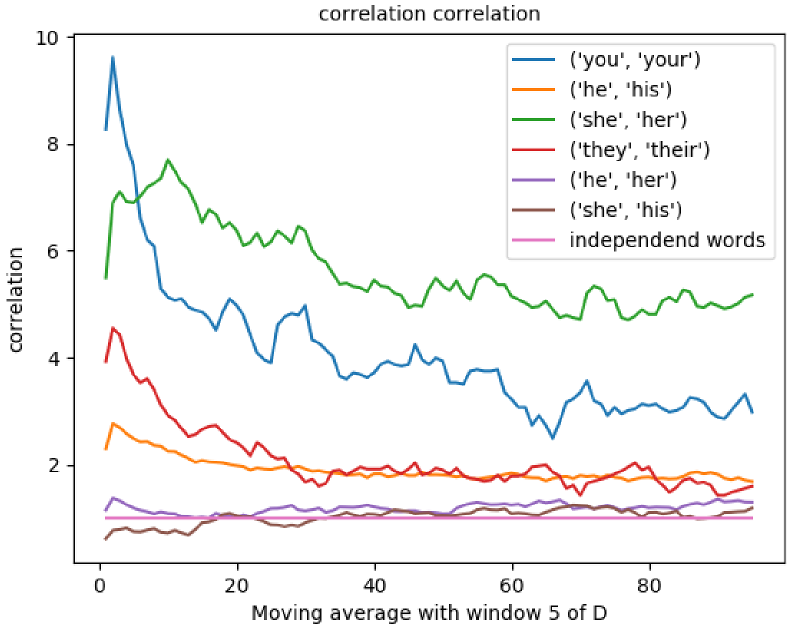
1 Long-Range Dependencies

a) since we can use $P_D(w_1, w_2) = P(w_1 \cap w_2)$ for independend words and further $P(w_1 \cap w_2) = P(w_1) * P(w_2)$ the correlation should be 1 for independend words.

d) correlation for ('you', 'your')
0.9221643776870978, 10.912288343411435, 11.373381240819898, 7.992112968564083,
10.143844869673593, 7.684737915525758, 5.994100997968329, 8.145836932239188,
5.9941118457142535, 5.22564069657605, 5.686731786140494, 5.37934574639068,
4.149784759378078, 5.225659610731501, 4.918272319333419, 5.8404536640794635,
4.610888643915833, 3.8424106801646123, 5.071986687335467, 4.303507750064686,
4.764602177524211, 6.301576324046464, 5.072005045491239, 4.457220588378105,
3.381342816412641, 2.9202532566578894, 4.610930367072644, 4.45723672144886,
4.149844840615155, 6.916414326250395, 3.6887576452374353, 4.918347977529823,
4.303558374573898, 5.072055531104863, 3.68877099691341, 3.381376473610213,
4.303573951585583, 3.68878101073382, 3.227686305100284, 3.3813887127572735,
3.996190276659316, 4.149893657901308, 3.381397892175713, 3.6888010385377474,
4.149904923591989, 4.303609000274172, 3.8425115130429846, 3.227712591711872,
3.8425184672296693, 5.994334233158624, 2.92031932029556, 3.381425430730066,
3.842532375678554, 3.535132984581986, 3.996240903138383, 2.920332533381782,
3.227738878751625, 5.07216568502325, 3.6888511089990415, 3.8425567157063796,
2.9203457465875715, 3.3814560296498795, 2.9203510319033663, 3.074056499545724,
3.0740592813034655, 2.920358959912929, 1.6907356646587284, 3.842584533258621,
2.151849285869343, 1.8444439141006141, 4.4574101593356765, 3.535190566810915,
4.150010072987964, 2.766675885614723, 2.9203827441998906, 2.6129763988252654,
3.2277972959285104, 3.074095444612248, 3.535212960406545, 2.305575756205301,
2.920398600606428, 3.3815172291508953, 3.5352257568742824, 3.381523349222836,
2.4592919340124464, 2.459294159503221, 3.0741204812475305, 3.6889479157452825,
3.6889512539995706, 3.3815417095715756, 2.30559870657892, 2.766720951599014,
2.7667234553078557, 3.2278469521914332, 3.227849873195655, 3.074145518290632,
3.5352705452409072, 3.5352737444533773, 1.537076932031147, 2.766740981396626
correlation for ('he', 'his')
0.017763055333367672, 2.966432924936185, 3.0552510465530127, 2.8332150167265215,
2.6022970252615036, 2.398023319446205, 2.5756570070666926, 2.486843498464508,
2.389148237173484, 2.273689599086713, 2.4335606010767328, 2.220404017458899,
2.415801756984041, 1.9095509108088327, 2.2381733253724394, 2.1049506273854806,
1.9273194070307509, 2.042782786790627, 2.078311324565894, 2.104958246275078,
2.0694333974026167, 1.8740379483597382, 1.882921348613653, 1.9806218900591637,
2.0072688439325206, 1.7497005313014293, 2.0783263695979466, 1.785230676490154,
1.9362220877522172, 2.140504336672696, 1.8740532104841179, 1.9184637903263937,
1.9717562353859626, 1.6964224402919585, 1.9184689983210685, 1.927352543280 7879,
1.7674815814644085, 1.9273560313733897, 1.616493618097905, 1.776368220717 6192,
1.9628886600988398, 1.8385444358178302, 1.589853873489232, 1.9628939887357395,
1.785258139166 7335, 1.687558972052119, 1.9451355226739542, 1.8030267964112145,
1.7586188607809188, 1.7852662165853521, 2.0073160698984744, 1.696450072072108,

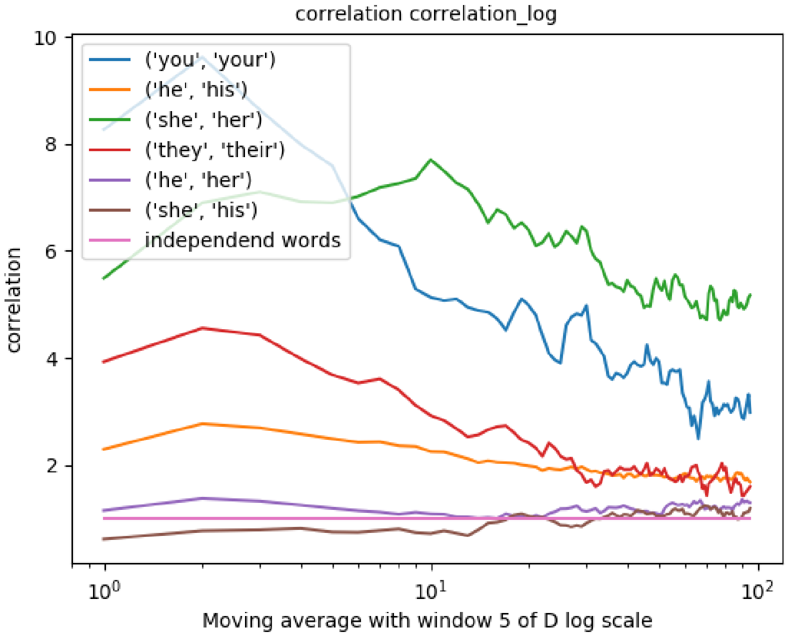
```
1.8119168998338706, 1.7319809568244298, 1.7408644857666888, 1.874095606578415,
1.820805436041363, 1.5543475104788842, 1.660932842761255, 1.8474563845846206,
1.962924184891323, 1.8829877998503017, 1.723113036486928, 1.7852888337463417,
1.7319981970624632, 1.7231177143001368, 1.82082191280028, 1.7053567005588977,
1.5099526116679713, 1.9629401714674592, 1.838592684110147, 1.8741227410854644,
1.5188401835550698, 1.7941871035863832, 1.8652457064811967, 1.90965804663282,
1.6876063125507623, 1.7586650119060683, 1.6076699758055348, 1.7853146826320514,
1.6520837386441822, 1.8563753432001875, 1.9007879566548809, 1.6432060289528887,
1.6432075159404642, 1.696502267890646, 1.7853259917549262, 1.936325464694138,
1.8297403976596842, 1.962975698124512, 1.8030969562067767, 1.6076888887141683,
2.05180369814955, 1.6609854492086644, 1.6165755364621135, 1.652106164186720 5,
1.8031067463797896, 1.820872992648048, 1.5366405502188893, 1.5721712342196605
correlation for ('she', 'her')
0.05300277247769686, 6.360338452551646, 7.57941018094985, 6.572361628473569,
6.890385361489725, 7.102403645484146, 7.367425373415592, 6.678391633345433,
6.466385051763431, 7.473451781332174, 7.950487812590686, 7.738481806575751,
7.155451980840737, 8.215526374962927, 6.4134167153926205, 6.837450453850416,
7.1554778799995855, 5.777391071668345, 6.466443564571976, 7.632530458124544,
6.36044780386304, 5.883419542351922, 6.307455487123243, 5.724418563182966,
6.201459055029108, 6.678500410206325, 6.73151047294114, 5.035386406274436,
6.201481501454243, 7.208566216920381, 6.2014927247277445, 6.095489817809137,
6.5725340987794185, 5.777474718024309, 5.406449123781668, 5.406454016026505,
5.777490401985701, 5.512472894671032, 4.770413552483067, 5.512482871076639,
5.088450331655062, 5.671507064273694, 5.141464327577091, 5.830531833072888,
4.982458984147811, 4.982463492768766, 5.088477958874432, 4.9294674833326555,
4.717451645346518, 5.194502017859047, 4.87647569480454, 6.731662757154935,
5.936589850792769, 3.9753985867261967, 4.770482620906494, 5.777589735718675,
5.565573130366723, 5.3535561413034785, 5.777605420304545, 5.300560227998862,
5.565593275761062, 4.823518537181442, 5.353580363807667, 4.7175156786316625,
4.9825491581173, 5.247583117268554, 4.39949285726152, 5.4596165585415655,
5.194591329188107, 4.240486554995467, 4.452514911907966, 4.611537474687913,
5.247616357800883, 5.035596011271312, 6.678796542947688, 5.141617864400059,
4.346526251631804, 4.134504322233284, 5.141631822747767, 4.982616790736441,
4.929614690072329, 4.717592520868478, 4.717596789955034, 4.717601059049318,
4.982639335350839, 6.201801380669037, 5.0356552506287615, 4.2935625727618065,
5.830769264157631, 4.770633715055666, 4.876652210664705, 4.92966376093934,
4.7176394812455555, 5.565759480829168, 4.505618895120138, 5.035696263308104,
5.247730328537448, 5.300742502450073, 5.777814556253836, 5.247744575227606
correlation for ('they', 'their')
0.0, 4.748491637314154, 6.078074795585484, 5.223350253875534, 3.608863440936858,
3.134015824043828, 4.083721284126168, 3.8937842826393165, 3.7038469374164595,
2.849115606856754, 3.513912428088081, 3.0390621472218404, 2.4692402289614237,
2.7541550552294316, 2.3742737477508755, 2.75416003955711, 2.2793069228091434,
2.6591938161820425, 3.229023984372534, 2.659198628671191, 2.8491439659881754,
1.6145163750007816, 1.99440438557818, 2.9441234237412166, 2.184351613547007,
2.089381694905331, 2.8491594348432683, 1.5195530736011276, 2.1843595198936465,
1.8994447795558689, 2.184363473088429, 1.709503395410494, 1.1396699615465329,
1.4245887410265867, 2.184371379520922, 1.5195640738351222, 1.994429651649185,
2.2793502358777586, 1.5195681989639305, 1.709515770763324, 1.8994636863278065,
2.3743317564298994, 2.089413836359457, 1.519575074228382, 1.7095235054498301,
2.1843931225053996, 1.8994739992711445, 1.8045019322006257, 1.994451308791,
2.27937498694177, 1.0447144810459048, 2.089430852823752, 2.279381174791763,
1.8045117296209152, 1.8994877500364638, 1.6145660485617082, 1.3296438314312309,
2.089442197287265, 1.7095451629444043, 1.7095467099293, 1.709548256916996,
1.8045247930134634, 1.5196012008006998, 2.1844287028682388, 1.7095544448957776,
1.8995066576638582, 2.4693608895250634, 1.6145835811365243, 2.2794141772258074,
1.0447324432919303, 1.5196122017315055, 1.3296618797516735, 1.8995186899863676,
1.3296642862314192, 2.1844504469939174, 1.7095714620684115, 1.5196204525341372,
2.0894800130553524, 1.614599652997941, 2.469389939199709, 2.089485685538626,
1.8995341603392986, 1.4246519094622179, 1.8995375982297267, 1.329677522025725,
1.5196328289060794, 1.2347027908076562, 1.8995444740479157, 2.564387360563825,
1.5196383295805413, 1.044752297019323, 1.3296859449411174, 1.6146201085570535,
1.614621569688252, 1.5196452054796188, 1.3296907580835384, 1.5196479558566722,
1.8045835806206072, 1.8045852136643716, 1.234717316170747
correlation for ('he', 'her')
0.0, 1.6022229898852482, 1.4167037150840878, 1.5010326754021832, 1.2311852185990009,
1.1299929354519975, 1.3492465169493955, 1.0456669968251582, 1.1974584508734445,
1.0119376347184341, 0.9950729078840352, 1.1468647282062197, 1.2311941311166785,
1.045672673997378, 0.9782108059923709, 0.7420916277690314, 1.1300041829781053,
1.1131394561558865, 1.1131404634097903, 0.8938863325041275, 1.1974714535232995,
0.9107537606038143, 1.0625470155143792, 0.9444870906601911, 1.197475787802669,
0.9782205428153522, 1.2649415017120984, 1.2480767443862848, 1.21434603932628,
1.2143471381705955, 1.0625547073897872, 1.45047281783898, 0.82643293473692,
1.1131555724370992, 1.3324146939074761, 1.2143537312782513, 1.0456944370619754,
0.9950972195945285, 1.450482005498162, 1.3155546372833895, 1.197493125233913,
1.1131636307527082, 1.1468968998094646, 1.2818271067664828, 1.1806312982663616,
1.130033836620238, 0.8939081721930029, 1.1300358817563743, 1.349297796211604,
1.1806366400446853, 0.9613764197021512, 0.9276447531760745, 1.0794421441233204,
1.2481061085594132, 1.1975082958980805, 1.467370648158511, 1.1975104631671973,
1.2481106262471495, 1.3661763812109418, 1.0794489817109465, 1.3661788537471289,
1.2143823022387368, 1.3155820179110709, 1.2481174028400754, 1.1637861990172635,
1.130054288314721, 1.4673852544767516, 1.6023186818636428, 1.062591245318848,
1.264990722472894, 1.298724983645655, 1.1300604239674217, 1.1131948578277173,
1.146929073217702, 1.3493295424701581, 1.1300645144395707, 1.3155986849381243,
1.2481332151764104, 0.9445340986497684, 1.0794685181529147, 1.2987367361641462,
1.3999382681623194, 1.3830727936215095, 0.8096043191446259, 1.1300737181101779,
1.2818758253277904, 1.3830777999706727, 1.1806772391396814, 1.130077808678569,
1.315614161841387, 1.3999496698482268, 1.2312821491578154, 1.4168190976004806,
1.4168203797340149, 1.062616246402402, 1.4505568236272839, 1.2987555406361617,
1.2818897455969416, 1.3999598048360586, 1.2481580639430037
correlation for ('she', 'his')
0.05582352995259124, 0.669882965582677, 0.7536190182022755, 0.9769144260883175,
```

0.614061052040338, 0.8373567377499689, 0.7536217459016723, 0.8931821366869412, 0.6419752416463641, 0.5861518379804718, 1.0048326314944196, 0.8931853695451089, 0.5582413611020671, 0.5582418662393553, 0.8373635570663367, 0.7815400271233442, 0.6698920579880463, 1.088575579255418, 1.200225442669945, 0.9211040801857664, 1.0048417240067462, 1.1164918147411436, 1.2002297869095886, 0.976932105909406, 0.7815463919338069, 1.032758666722523, 0.9211099146270338, 0.9769356419504153, 0.6698993320900574, 0.7536374305564002, 0.8932007259410015, 1.0606768218470526, 0.8652897692333679, 1.0048535445187794, 0.9769418300837757, 1.088593310007424, 1.0327689466012078, 1.004857181655355, 1.172334439434215, 1.2281610002854089, 0.8652960332021398, 0.8652968162046131, 1.2560771601413758, 1.1444268926053627, 1.116515051897465, 1.1444289637863145, 1.1165170725654652, 1.1723439870473185, 1.0606931385787406, 0.921129085454473, 1.2002601974676634, 1.1444351773741515, 0.9211315860560141, 1.060697937713011, 0.8932201242484855, 1.2560919363697254, 1.2560930730170752, 0.9769621630739302, 1.3956614959088285, 1.032790441554317, 1.116531217446262, 1.228185450589465, 1.1165332381727757, 0.9211407550444896, 1.0886218774339385, 1.0607094558124472, 1.1165372796477457, 1.0886248327705332, 1.2002797479259177, 1.3119348651584792, 1.2561089862959705, 1.2002830063975474, 1.256111259654442, 1.1444579611068288, 1.2840271670882786, 0.9769780764392273, 1.3398568601558092, 1.31194436277936, 1.0048944638233834, 0.9211540920808556, 1.2282065675446023, 1.032811002781503, 0.8932427566721652, 1.4515207931156124, 0.949072146650103, 1.0607286532007394, 1.1165574874614683, 1.0886445354245713, 0.8374196312123035, 1.0607324927617852, 0.8374211468339918, 1.144476603017421, 1.1444776386969275, 0.9490798763137192, 1.4515352420293006, 0.8653391004426221, 1.20031016101583, 1.1723970321743473, 1.2561408140634114, 0.7536851704795334



d)

The plot shows that the less context 2 words have the lower the correlation is, down to about 1. Further you can observe that a low correlation leads to fewer differences between the values. On the other side the words with high dependency like he and his have a lot of fluctuations.

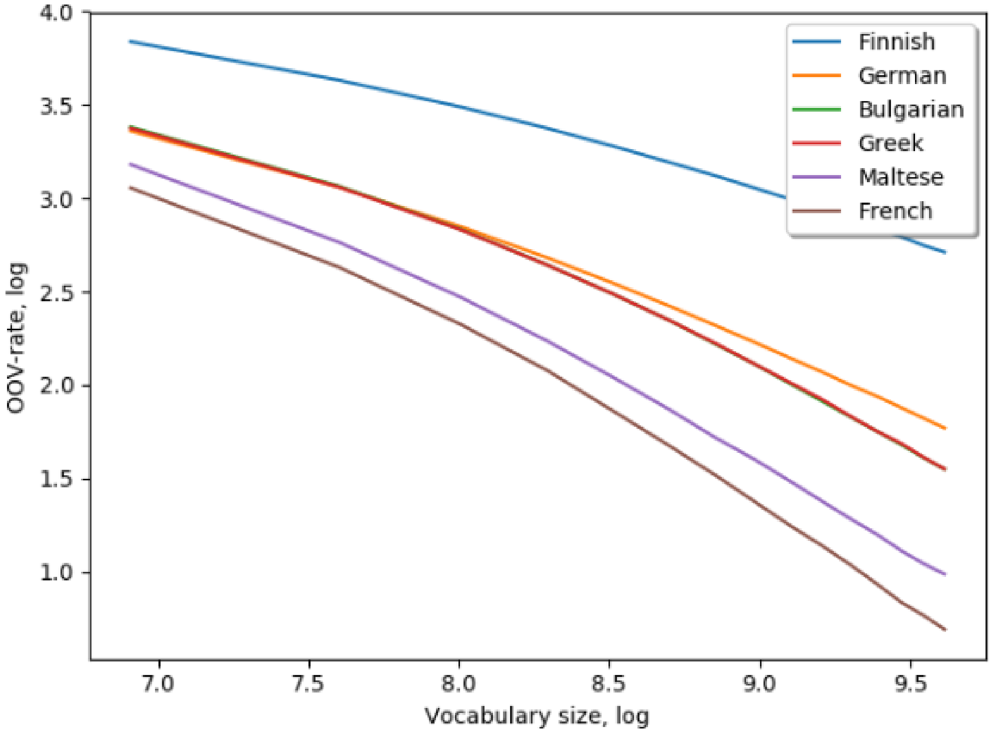


e)

same observations as in d.

2.2 Out-of-vocabulary (OOV) Words

d) We handled six language corpora and calculated OOV-rates for Finnish, German, Bulgarian, Greek, Maltese and French on different partitions of the vocabulary.



Full vocabulary size of these language corpora are different: Finnish – 67462, German – 42696, Bulgarian – 38319, Greek – 39397, Maltese – 30508 and French – 26651.

Morphologically rich languages have higher OOV rates for the same vocabulary size. We can conclude that Finnish is highly diversified language with the richest morphology. Other languages are less diversified in the following order: German, Bulgarian, Greek, Maltese and French. Moreover, Bulgarian and Greek have the graph lines which are very close to each other.

e) Morphologically rich languages such as Finnish and also German, Bulgarian and Greek are more challenging for statistical modelling than others. The reasons are:

- Larger corpora are necessary for development of the LM.
- The LM itself is heavier. More complex model can’t be constructed because the process will be time- and resources-consuming.
- Morphological diversity introduces some kinds of ambiguities.
- Large amount of out-of-vocabulary words.