EMNLP 2016 Submission number: 28

Supplemental Materials (SimpleScience_28_Supplemental_Materials.pdf)

1. Example Simplifications

Format:
Complex word -> Simple word
Original sentence
Replaced sentence

catechols -> compounds
This pathway includes an iron-dependent extradiol dioxygenase HsaC that cleaves catechols.
This pathway includes an iron-dependent extradiol dioxygenase HsaC that cleaves compounds.

fecundity -> fertility
Finally we show that the transient immune activation that renders mosquitoes resistant to the human malaria parasite has little to no effect on mosquito fitness as a measure of survival or fecundity under laboratory conditions.
Finally we show that the transient immune activation that renders mosquitoes resistant to the human malaria parasite has little to no effect on mosquito fitness as a measure of survival or fertility under laboratory conditions.

enteric -> digestive
Proteus mirabilis is an enteric bacterium that forms biofilms on urinary catheters but in laboratory experiments it can swarm over hard surfaces and form a variety of spatial patterns.
Proteus mirabilis is an digestive bacterium that forms biofilms on urinary catheters but in laboratory experiments it can swarm over hard surfaces and form a variety of spatial patterns.

gastrointestinal -> digestive
For instance the chemical environment that symbiotic Escherichia coli encounter in the human gut relates to health of the gastrointestinal tract gut metabolism immune response and tissue homeostasis.
For instance the chemical environment that symbiotic Escherichia coli encounter in the human gut relates to health of the digestive tract gut metabolism immune response and tissue homeostasis.

apatites -> minerals
The microstructure of vertebrate bones and teeth is controlled by polyproline-rich protein matrices (such as amelogenin) that serve as a scaffold to control the assembly of biological apatites.
The microstructure of vertebrate bones and teeth is controlled by polyproline-rich protein matrices (such as amelogenin) that serve as a scaffold to control the assembly of biological minerals.

copulatory -> sexual
During mating DA is used as a feedback mechanism to adjust the activity of multiple sensory-motor neurons and muscles that promote the rhythmic thrusting of the male copulatory organs against his partner's vulval genitalia.
During mating DA is used as a feedback mechanism to adjust the activity of multiple sensory-motor neurons and muscles that promote the rhythmic thrusting of the male sexual organs against his partner's vulval genitalia.
sporulating -> grown

nitrosation -> oxidation
Dimethylamine undergoes nitrosation under weak acid conditions to give dimethylnitrosamine.
Dimethylamine undergoes oxidation under weak acid conditions to give dimethylnitrosamine.

emanating -> stemming

In vitro experiments have established that the invasive growth of malignant tumors is characterized by the dendritic invasive branches composed of chains of tumor cells emanating from the primary tumor mass.
In vitro experiments have established that the invasive growth of malignant tumors is characterized by the dendritic invasive branches composed of chains of tumor cells stemming from the primary tumor mass.

deteriorates -> worsens

Humans can remember several visual items for a few seconds and recall them; however performance deteriorates surprisingly quickly with the number of items that must be stored.
Humans can remember several visual items for a few seconds and recall them; however performance worsens surprisingly quickly with the number of items that must be stored.

isomaltose -> starch

By resurrecting the ancient genes and proteins using high-confidence predictions from many fungal genome sequences available we show that the very first preduplication enzyme was promiscuous preferring maltose-like substrates but also showing trace activity towards isomaltose-like sugars.
By resurrecting the ancient genes and proteins using high-confidence predictions from many fungal genome sequences available we show that the very first preduplication enzyme was promiscuous preferring maltose-like substrates but also showing trace activity towards starch-like sugars.

attenuated -> reduced

Many cases of drug resistance have been found to be associated with secondary mutations in drug target which lead to the attenuated drug-target interactions.
Many cases of drug resistance have been found to be associated with secondary mutations in drug target which lead to the reduced drug-target interactions.

prognosis -> outcome

With the increasing awareness of heterogeneity in breast cancers better prediction of breast cancer prognosis is much needed early on for more personalized treatment and management.
With the increasing awareness of heterogeneity in breast cancers better prediction of breast cancer outcome is much needed early on for more personalized treatment and management.

hemorrhage -> bleeding

Herein we show that TcdB from the epidemic BI/NAP1/027 strain of C. difficile is more lethal causes more extensive brain hemorrhage and is antigenically variable from TcdB produced by previously studied strains of this pathogen (TcdB003).
Herein we show that TcdB from the epidemic BI/NAP1/027 strain of C. difficile is more lethal causes more extensive brain bleeding and is antigenically variable from TcdB produced by previously studied strains of this pathogen (TcdB003).

sialadenitis -> inflammation

Chronic sclerosing sialadenitis Chronic sclerosing sialadenitis is a chronic -LRB- long-lasting -RRB- inflammatory condition affecting the salivary gland.
Chronic sclerosing inflammation Chronic sclerosing inflammation is a chronic -LRB- long-lasting -RRB- inflammatory condition affecting the salivary gland.

rodenticides -> poison

The events are related to post-war poverty and social stresses and to the ready availability of thallium sulphate rodenticides -LRB- rat poisons -RRB- which could be easily administered to humans in food and drink being virtually tasteless and odourless.
The events are related to post-war poverty and social stresses and to the ready availability of thallium sulphate poison -LRB- rat poisons -RRB- which could be easily administered to humans in food and drink being virtually tasteless and odourless.

macrophages -> cells

Complement Receptor 3 (CR3) and Toll-like Receptor 2 (TLR2) are pattern recognition receptors expressed on the surface of human macrophages.
Complement Receptor 3 (CR3) and Toll-like Receptor 2 (TLR2) are pattern recognition receptors expressed on the surface of human cells.

hydrangea -> flower

Perfumed water was once used but since the nineteenth century this has generally been replaced by sweet hydrangea tea known as `` amacha ''.
Perfumed water was once used but since the nineteenth century this has generally been replaced by sweet flower tea known as `` amacha ''.

annular -> spherical

Flowers are 4 5 cm in diameter ; calyx with 6 ovate green lobes 5-10 x 5 9 mm ; there are 6 petals widely obovate 15-25 x 15 19 mm yellow or occasionally white frequently purple at edges and apex ; the staminal ring has 370-510 stamens with filaments 1.5 2 mm long wider at apex yellow with anthers 0.5 mm long ; hood flat 10-20 x 16 20 mm yellow sometimes white with well-developed appendages proximal with anthers distal antherless the pollen of hood turning black with age ; hypanthium usually puberulous sometimes glabrous ; ovary 4-locular with 12-26 ovules in each locule the ovules inserted on lower part of septum the style 1.5 2 mm long with annular expansion towards the tip.
Flowers are 4 5 cm in diameter ; calyx with 6 ovate green lobes 5-10 x 5 9 mm ; there are 6 petals widely obovate 15-25 x 15 19 mm yellow or occasionally white frequently purple at edges and apex ; the staminal ring has 370-510 stamens with filaments 1.5 2 mm long wider at apex yellow with anthers 0.5 mm long ; hood flat 10-20 x 16 20 mm yellow sometimes white with well-developed appendages proximal with anthers distal antherless the pollen of hood turning black with age ; hypanthium usually puberulous sometimes glabrous ; ovary 4-locular with 12-26 ovules in each locule the ovules inserted on lower part of septum the style 1.5 2 mm long with spherical expansion towards the tip.

discernible -> obvious

However these functional sequences are embedded in a background of DNA that serves no discernible function.
However these functional sequences are embedded in a background of DNA that serves no obvious function.

2. Submitted Datasets (SimpleScience.zip)

1) SimpleSciGold Raw Dataset (SimpleSciGold.csv)

A set of 293 sentences from PLOS journal abstracts, each containing one (complex) term from the MeSH ontology or consumer health vocabulary set (Vydiswaran et al. 2014), with an average of 21 simplifications per complex term.

2) Rules Generated by Simple Science (for complex words from SimpleSciGold set)

simplifications_cosim04_wiki3000: A text file containing rules generated according to our approach, with α=0.4 (cosine similarity threshold) and k=3000 (simple word in general corpus threshold)

3) Simplification Generation (Grid Search Analysis): F-measure, Potential, and Precision by frequency thresholds

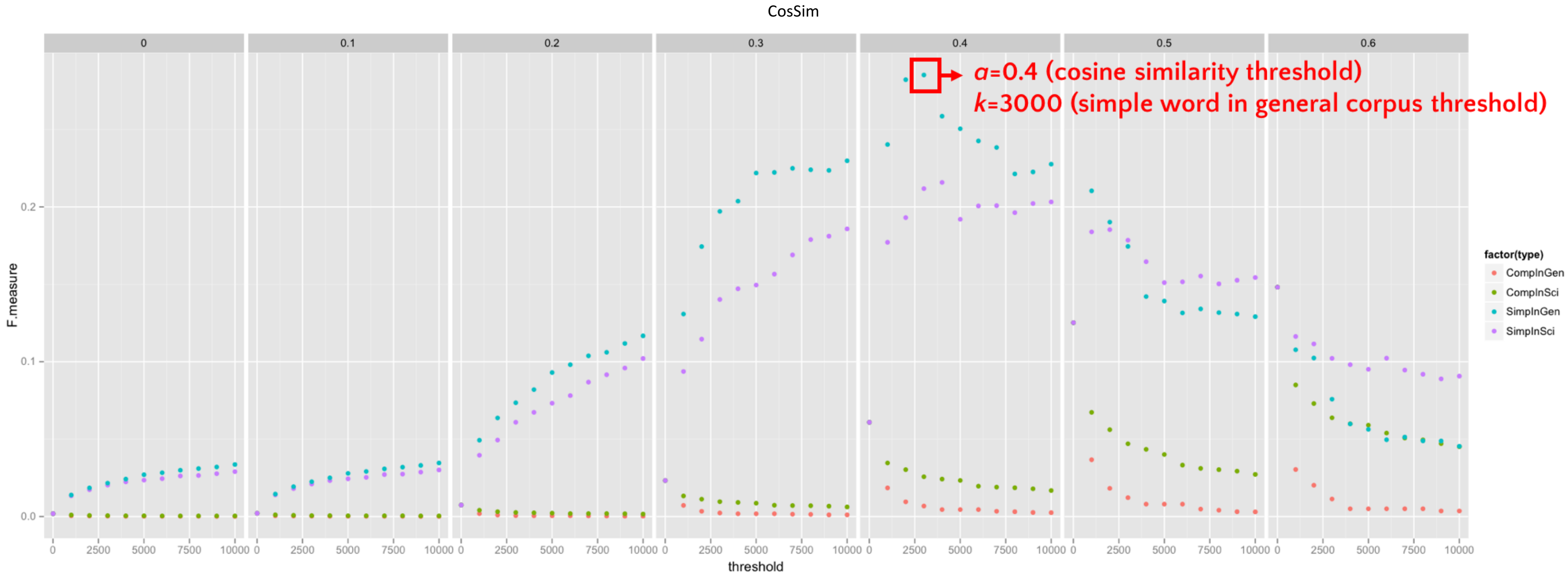
CosSim: The cosine similarity used to filter word2vec results

CompInSci: The number of times the complex word appears in the scientific corpus

CompInGen: The number of times the complex word appears in the general corpus

SimpInSci: The number of times the simple word appears in the scientific corpus

SimpInGen: The number of times the simple word appears in the general corpus



α=0.4 (cosine similarity threshold)
k=3000 (simple word in general corpus threshold)

Fig 1: Summary of grid search results. (F-measure)

3) Simplification Generation (Grid Search Analysis): F-measure, Potential, and Precision by frequency thresholds

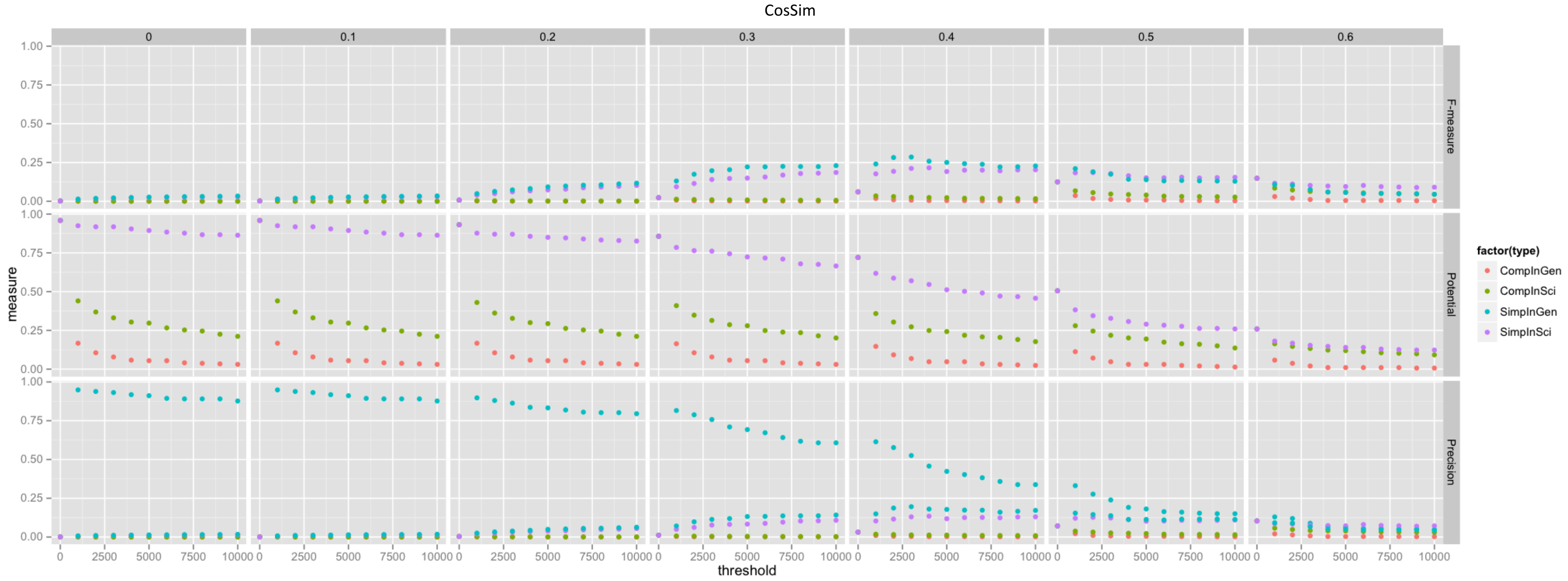


Fig 2: Summary of grid search results. (F-measure, Potential, and Precision)

3) Simplification Generation: Varying threshold on *k* (# times the simple word appears in the general corpus)

	Threshold: 0 or above		
CosSim	**Pot.**	**Prec.**	**F**
0	0.959	0.001	0.002
0.1	0.959	0.001	0.002
0.2	0.932	0.004	0.007
0.3	0.857	0.012	0.023
0.4	0.720	0.032	0.061
0.5	0.505	0.071	0.125
0.6	0.259	0.104	0.148

	1000 or above		
CosSim	**Pot.**	**Prec.**	**F**
0	0.949	0.007	0.014
0.1	0.949	0.007	0.015
0.2	0.898	0.025	0.049
0.3	0.816	0.071	0.131
0.4	0.614	0.149	0.240
0.5	0.331	0.154	0.210
0.6	0.130	0.092	0.108

	2000 or above		
CosSim	**Pot.**	**Prec.**	**F**
0	0.939	0.009	0.019
0.1	0.939	0.010	0.019
0.2	0.881	0.033	0.064
0.3	0.788	0.098	0.174
0.4	0.577	0.187	0.282
0.5	0.276	0.145	0.190
0.6	0.119	0.090	0.102

	3000 or above		
CosSim	**Pot.**	**Prec.**	**F**
0	0.932	0.011	0.022
0.1	0.932	0.011	0.022
0.2	0.863	0.038	0.073
0.3	0.758	0.113	0.197
0.4	0.526	0.196	0.285
0.5	0.239	0.137	0.175
0.6	0.089	0.066	0.076

	4000 or above		
CosSim	**Pot.**	**Prec.**	**F**
0	0.918	0.012	0.024
0.1	0.918	0.013	0.025
0.2	0.836	0.043	0.082
0.3	0.710	0.119	0.204
0.4	0.457	0.180	0.259
0.5	0.191	0.113	0.142
0.6	0.068	0.053	0.060

	5000 or above		
CosSim	**Pot.**	**Prec.**	**F**
0	0.911	0.014	0.027
0.1	0.911	0.014	0.028
0.2	0.833	0.049	0.093
0.3	0.693	0.132	0.222
0.4	0.423	0.178	0.250
0.5	0.181	0.113	0.139
0.6	0.061	0.052	0.056

	6000 or above		
CosSim	Pot.	Prec.	F
0	0.894	0.014	0.028
0.1	0.894	0.015	0.029
0.2	0.819	0.052	0.098
0.3	0.672	0.133	0.222
0.4	0.403	0.174	0.243
0.5	0.164	0.110	0.132
0.6	0.055	0.045	0.050

	7000 or above		
CosSim	Pot.	Prec.	F
0	0.891	0.015	0.030
0.1	0.891	0.016	0.031
0.2	0.805	0.055	0.104
0.3	0.642	0.136	0.225
0.4	0.382	0.173	0.238
0.5	0.160	0.115	0.134
0.6	0.055	0.048	0.051

	8000 or above		
CosSim	Pot.	Prec.	F
0	0.891	0.016	0.031
0.1	0.891	0.016	0.032
0.2	0.802	0.057	0.106
0.3	0.618	0.137	0.224
0.4	0.358	0.160	0.221
0.5	0.154	0.115	0.132
0.6	0.051	0.047	0.049

	9000 or above		
CosSim	Pot.	Prec.	F
0	0.891	0.016	0.032
0.1	0.891	0.017	0.033
0.2	0.802	0.060	0.112
0.3	0.608	0.137	0.224
0.4	0.338	0.166	0.223
0.5	0.150	0.116	0.131
0.6	0.051	0.047	0.049

	10000 or above		
CosSim	Pot.	Prec.	F
0	0.877	0.017	0.034
0.1	0.877	0.018	0.035
0.2	0.795	0.063	0.117
0.3	0.608	0.142	0.230
0.4	0.338	0.172	0.228
0.5	0.150	0.113	0.129
0.6	0.048	0.043	0.045

3) Simplification Generation: Varying threshold on # times the simple word appears in the general corpus

	Threshold: 0 or above		
CosSim	Pot.	Prec.	F
0	0.959	0.001	0.002
0.1	0.959	0.001	0.002
0.2	0.932	0.004	0.007
0.3	0.857	0.012	0.023
0.4	0.720	0.032	0.061
0.5	0.505	0.071	0.125
0.6	0.259	0.104	0.148

	1000 or above		
CosSim	Pot.	Prec.	F
0	0.918	0.009	0.017
0.1	0.918	0.009	0.018
0.2	0.870	0.025	0.049
0.3	0.765	0.062	0.115
0.4	0.587	0.116	0.193
0.5	0.345	0.127	0.185
0.6	0.167	0.084	0.111

	2000 or above		
CosSim	Pot.	Prec.	F
0	0.212	0.000	0.000
0.1	0.212	0.000	0.000
0.2	0.212	0.001	0.002
0.3	0.201	0.003	0.006
0.4	0.177	0.009	0.017
0.5	0.137	0.015	0.027
0.6	0.092	0.030	0.045

	3000 or above		
CosSim	Pot.	Prec.	F
0	0.918	0.010	0.020
0.1	0.918	0.011	0.021
0.2	0.870	0.032	0.061
0.3	0.761	0.077	0.140
0.4	0.570	0.130	0.212
0.5	0.328	0.123	0.178
0.6	0.154	0.077	0.102

	4000 or above		
CosSim	Pot.	Prec.	F
0	0.904	0.011	0.022
0.1	0.904	0.012	0.023
0.2	0.857	0.035	0.067
0.3	0.744	0.082	0.147
0.4	0.546	0.134	0.216
0.5	0.307	0.112	0.165
0.6	0.147	0.074	0.098

	5000 or above		
CosSim	Pot.	Prec.	F
0	0.894	0.012	0.024
0.1	0.894	0.012	0.024
0.2	0.850	0.038	0.073
0.3	0.724	0.083	0.150
0.4	0.512	0.118	0.192
0.5	0.290	0.102	0.151
0.6	0.140	0.072	0.095

	6000 or above		
CosSim	Pot.	Prec.	F
0	0.884	0.012	0.025
0.1	0.884	0.013	0.025
0.2	0.846	0.041	0.078
0.3	0.717	0.088	0.157
0.4	0.502	0.125	0.201
0.5	0.283	0.103	0.152
0.6	0.140	0.081	0.102

	7000 or above		
CosSim	Pot.	Prec.	F
0	0.877	0.013	0.026
0.1	0.877	0.014	0.027
0.2	0.840	0.046	0.087
0.3	0.710	0.096	0.169
0.4	0.491	0.126	0.201
0.5	0.276	0.108	0.155
0.6	0.130	0.074	0.095

	8000 or above		
CosSim	Pot.	Prec.	F
0	0.867	0.013	0.026
0.1	0.867	0.014	0.027
0.2	0.833	0.048	0.092
0.3	0.679	0.103	0.179
0.4	0.471	0.124	0.196
0.5	0.263	0.105	0.150
0.6	0.126	0.072	0.092

	9000 or above		
CosSim	Pot.	Prec.	F
0	0.867	0.014	0.028
0.1	0.867	0.015	0.029
0.2	0.829	0.051	0.096
0.3	0.676	0.105	0.181
0.4	0.468	0.129	0.202
0.5	0.263	0.108	0.153
0.6	0.123	0.070	0.089

	10000 or above		
CosSim	Pot.	Prec.	F
0	0.863	0.015	0.029
0.1	0.863	0.015	0.030
0.2	0.826	0.054	0.102
0.3	0.666	0.108	0.186
0.4	0.457	0.131	0.203
0.5	0.259	0.110	0.154
0.6	0.123	0.072	0.091

3) Simplification Generation: Varying threshold on # times complex word appears in general corpus

	Threshold: 0 or above		
CosSim	Pot.	Prec.	F
0	0.959	0.001	0.002
0.1	0.959	0.001	0.002
0.2	0.932	0.004	0.007
0.3	0.857	0.012	0.023
0.4	0.720	0.032	0.061
0.5	0.505	0.071	0.125
0.6	0.259	0.104	0.148

	1000 or above		
CosSim	Pot.	Prec.	F
0	0.167	0.000	0.000
0.1	0.167	0.000	0.001
0.2	0.167	0.001	0.002
0.3	0.164	0.004	0.007
0.4	0.147	0.010	0.018
0.5	0.113	0.022	0.037
0.6	0.058	0.021	0.030

	2000 or above		
CosSim	Pot.	Prec.	F
0	0.106	0.000	0.000
0.1	0.106	0.000	0.000
0.2	0.106	0.001	0.001
0.3	0.106	0.002	0.003
0.4	0.092	0.005	0.009
0.5	0.072	0.010	0.018
0.6	0.038	0.014	0.020

	3000 or above		
CosSim	Pot.	Prec.	F
0	0.078	7.672	0.000
0.1	0.078	8.222	0.000
0.2	0.078	0.000	0.001
0.3	0.078	0.001	0.002
0.4	0.068	0.004	0.007
0.5	0.048	0.007	0.012
0.6	0.020	0.008	0.011

	4000 or above		
CosSim	Pot.	Prec.	F
0	0.058	6.622	0.000
0.1	0.058	7.074	0.000
0.2	0.058	0.000	0.000
0.3	0.058	0.001	0.002
0.4	0.048	0.002	0.004
0.5	0.031	0.005	0.008
0.6	0.010	0.003	0.005

	5000 or above		
CosSim	Pot.	Prec.	F
0	0.055	6.461	0.000
0.1	0.055	6.914	0.000
0.2	0.055	0.000	0.000
0.3	0.055	0.001	0.002
0.4	0.048	0.002	0.004
0.5	0.031	0.005	0.008
0.6	0.010	0.003	0.005

	6000 or above		
CosSim	Pot.	Prec.	F
0	0.055	6.461	0.000
0.1	0.055	6.914	0.000
0.2	0.055	0.000	0.000
0.3	0.055	0.001	0.002
0.4	0.048	0.002	0.004
0.5	0.031	0.005	0.008
0.6	0.010	0.003	0.005

	7000 or above		
CosSim	Pot.	Prec.	F
0	0.041	4.552	0.000
0.1	0.041	5.004	0.000
0.2	0.041	0.000	0.000
0.3	0.041	0.001	0.001
0.4	0.034	0.002	0.003
0.5	0.024	0.003	0.005
0.6	0.010	0.003	0.005

	8000 or above		
CosSim	Pot.	Prec.	F
0	0.038	4.455	0.000
0.1	0.038	4.905	0.000
0.2	0.038	0.000	0.000
0.3	0.038	0.001	0.001
0.4	0.031	0.002	0.003
0.5	0.020	0.002	0.004
0.6	0.010	0.003	0.005

	9000 or above		
CosSim	Pot.	Prec.	F
0	0.034	1.718	0.000
0.1	0.034	2.149	0.000
0.2	0.034	0.000	0.000
0.3	0.034	0.001	0.001
0.4	0.027	0.001	0.003
0.5	0.017	0.002	0.003
0.6	0.007	0.002	0.004

	10000 or above		
CosSim	Pot.	Prec.	F
0	0.031	1.665	0.000
0.1	0.031	2.097	0.000
0.2	0.031	0.000	0.000
0.3	0.031	0.001	0.001
0.4	0.024	0.001	0.003
0.5	0.014	0.002	0.003
0.6	0.007	0.002	0.004

3) Simplification Generation: Varying threshold on the # times complex word appears in scientific corpus

	Threshold: 0 or above		
CosSim	Pot.	Prec.	F
0	0.959	0.001	0.002
0.1	0.959	0.001	0.002
0.2	0.932	0.004	0.007
0.3	0.857	0.012	0.023
0.4	0.720	0.032	0.061
0.5	0.505	0.071	0.125
0.6	0.259	0.104	0.148

	1000 or above		
CosSim	Pot.	Prec.	F
0	0.440	0.000	0.001
0.1	0.440	0.001	0.001
0.2	0.430	0.002	0.004
0.3	0.410	0.007	0.013
0.4	0.358	0.018	0.035
0.5	0.280	0.038	0.067
0.6	0.164	0.057	0.085

	2000 or above		
CosSim	Pot.	Prec.	F
0	0.369	0.000	0.001
0.1	0.369	0.000	0.001
0.2	0.362	0.002	0.003
0.3	0.348	0.006	0.011
0.4	0.304	0.016	0.030
0.5	0.246	0.032	0.056
0.6	0.147	0.049	0.073

	3000 or above		
CosSim	Pot.	Prec.	F
0	0.331	0.000	0.001
0.1	0.331	0.000	0.001
0.2	0.328	0.001	0.003
0.3	0.314	0.005	0.010
0.4	0.273	0.013	0.026
0.5	0.218	0.026	0.047
0.6	0.133	0.042	0.064

	4000 or above		
CosSim	Pot.	Prec.	F
0	0.304	0.000	0.001
0.1	0.304	0.000	0.001
0.2	0.300	0.001	0.002
0.3	0.287	0.005	0.009
0.4	0.249	0.013	0.024
0.5	0.201	0.024	0.043
0.6	0.123	0.040	0.060

	5000 or above		
CosSim	Pot.	Prec.	F
0	0.297	0.000	0.000
0.1	0.297	0.000	0.001
0.2	0.294	0.001	0.002
0.3	0.280	0.004	0.009
0.4	0.242	0.012	0.023
0.5	0.195	0.022	0.040
0.6	0.119	0.039	0.059

	6000 or above		
CosSim	Pot.	Prec.	F
0	0.266	0.000	0.000
0.1	0.266	0.000	0.000
0.2	0.263	0.001	0.002
0.3	0.249	0.004	0.007
0.4	0.218	0.010	0.020
0.5	0.174	0.018	0.033
0.6	0.113	0.035	0.054

	7000 or above		
CosSim	Pot.	Prec.	F
0	0.253	0.000	0.000
0.1	0.253	0.000	0.000
0.2	0.253	0.001	0.002
0.3	0.239	0.004	0.007
0.4	0.208	0.010	0.019
0.5	0.164	0.017	0.031
0.6	0.106	0.033	0.051

	8000 or above		
CosSim	Pot.	Prec.	F
0	0.246	0.000	0.000
0.1	0.246	0.000	0.000
0.2	0.246	0.001	0.002
0.3	0.235	0.004	0.007
0.4	0.205	0.010	0.019
0.5	0.160	0.017	0.030
0.6	0.102	0.033	0.049

	9000 or above		
CosSim	Pot.	Prec.	F
0	0.225	0.000	0.000
0.1	0.225	0.000	0.000
0.2	0.225	0.001	0.002
0.3	0.215	0.003	0.007
0.4	0.191	0.009	0.018
0.5	0.150	0.016	0.029
0.6	0.099	0.031	0.047

	10000 or above		
CosSim	Pot.	Prec.	F
0	0.925	0.007	0.013
0.1	0.925	0.007	0.014
0.2	0.877	0.020	0.040
0.3	0.785	0.050	0.094
0.4	0.618	0.103	0.177
0.5	0.382	0.121	0.184
0.6	0.181	0.086	0.116

4) Applying Simplifications:

Varying context similarity threshold using distributed representations (word2vec). Context similarity is the cosine similarity between the distributed word representations of the simple word in the pair and the sum of the distributed word representations for all words in a window of length l surrounding the complex word in the sentence (Paetzold and Specia 2015)

	Window: 1		
Context Simil.	Pot.	Prec.	F
0	0.526	0.196	0.285
0.1	0.290	0.120	0.169
0.2	0.191	0.079	0.111
0.3	0.126	0.058	0.080
0.4	0.055	0.026	0.035
0.5	0.024	0.009	0.013
0.6	0.010	0.008	0.009

	Window: 2		
Context Simil.	Pot.	Prec.	F
0	0.526	0.196	0.285
0.1	0.379	0.146	0.211
0.2	0.270	0.106	0.152
0.3	0.160	0.071	0.099
0.4	0.075	0.039	0.051
0.5	0.038	0.028	0.032
0.6	0.010	0.010	0.010

	Window: 3		
Context Simil.	Pot.	Prec.	F
0	0.526	0.196	0.285
0.1	0.413	0.158	0.229
0.2	0.297	0.111	0.161
0.3	0.191	0.082	0.115
0.4	0.096	0.051	0.066
0.5	0.041	0.023	0.029
0.6	0.020	0.014	0.017

	Window: 4		
Context Simil.	Pot.	Prec.	F
0	0.526	0.196	0.285
0.1	0.427	0.157	0.230
0.2	0.307	0.118	0.170
0.3	0.201	0.090	0.124
0.4	0.116	0.057	0.077
0.5	0.041	0.023	0.030
0.6	0.014	0.009	0.011

	Window: whole sentence		
Context Simil.	Pot.	Prec.	F
0	0.526	0.196	0.285
0.1	0.457	0.179	0.257
0.2	0.362	0.149	0.211
0.3	0.201	0.083	0.118
0.4	0.116	0.063	0.082
0.5	0.031	0.018	0.023
0.6	0.017	0.013	0.015

4) Applying Simplification: co-occurrence matrix as context vector (Biran et al. 2011)

Varying context similarity threshold using co-occurrence matrices defined over both corpora (adapted from Biran et al. 2011, who only define co-occurrence matrices over the complex corpus). Context similarity is the cosine similarity between a co-occurrence matrix representing the minimum occurrence of words across the matrices for both the complex and simple word in the rule, and the co-occurrence matrix for the complex word defined only on the input sentence (Biran et al. 2011)

	Window: 10 (Biran et al, 2011)		
Context Similarity	**Pot.**	**Prec.**	**F**
0	0.526	0.196	0.285
0.1	0.481	0.157	0.237
0.2	0.392	0.136	0.202
0.3	0.215	0.071	0.106
0.4	0.085	0.022	0.035
0.5	0.020	0.009	0.012
0.6	0.000	0.000	0.000

	Window: whole sentence		
Context Similarity	**Pot.**	**Prec.**	**F**
0	0.526	0.196	0.285
0.1	0.485	0.156	0.237
0.2	0.427	0.146	0.218
0.3	0.273	0.088	0.133
0.4	0.119	0.029	0.047
0.5	0.024	0.010	0.014
0.6	0.003	0.001	0.002

4) Applying Simplification: Top-k Results (Cosine similarity: 0.4, simple word in general corpus: 3000 or above)

	Ranked by Cosine Similarity		
Top K	**Pot.**	**Prec.**	**F**
k=1	0.389	0.389	0.389
k=2	0.485	0.264	0.342
k=3	0.544	0.199	0.292
k=4	0.552	0.145	0.230
k=5	0.586	0.119	0.198

	Ranked by the Count of Simple Word in General corpus		
Top K	**Pot.**	**Prec.**	**F**
k=1	0.314	0.314	0.314
k=2	0.423	0.232	0.300
k=3	0.490	0.166	0.248
k=4	0.536	0.131	0.210
k=5	0.561	0.100	0.169

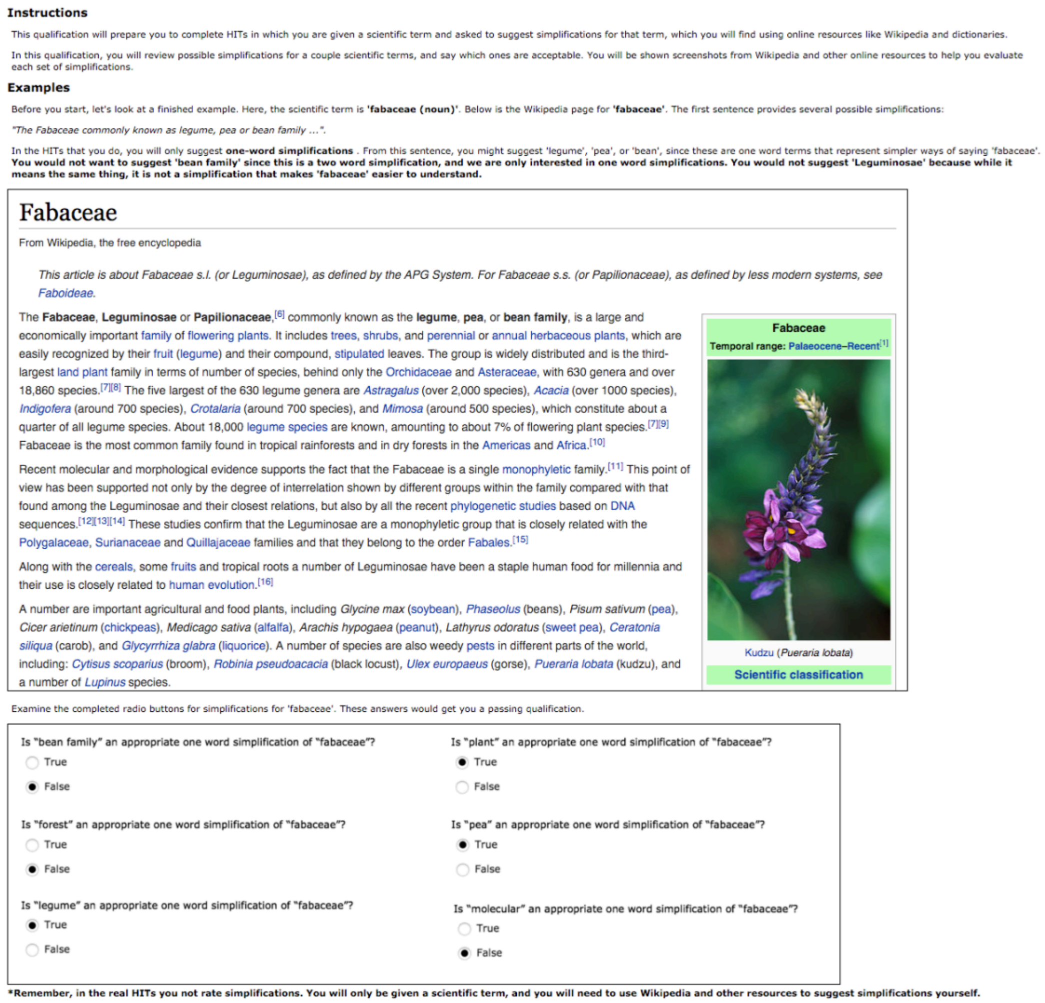
4) Applying Simplification: Top-k Results (No threshold)

	Ranked by Cosine Similarity		
Top K	**Pot.**	**Prec.**	**F**
k=1	0.186	0.186	0.186
k=2	0.251	0.143	0.182
k=3	0.292	0.125	0.175
k=4	0.326	0.115	0.170
k=5	0.351	0.105	0.162

	Ranked by the Count of Simple Word in General corpus		
Top K	**Pot.**	**Prec.**	**F**
k=1	0.000	0.000	0.000
k=2	0.014	0.007	0.009
k=3	0.021	0.009	0.012
k=4	0.031	0.009	0.014
k=5	0.052	0.012	0.019

5) SimpleSciGold Study Interface

Qualification Interface:



Instructions

This qualification will prepare you to complete HITs in which you are given a scientific term and asked to suggest simplifications for that term, which you will find using online resources like Wikipedia and dictionaries.

In this qualification, you will review possible simplifications for a couple scientific terms, and say which ones are acceptable. You will be shown screenshots from Wikipedia and other online resources to help you evaluate each set of simplifications.

Examples

Before you start, let's look at a finished example. Here, the scientific term is **'fabaceae (noun)'**. Below is the Wikipedia page for **'fabaceae'**. The first sentence provides several possible simplifications:

"The Fabaceae commonly known as legume, pea or bean family ...".

In the HITs that you do, you will only suggest **one-word simplifications** . From this sentence, you might suggest 'legume', 'pea', or 'bean', since these are one word terms that represent simpler ways of saying 'fabaceae'. **You would not want to suggest 'bean family' since this is a two word simplification, and we are only interested in one word simplifications. You would not suggest 'Leguminosae' because while it means the same thing, it is not a simplification that makes 'fabaceae' easier to understand.**

Fabaceae

From Wikipedia, the free encyclopedia

This article is about Fabaceae s.l. (or Leguminosae), as defined by the APG System. For Fabaceae s.s. (or Papilionaceae), as defined by less modern systems, see Faboideae.

The **Fabaceae**, **Leguminosae** or **Papilionaceae**,[6] commonly known as the **legume**, **pea**, or **bean family**, is a large and economically important family of flowering plants. It includes trees, shrubs, and perennial or annual herbaceous plants, which are easily recognized by their fruit (legume) and their compound, stipulated leaves. The group is widely distributed and is the third-largest land plant family in terms of number of species, behind only the Orchidaceae and Asteraceae, with 630 genera and over 18,860 species.[7][8] The five largest of the 630 legume genera are *Astragalus* (over 2,000 species), *Acacia* (over 1000 species), *Indigofera* (around 700 species), *Crotalaria* (around 700 species), and *Mimosa* (around 500 species), which constitute about a quarter of all legume species. About 18,000 legume species are known, amounting to about 7% of flowering plant species.[7][9] Fabaceae is the most common family found in tropical rainforests and in dry forests in the Americas and Africa.[10]

Recent molecular and morphological evidence supports the fact that the Fabaceae is a single monophyletic family.[11] This point of view has been supported not only by the degree of interrelation shown by different groups within the family compared with that found among the Leguminosae and their closest relations, but also by all the recent phylogenetic studies based on DNA sequences.[12][13][14] These studies confirm that the Leguminosae are a monophyletic group that is closely related with the Polygalaceae, Surianaceae and Quillajaceae families and that they belong to the order Fabales.[15]

Along with the cereals, some fruits and tropical roots a number of Leguminosae have been a staple human food for millennia and their use is closely related to human evolution.[16]

A number are important agricultural and food plants, including *Glycine max* (soybean), *Phaseolus* (beans), *Pisum sativum* (pea), *Cicer arietinum* (chickpeas), *Medicago sativa* (alfalfa), *Arachis hypogaea* (peanut), *Lathyrus odoratus* (sweet pea), *Ceratonia siliqua* (carob), and *Glycyrrhiza glabra* (liquorice). A number of species are also weedy pests in different parts of the world, including: *Cytisus scoparius* (broom), *Robinia pseudoacacia* (black locust), *Ulex europaeus* (gorse), *Pueraria lobata* (kudzu), and a number of *Lupinus* species.

Fabaceae

Temporal range: Palaeocene–Recent[1]

Kudzu (*Pueraria lobata*)

Scientific classification

Examine the completed radio buttons for simplifications for 'fabaceae'. These answers would get you a passing qualification.

Is "bean family" an appropriate one word simplification of "fabaceae"?

○ True

● False

Is "forest" an appropriate one word simplification of "fabaceae"?

○ True

● False

Is "legume" an appropriate one word simplification of "fabaceae"?

● True

○ False

Is "plant" an appropriate one word simplification of "fabaceae"?

● True

○ False

Is "pea" an appropriate one word simplification of "fabaceae"?

● True

○ False

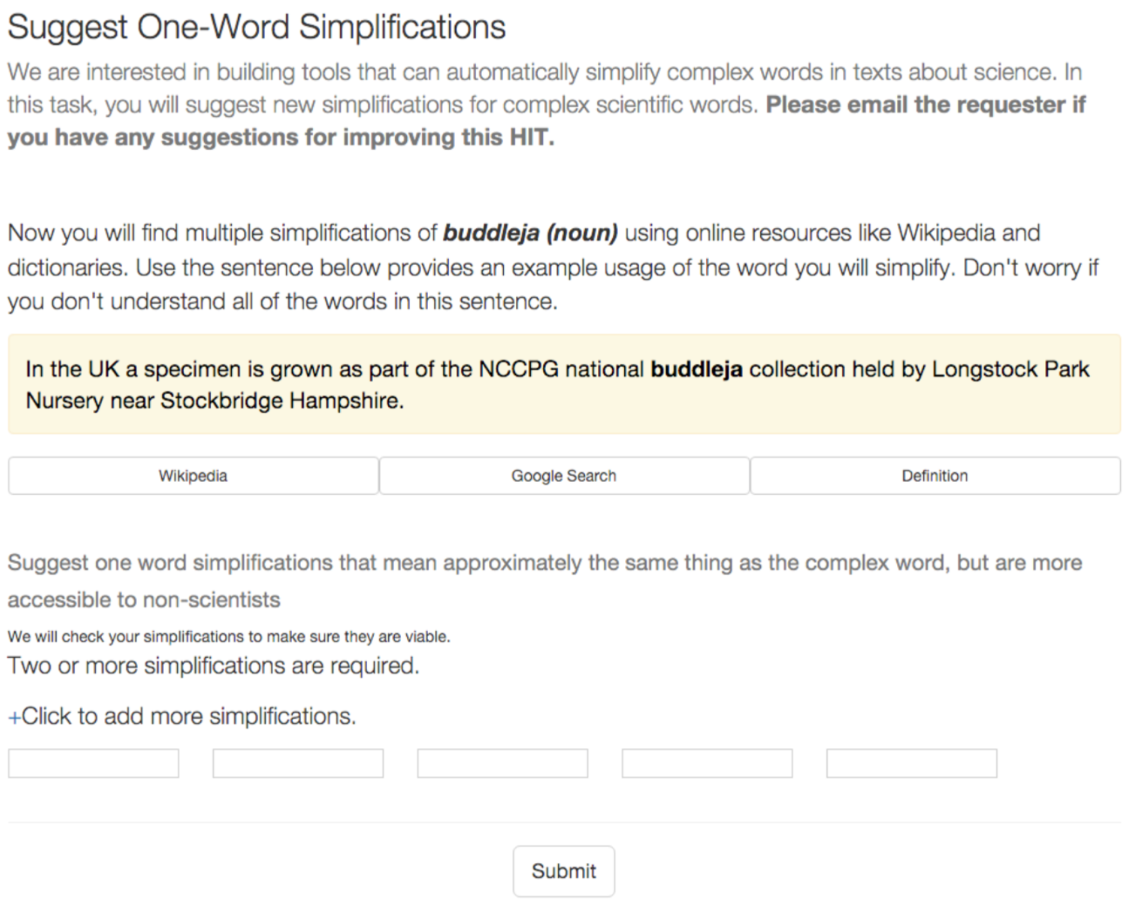
Is "molecular" an appropriate one word simplification of "fabaceae"?

○ True

● False

***Remember, in the real HITs you not rate simplifications. You will only be given a scientific term, and you will need to use Wikipedia and other resources to suggest simplifications yourself.**

Task Interface:



Suggest One-Word Simplifications

We are interested in building tools that can automatically simplify complex words in texts about science. In this task, you will suggest new simplifications for complex scientific words. **Please email the requester if you have any suggestions for improving this HIT.**

Now you will find multiple simplifications of ***buddleja (noun)*** using online resources like Wikipedia and dictionaries. Use the sentence below provides an example usage of the word you will simplify. Don't worry if you don't understand all of the words in this sentence.

In the UK a specimen is grown as part of the NCCPG national **buddleja** collection held by Longstock Park Nursery near Stockbridge Hampshire.

Suggest one word simplifications that mean approximately the same thing as the complex word, but are more accessible to non-scientists

We will check your simplifications to make sure they are viable.

Two or more simplifications are required.

+Click to add more simplifications.