.

V. H. Yngve.1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, *104*(5), 444-466.