

SUPPLEMENTARY MATERIAL

Common English complexities

- Hyphenated compound words

Table 2: Tokenizers output for sentence (1)

Tokenizer	Output
1, 2, 3, 4, 6, 7, 8, 9, 10, 11	Normal chest x-ray.
5	Normal chest x - ray.
12	Normal chest x- ray.

Table 3: Tokenizers output for sentence (2)

Tokenizer	Output
1, 2, 6, 8, 9, 11, 12	2-year 2-month old female with pneumonia.
3, 4, 5, 7	2-year 2-month old female with pneumonia.
10	2-year 2-month old female with pneumonia.

Table 4: Tokenizers output for sentence (3)

Tokenizer	Output
1, 2, 4, 6, 8, 9, 10, 11, 12	This may occur through the ability of IL-10 to induce expression of the gene.
5, 7	This may occur through the ability of IL - 10 to induce expression of the gene.
3	This may occur through the ability of IL- 10 to induce expression of the gene.

- Words with letters and slashes

Table 5: Tokenizers output for sentence (4)

Tokenizer	Output
2, 6, 8, 9, 11, 12	The maximal effect is observed at the IL-10 concentration of 20 U/ml.
3, 5, 7	The maximal effect is observed at the IL - 10 concentration of 20 U / ml.
1, 4, 10	The maximal effect is observed at the IL-10 concentration of 20 U / ml.

Table 6: Tokenizers output for sentence (5)

Tokenizer	Output
-----------	--------

These results indicate that within the TCR/CD3 signal transduction pathway both PKC and calcineurin are required for the effective activation of the IKK complex and NF-kappaB in T lymphocytes.

These results indicate that within the TCR / CD3 signal transduction pathway both PKC and calcineurin are required for the effective activation of the IKK complex and NF - kappaB in T lymphocytes.

- Words with letters and apostrophes

Table 7: Tokenizers output for sentence (6)

Tokenizer	Output
1, 2, 4, 8, 9, 10, 11, 12	The false positive rate of our predictor was estimated by the method of D'Haeseleer and Church 1855 and used to compare it to other prediction datasets.

The false positive rate of our predictor was estimated by the method of D ' Haeseleer and Church 1855 and used to compare it to other prediction datasets.

Table 8: Tokenizers output for sentence (7)

Tokenizer	Output
1, 2, 4, 6, 8, 9, 10, 11, 12	Small, scarred right kidney, below more than 2 standard deviations in size for patient's age.
3, 5, 7	Small, scarred right kidney, below more than 2 standard deviations in size for patient ' s age.

- Words with letters and brackets

Table 9: Tokenizers output for sentence (8)

Tokenizer	Output
1, 2, 5, 7, 8, 11, 12	Of these, Diap1 has been most extensively characterized: it can block cell death caused by the ectopic expression of reaper, hid, and grim (reviewed in [26]).
6	Of these, Diap1 has been most extensively characterized: it can block cell death caused by the ectopic expression of reaper, hid, and grim (reviewed in [26]).

9	Of these, Diap1 has been most extensively characterized; it can block cell death caused by the ectopic expression of reaper, hid, and grim (reviewed in [26]).
4	Of these, Diap1 has been most extensively characterized; it can block cell death caused by the ectopic expression of reaper, hid, and grim (reviewed in [26]).
10	Of these, Diap1 has been most extensively characterized ; it can block cell death caused by the ectopic expression of reaper, hid, and grim (reviewed in [26]).
3	Of these, Diap1 has been most extensively characterized; it can block cell death caused by the ectopic expression of reaper, hid, and grim (reviewed in [26]).

- **Abbreviations in capital letters and acronyms**

Table 10: Tokenizers output for sentence (9)

Tokenizer	Output
4, 6, 8, 11	Mutants in Toll signaling pathway were obtained from Dr. S. Govind : cactE8, cactIII ^G and cactD13 mutations in the cact gene on Chromosome II.
9	Mutants in Toll signaling pathway were obtained from Dr. S. Govind: cactE8, cactIII^G and cactD13 mutations in the cact gene on Chromosome II.
2, 5, 7	Mutants in Toll signaling pathway were obtained from Dr. S. Govind : cactE8, cactIII ^G and cactD13 mutations in the cact gene on Chromosome II.
1	Mutants in Toll signaling pathway were obtained from Dr. S. Govind : cactE8, cactIII^G and cactD13 mutations in the cact gene on Chromosome II .
10	Mutants in Toll signaling pathway were obtained from Dr. S. Govind : cactE8, cactIII^G and cactD13 mutations in the cact gene on Chromosome II.
3	Mutants in Toll signaling pathway were obtained from Dr. S. Govind : cactE8 , cactIII ^G and cactD13 mutations in the cact gene on Chromosome II.

12	Mutants in Toll signaling pathway were obtained from Dr. S. Govind: cactE8, cactIII^G and cactD13 mutations in the cact gene on Chromosome II .
----	---

Table 11: Tokenizers output for sentence (10)

Tokenizer	Output
2, 6, 8, 9, 12	The transcripts were detected in all the CD4⁻CD8⁻, CD4⁺CD8⁺, CD4⁺CD8⁻ and CD4⁻CD8⁺ cell populations.
1, 3, 4, 7, 10, 11	The transcripts were detected in all the CD4⁻CD8⁻, CD4⁺CD8⁺, CD4⁺CD8⁻ and CD4⁻CD8⁺ cell populations.
5	The transcripts were detected in all the CD4⁻CD8⁻, CD4⁺CD8⁺ and CD4⁺CD8⁻ and CD4⁻CD8⁺ cell populations.

- **Words with letters and periods**

Table 12: Tokenizers output for sentence (11)

Tokenizer	Output
2, 8	Two stop codons of an iORF (i.e. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).
1, 6, 11, 12	Two stop codons of an iORF (i.e. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).
9	Two stop codons of an iORF (i.e. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).
4, 7	Two stop codons of an iORF (i.e. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).
5	Two stop codons of an iORF (i.e. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).
10	Two stop codons of an iORF (i.e. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).

3	Two stop codons of an iORF (i.e. the inframe and C-terminal stops) can be any combination of canonical stop codons (TAA, TAG, TGA).
---	---

- Words with letters and numbers

Table 13: Tokenizers output for sentence (12)

Tokenizer	Output
1, 2, 4, 5, 6, 7, 8, 9, 11, 12	Selenocysteine and pyrrolysine are the 21st and 22nd amino acids which are genetically encoded by stop codons.
10	Selenocysteine and pyrrolysine are the 21st and 22nd amino acids which are genetically enc_oded by stop codons.
3	Selenocysteine and pyrrolysine are the 21st and 22nd amino acids, which are genetically encoded by stop codons.

- Words with numbers and one type of punctuation

Table 14: Tokenizers output for sentence (13)

Tokenizer	Output
1, 5, 6, 8, 9, 10, 11, 12	A total of 26,003 iORF satisfied the above criteria.
2, 3, 4, 7	A total of 26⁰⁰³ iORF satisfied the above criteria.

Table 15: Tokenizers output for sentence (14)

Tokenizer	Output
1, 2, 6, 8, 9, 11, 12	The patient had prior x-ray on 1/2 which demonstrated no pneumonia.
4, 5, 7	The patient had prior x-ray on 1/2 which demonstrated no pneumonia.
3, 10	The patient had prior x-ray on 1/2 which demonstrated no pneumonia.

Table 16: Tokenizers output for sentence (15)

Tokenizer	Output
3, 4, 5, 6, 7, 8, 9, 10, 11,	Indeed, it has been estimated recently that the current yeast and human protein interaction maps are only 50% and 10% complete, respectively 18 .
1	Indeed, it has been estimated recently that the current yeast and human protein interaction maps

	are only 50% and 10% complete, respectively 18 .
12	Indeed, it has been estimated recently that the current yeast and human protein interaction maps are only 50% and 10% complete, respectively 18 .

Table 17: Tokenizers output for sentence (16)

Tokenizer	Output
1, 2, 4, 5, 6, 8, 9, 10, 11, 12	The dotted line indicates significance level 0.05 after a correction for multiple testing.
3, 7	The dotted line indicates significance level 0.05 after a correction for multiple testing.

Table 18: Tokenizers output for sentence (17)

Tokenizer	Output
1, 2, 8, 9, 10, 11, 12	E-selectin is induced within 1-2h , peaks at 4-6h , and gradually returns to basal level by 24h .
6	E-selectin is induced within 1-2h , peaks at 4-6h , and gradually returns to basal level by 24h .
4, 7	E-selectin is induced within 1-2h , peaks at 4-6h , and gradually returns to basal level by 24h .
5	E-selectin is induced within 1-2h , peaks at 4-6h , and gradually returns to basal level by 24h .
3	E-selectin is induced within 1-2h , peaks at 4-6h , and gradually returns to basal level by 24h .

- Numeration

Table 19: Tokenizers output for sentence (18)

Tokenizer	Output
1, 2, 3, 5, 7, 8, 9, 10, 11, 12	1 . Bioactivation of sulphamethoxazole (SMX) to chemically-reactive metabolites and subsequent protein conjugation is thought to be involved in SMX hypersensitivity.
4, 6	1 . Bioactivation of sulphamethoxazole (SMX) to chemically-reactive metabolites and subsequent protein conjugation is thought to be involved in SMX hypersensitivity.

- A hypertext markup symbol

Table 20: Tokenizers output for sentence (19)

Tokenizer	Output
2, 4, 5, 8	Bcd_mRNA_transcripts_of_<_or_=_2.6_kb_were_selectively_expressed_in_PBL_and_testis_of_healthy_individuals_.
6	Bcd_mRNA_transcripts_of_<_or_=_2.6kb_were_selectively_expressed_in_PBL_and_testis_of_healthy_individuals_.
9, 12	Bcd_mRNA_transcripts_of_<_or_=_2.6_kb_were_selectively_expressed_in_PBL_and_testis_of_healthy_individuals_.
3, 7	Bcd_mRNA_transcripts_of_<_or_=_2_6_kb_were_selectively_expressed_in_PBL_and_testis_of_healthy_individuals_.
11	Bcd_mRNA_transcripts_of_<_or_=_2.6_kb_were_selectively_expressed_in_PBL_and_testis_of_healthy_individuals_.
1	Bcd_mRNA_transcripts_of_or_=_2.6_kb_were_selectively_expressed_in_PBL_and_testis_of_healthy_individuals_.
10	Bc_d_mRNA_transcripts_of_<_or_=_2.6_kb_were_selectively_expressed_in_PBL_and_testis_of_healthy_individuals_.

- A URL

Table 21: Tokenizers output for sentence (20)

Tokenizer	Output
2, 6, 8,	Names_of_all_available_Trace_Databases_were_taken_from_a_list_of_databases_at_http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml
9	Names_of_all_available_Trace_Databases_were_taken_from_a_list_of_databases_at_http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml
3, 5, 7	Names_of_all_available_Trace_Databases_were_taken_from_a_list_of_databases_at_http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml
11, 12	Names_of_all_available_Trace_Databases_were_taken_from_a_list_of_databases_at_http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml

1	Names_of_all_available_Trace_Databases_were_taken_from_a_list_of_databases_at_http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml
4	Names_of_all_available_Trace_Databases_were_taken_from_a_list_of_databases_at_http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml_1
10	Names_of_all_available_Trace_Databases_were_taken_from_a_list_of_databases_at_http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml

Biomedical English complexities

- A DNA sequence

Table 22: Tokenizers output for sentence (21)

Tokenizer	Output
1, 2, 4, 5, 6, 7, 8, 9, 11, 12	Footprinting_analysis_revealed_that_the_identical_sequence_CCGAAACTGAAAAGG_designated_E6_was_protected_by_nuclear_extracts_from_B_cells_T_cells_or_HeLa_cells.
10	Footprinting_analysis_revealed_that_the_identical_sequence_CCGAAACTGAAAAGG_designated_E6_was_protected_by_nuclear_extracts_from_B_cells_T_cells_or_HeLa_cells.
3	Footprinting_analysis_revealed_that_the_identical_sequence_CCGAAACTGAAAAGG_designated_E6_was_protected_by_nuclear_extracts_from_B_cells_T_cells_or_HeLa_cells.

- Temporal expressions

Table 23: Tokenizers output for sentence (22)

Tokenizer	Output
2, 6, 8, 9, 11, 12	This_was_last_documented_on_the_Nuclearv_Cystogram_dated_1/2/01.
1, 3, 4, 7, 10	This_was_last_documented_on_the_Nuclearv_Cystogram_dated_1_2_01.
1	This_was_last_documented_on_the_Nuclearv_Cystogram_dated_1/2/01.

- Chemical substances

Table 24: Tokenizers output for sentence (23)

Tokenizer	Output
6, 8,	These results reveal a central role for CaMKIV/Gr as a Ca(2+)-regulated activator of gene transcription in T lymphocytes.
9	These results reveal a central role for CaMKIV/Gr as a Ca(2+)-regulated activator of gene transcription in T lymphocytes.
1, 3, 4, 7	These results reveal a central role for CaMKIV/Gr as a Ca(2+)-regulated activator of gene transcription in T lymphocytes.
11	These results reveal a central role for CaMKIV/Gr as a Ca(2+)-regulated activator of gene transcription in T lymphocytes.
10	These results reveal a central role for CaMKIV/Gr as a Ca(2+)-regulated activator of gene transcription in T lymphocytes.
2	These results reveal a central role for CaMKIV/Gr as a Ca(2+)-regulated activator of gene transcription in T lymphocytes.
12	These results reveal a central role for CaMKIV/Gr as a Ca(2+)-regulated activator of gene transcription in T lymphocytes.

Table 25: Tokenizers output for sentence (24)

Tokenizer	Output
1, 2, 6, 8, 11	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.
9	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.

7	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.
5	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.
4	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.
10	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.
3	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.
12	Expression of a highly specific protein inhibitor for cyclic AMP-dependent protein kinases in interleukin-1(IL-1)-responsive cells blocked IL-1-induced gene transcription that was driven by the kappa immunoglobulin enhancer or the human immunodeficiency virus long terminal repeat.