Université de Paris

Le Corpus d'Étude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp) source de visualisations linguistiques

•••

Thomas Gaillat - Université Rennes 2
Pascale Janvier - Université Rennes 1
Bénédicte Dumont - Université Rennes 1
Antoine Lafontaine - Université Rennes 1
Anas Kerfati - ENSAI

UNIVERSITÉ DE RENNES 1

UNIVERSITÉ RENNES 2 LIDILE EA 3874

ENSAI École nationale de la statistique et de l'analyse de l'information

Financé dans le cadre du projet DUNE DESIR Universités de Rennes

Contexte et besoins

Utilité des corpus en LANSAD

- Besoin de définir des objectifs linguistiques et contenus
- Activités linguistiques par exploration par les apprenants

  Mais peu d'usage pour nourrir des systèmes iCALL sauf correction automatique

Besoin : Outils de diagnostic de production et guidage en temps réel pour assister les enseignants

Corpus LANSAD peut aider à positionner les productions d'apprenants.

Question de recherche

Quelle exploitation automatisée d'un corpus d'apprenants pour le guidage linguistique des étudiants de LANSAD ?

Outline

Description du corpus

1. Structure du corpus
2. Collecte (métadonnées)
3. Tâche d'élicitation
4. Annotation CECR et évaluation

Outils compagnons

1. Outil BDD MOODLE
2. VizLing : Outils d'annotation, calcul de complexité linguistique & de visualisation

Les corpus en LANSAD

- Pour l'élaboration de contenus d'enseignement
  - Analyses par les chercheurs et enseignants (Granger 2008)
- Activités pédagogiques en classe Data Driven Learning
  - Analyses exploratoires par les enseignants et les apprenants (Boulton, 2017) (Leńko-Szymańska and Boulton 2015) (Granger, Gilquin, and Meunier 2015)
- Correction grammaticale
  - Analyse fondées sur des apprentissages automatiques à partir de corpus d'apprenants (Leacock 2010)

Notre proposition : Un corpus comme étalon des niveaux par spécialité et en fonction de traits critériés linguistiques (Hawkins and Buttery 2010) (Vajjala 2017)

Description du corpus

Corpus CELVA.Sp

- Domaines : Médecine, Pharmacie, Informatique, Science Physiques de la Matière, Biologie
- 279 apprenants d'anglais; 8 espagnol et 17 allemand
- > 81 000 mots
- Niveaux L1 à M1
- Double mappage sur les niveaux CEFR avec DIALANG écrit et annotation d'experts

Niveaux CECR	A1	A2	B1	B2	C1	C2
Répartition des textes anglais L2	27	63	125	43	19	3

Collecte

- Interface MOODLE

Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp)

Généralités

Corpus d'Etude des Langues Vivantes Appliquées
à une Spécialité (CELVA.Sp)

Collecte

- Interface MOODLE
- Méta-données

Collecte

- Interface MOODLE
- Méta-données

- L1
- L2
- Age
- Sexe
- Domaine_de_specialite

- Acceptation_donnees
- Sejours_duree_semaines
- Sejours_duree_mois
- Sejours_frequence
- Lang_exposition
- Section_renforcee
- Annee_naissance
- Niveau_etudes_actuel
- Note_dialang_ecrit
- Lecture_regularite

CELVA.Sp, A Learner Corpus dedicated to Languages for Specific Purposes.

CELVA.Sp est une collection de textes écrits par des apprenants d'une langue étrangère. Il peut être exploité pour l'analyse linguistique des aspects particuliers de ce type de langue.

Merci de votre participation.

Affichage liste | Affichage fiche | Recherche | Ajouter une fiche | Exporter | Modèles | Champs | Préréglages

Nouvelle fiche

Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité - CELVA.Sp

Préliminaire indispensable : Test de niveau DIALANG à effectuer (45 minutes): ATTENTION: Ne faire que le module écrit et passer les étapes sur le vocabulaire et les listes de compétences.

Questionnaire (10 minutes)

ID_etudiant:

Votre année de naissance :

Votre age :

Sexe :

- F
- M

Votre niveau d'études actuel:

- L1
- L2
- L3
- M1

Collecte

- Interface MOODLE
- Méta-données
- Protection des données

MERCI pour votre contribution. :-)

Corpus recueilli et géré par l'université de Rennes 1.
Copyright Université de Rennes 1 SCELVA.
logo scelva

Contact : **Thomas Gaillat** - thomas.gaillat at univ-rennes1.fr

DONNÉES PERSONNELLES

Les informations recueillies sur ce formulaire sont enregistrées dans un fichier informatisé par **Université de Rennes 1 - SCELVA** pour **constituer un corpus de langue d'apprenants.**

Elles sont conservées pendant 20 ans et sont destinées **à des chercheurs universitaires de l'Union Européenne. Ces données sont anonymisées avant tout traitement par modélisation.**

Conformément à la loi « informatique et libertés » , vous pouvez exercer votre droit d'accès aux données vous concernant et les faire rectifier ou effacer en contactant le **SCELVA : infos-scelva@listes.univ-rennes1.fr**

ENREGISTRER ET AFFICHER | ENREGISTRER ET AJOUTER UNE FICHE

◄ AIDE TEST DIALANG | Aller à… | STUDENT PLOTS ►

Elicitation

Deux tâches écrites

1. Au choix, décrivez une expérience scientifique/découverte/invention/technologie de votre domaine spécialisé
2. Donnez votre opinion sur les conséquences de cette expérience/découverte/invention/technologie:

Durée pour les deux tâches : 45 minutes

Annotation

Annotation en catégories du CECR pour le subset anglais

- 2 expertes examinatrices CLES du SCELVA Université de Rennes 1
- 1 référentiel (Conseil de l'Europe 2018)
- Accord Inter-annotateurs - Weighted kappa sur 50 textes extraits au hasard

Rater B	Rater A
A1: 9	A1: 8
A2: 8	A2:10
B1:19	B1:16
B2:11	B2:10
C1: 3	C1: 5
	C2: 1

```
> #Weighted kappa
> kappa2(irr, "squared")
 Cohen's Kappa for 2 Raters (Weights:
squared)
 Subjects = 50 Raters = 2 Kappa = 0.714
z = 5.1 p-value = 3.34e-07
```

	Rater A					
Rater B	A1	A2	B1	B2	C1	C2
A1	5	3	1	0	0	0
A2	3	3	2	0	0	0
B1	0	4	9	4	1	1
B2	0	0	4	4	3	0
C1	0	0	0	2	1	0

Format

CELVA.Sp

Format .csv

Ou

.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<celva_sp>
  <writing>
        <document>12004489</document>
        <texte>the VR ( virtual reality) is a technology witch you can
create and explore a virtual world with all the rules you want or decide
to apply. It use a (lookslike a big pair of sunclasses) based on the 3D
vision, each eyes see a differents images at the same to recreate the
sensation of 3D vison. Some point of this technology is not perfect, the
real problems is the different perception between internal ears and your
vision. To be simple your body think you are poisoned, in fact your vision
say at your brain your on movement but your internal detect any of it, and
to figth again the poison you will throw. It&amp;apos;s basicaly like the sea
sike. With this technology, the scientist can experiment some test they
can’t do normaly, they can create a virtual room with the rules of space
to test that they want to test. It’s the practical theory i thinks. Itcan
remplace the real practice but it would help. An other aspect oh this is
in medical. With it a surgeon can opperate a people across the world with
this. He/She can see on a direct time with a hight accuracy (3D) what
his/her doing with the robots, it has been controle by the
surgeon.</texte>
        <CECR.niveau>A2</CECR.niveau>
        <nb_annees_L2>13</nb_annees_L2>
        <L1>French</L1>
        <Domaine_de_specialite>Informatique et
electronique</Domaine_de_specialite>
        <Acceptation_donnees>Oui</Acceptation_donnees>
        <Sejours_duree_semaines>0</Sejours_duree_semaines>
        <Sejours_duree_mois>0</Sejours_duree_mois>
        <Sejours_frequence>0</Sejours_frequence>
        <Lang_exposition>5</Lang_exposition>
        <Section_renforcee>Non</Section_renforcee>
        <Annee_naissance>1993</Annee_naissance>
        <Niveau_etudes_actuel>L3</Niveau_etudes_actuel>
        <Age>25</Age>
        <L2>Anglais</L2>
        <Note_dialang_ecrit>B1</Note_dialang_ecrit>
        <Lecture_regularite>mensuelle</Lecture_regularite>
        <Sexe>M</Sexe>
  </writing>
</celva_sp>
```

Outils compagnons

Annotation linguistique & métriques

VizLing (Gaillat et al. 2019) Programme R d'annotation et d'extraction de traits

- Entrée : CELVA.Sp.csv Sortie : Jeux de données et visualisations

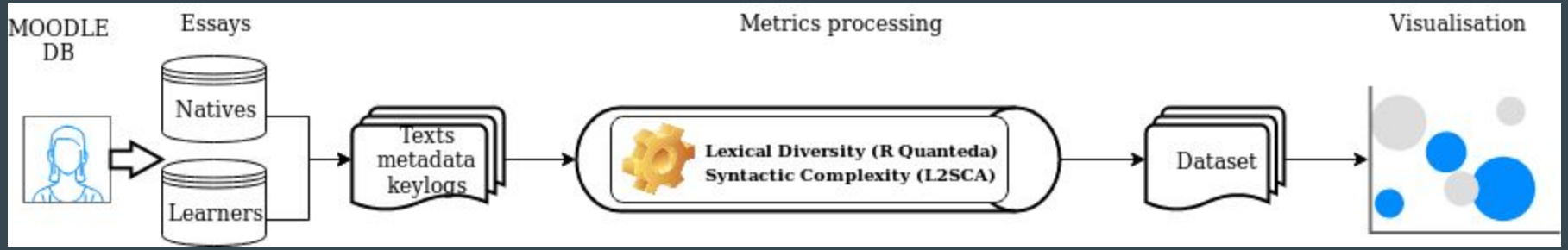
Intègre

- Stanford CoreNLP (Manning 2014)
- L2SCA (Lu 2010)
- R Quanteda library: texstat (lexdiv & readability) (Benoit 2018)

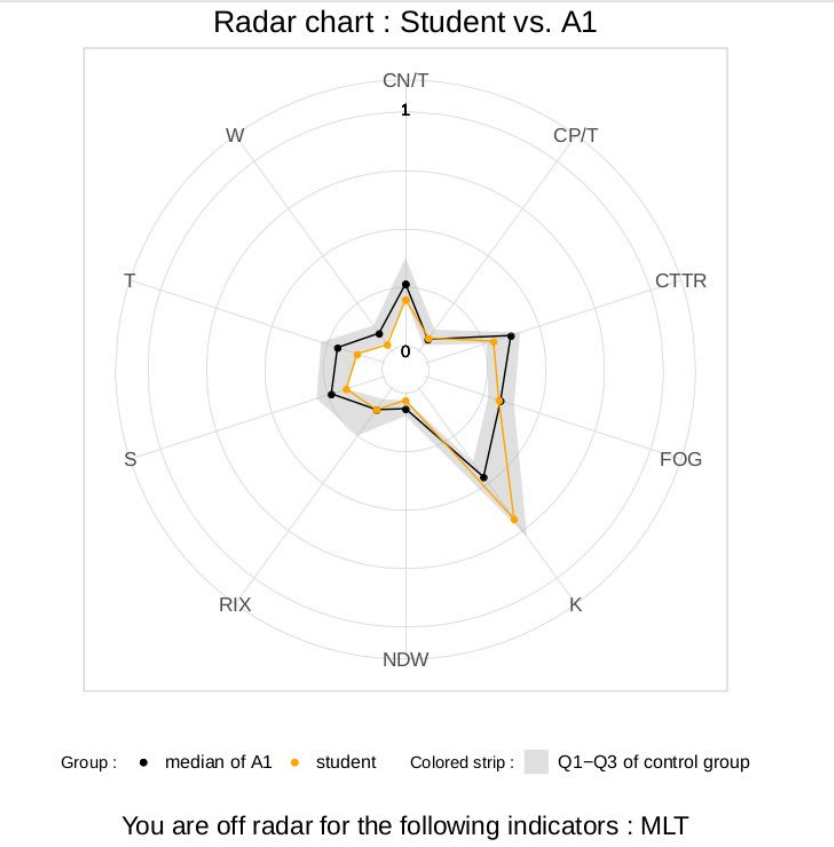
Métriques

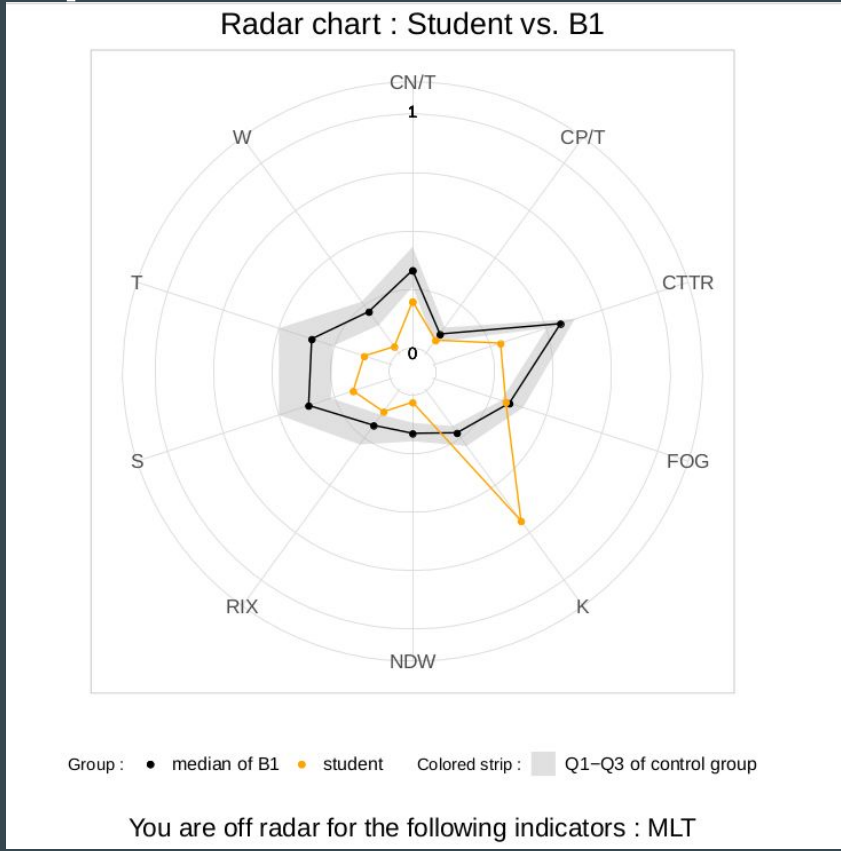
- Syntactiques (coordination, subordination)
- Diversité lexicale
- Lisibilité (niveau de difficulté d'un texte)

Traitement et jeux de données



Visualisations de profils linguistiques





Variable	Description	Category	Interpretation
CTTR	Carroll's Corrected Type Token ratio	Text.variation.words	The higher the richer the vocabulary. The index is correlaled with the length of your text. Longer texts have higher values.
W	Number of words	Text.size.words	The number of words in your text.
S	Number of sentences	Text.size.components	The number of sentences in your text. Even those including no grammatical construction.
T	Minimally terminable unit	Text.size.components	The number of sentences that include minimal grammar constructions such as subject + verb.
FOG	Gunning's Fog Index (Gunning 1952)	Sentence.size.word_syllables	How difficult or easy it is to read your text. The higher the more difficult a text is considered. It is linked to the number of words per sentence and the percentage of words with more than 3 syllables. For example, the New York Time articles are within the [11−12] interval.
RIX	Anderson's (1983) Readability Index	Sentence.size.word_syllables	How difficult or easy it is to read your text. The higher the more difficult a text. It is linked to the number of long words per sentence. RIX scores are normally between 1.5 (very easy) and 7.2 or above (very difficult).
NDW	Number of different words	Text.size.types	The number of different words in your text. For example "The experiment in the lab" includes 5 words but only 4 different words.
MLT	Mean Length of T−units	Sentence.size.words	How long your sentences are on average. Not too short not too long is good, e.g. 10 words on average.
CN/T	Average number of complex nominals in T−units	Sentence.component.component	The ratio of complex noun constructions relative to the number of sentences. This includes a noun used in the following cases: − compound words, e.g. "the science community" − adjective + nouns, e.g. "the scientific equipment" − genitive constructions, e.g. "the scientist's equipment" − participle + noun, e.g. "the finishing line" − subordinate clause, e.g. "This is the experiment that is conducted" or "The results suggest that the experiment is successful" − Infinitive/gerund as subject, e.g. "To experiment/experimenting this answers an important question"
CP/T	Average number of coordinate phrases per T−units	Sentence.component.component	The number of coordinate constructions in relation to the number of sentences, e.g. The clause "This experiment gives strong evidence and good results ..." includes "and" indicating coordination between "evidence" and "results".
K	Yule's K	Text.repetitions.types	How many times you repeat words. The index is linked to the number of times you use each word in your text. A small value indicates a rich vocabulary.

Perspectives

Continuer et élargir la collecte

Ajouter des métadonnées collectées ?

Distribution FAIR (Findable, Accessible, Interoperable, Reusable) des ressources:

1. Outil de collecte du corpus Moodle
2. Hébergement du corpus: Nakala Huma-num
3. Scripts de création jeux de données

Perspective

Déploiement

Merci

thomas.gaillat@univ-rennes2.fr

References

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. “Quanteda: An R Package for the Quantitative Analysis of Textual Data.” Journal of Open Source Software 3 (30): 774. https://doi.org/10.21105/joss.00774.

Conseil de l’Europe. 2018. Un Cadre Européen Commun de Référence Pour Les Langues: Apprendre, Enseigner, Évaluer Volume Complémentaire Avec de Nouveaux Descripteurs. Division des Politiques éducatives Service de l'Éducation. Strasbourg: Conseil de l’Europe. https://rm.coe.int/cecr-volume-complementaire-avec-de-nouveaux-descripteurs/16807875d5.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, eds. 2015. The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press.

Granger, Sylviane. 2008. “Learner Corpora in Foreign Language Education.” In Encyclopedia of Language and Education, edited by Nancy H. Hornberger, 1427–41. Springer US. https://doi.org/10.1007/978-0-387-30424-3_109.

Hawkins, John A., and Paula Buttery. 2010. “Criterial Features in Learner Corpora: Theory and Illustrations.” English Profile Journal 1 (01). https://doi.org/10.1017/S2041536210000103.

Leacock, Claudia. 2010. Automated Grammatical Error Detection for Language Learners. California: Morgan & Claypool Publishers.

Leńko-Szymańska, Agnieszka, and Alex Boulton. 2015. Multiple Affordances of Language Corpora for Data-Driven Learning. John Benjamins Publishing Company.

Lu, Xiaofei. 2010. “Automatic Analysis of Syntactic Complexity in Second Language Writing.” International Journal of Corpus Linguistics 15 (4): 474–496.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. “The Stanford CoreNLP Natural Language Processing Toolkit.” In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55–60. http://acl2014.org/acl2014/.

Vajjala, Sowmya. 2017. “Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features.” International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-017-0142-3.

Many thanks to:

Xiaofei Lu

Detmar Meurers

Schmid (Treetagger)

Syntactic complexity metrics

meanSentenceLength meanWordSyllables W S VP C T DC CT CP CN MLS MLT MLC C/S VP/T C/T DC/C DC/T T/S CT/T CP/T CP/C CN/T CN/C

Readability metrics

ARI ARI.simple Bormuth Bormuth.GP Coleman Coleman.C2 Coleman.Liau Coleman.Liau.grade Coleman.Liau.short Dale.Chall Dale.Chall.old Dale.Chall.PSK Danielson.Bryan Danielson.Bryan.2 Dickes.Steiwer DRP ELF Farr.Jenkins.Paterson Flesch Flesch.PSK Flesch.Kincaid FOG FOG.PSK FOG.NRI FORCAST FORCAST.RGL Fucks Linsear.Write LIW nWS nWS.2 nWS.3 nWS.4 RIX Scrabble SMOG SMOG.C SMOG.simple SMOG.de Spache Spache.old Strain Traenkle.Bailer Traenkle.Bailer.2 Wheeler.Smith

Lexical diversity metrics

TTR C.x R CTTR U S.x K D Vm Maas (a, log V0 log eV0)

https://www.rdocumentation.org/packages/quanteda/versions/0.9.7-17/topics/lexdiv

https://quanteda.io/reference/textstat_lexdiv.html

L2SCA component definitions

- **Sentence**: a group of words (including sentence fragments) punctuated with a

  sentence-final punctuation mark, such as a period, question mark, or exclamation mark.
- **Clause**: a structure with a subject and a finite verb, such as an independent,adjective, adverbial, or nominal clause (see, e.g., Hunt 1965; Polio 1997). Non-finite verb phrases are not counted as clauses.
- **Dependent clause:** a finite adjective, adverbial, or nominal clause (e.g.,Cooper 1976; Hunt 1965; Kameen 1979).
- **T-unit:** "a main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it" (Hunt 1970, p. 4).
- **Complex T-unit**: a T-unit with one or more dependent clauses (see, e.g., Casanave 1994).
- **Coordinate phrase:** a coordinate adjective, noun, or verb phrase.
- **Complex nominal:** a noun plus an adjective, possessive, prepositional phrase,adjective clause, participle, or appositive; a nominal clause; or a gerund or infinitive in subject position (see, e.g., Cooper 1976).
- **Verb phrase:** a finite or nonfinite verb phrase.

Metrics and scopes: a taxonomy for learner feedback

- We classify according to the types of variables used in each formula
- 3 exemples
  - ARI = 0.5ASL+4.71AWL−21.34
    - Word.size.characters (one of the variables focuses on word size in relation to characters)
    - Sentence.size.words (one of the variables focuses on sentence size in relation to words)
  - CN/C = Complex Nominals / Clauses
    - Sentence.component.component (one of the variables focuses on a specific sentence component in relation to another)
  - W = Total number of words in a text
    - Text.size.words (one of the variables focuses on text size in relation to words

Metrics and scopes: a taxonomy for learner feedback

Word.size.characters: ARI ARI.simple Bormuth Bormuth.GP Coleman.Liau Coleman.Liau.grade Coleman.Liau.short Dickes.Steiwer DRP, Fucks, nWS nWS.2 , Traenkle.Bailer Traenkle.Bailer.2 Wheeler.Smith
Word.size.syllables: Coleman Coleman.C2 meanWordSyllables, Farr.Jenkins.Paterson, Flesch Flesch.PSK Flesch.Kincaid FOG FOG.PSK FOG.NRI FORCAST FORCAST.RGL, Linsear.Write, LIW, nWS nWS.2 nWS.3 nWS.4 Wheeler.Smith

Sentence.size.words: (n words/nsent) MLS MLT MLC ARI family, Bormuth family, Dale.Chall family, Farr.Jenkins.Paterson Fucks WS.3 nWS.4 Flesch Flesch.PSK Flesch.Kincaid FOG FOG.PSK
Sentence.size.characters: Danielson.Bryan family, Dickes.Steiwer,
Sentence.size.syllables : DRP ELF Flesch Flesch.PSK Flesch.Kincaid FOG FOG.PSK FOG.NRI RIX SMOG SMOG.C SMOG.simple Strain
Sentence.components: Verb Phrase (VP) Clauses (C) T-Units (T) Dependent Clauses (DC) Coordinate Phrases (CP) Complex Nominals (CN)
Sentence.components.components: C/S (Sentences) VP/T C/T DC/C DC/T T/S CT CT/T CP/T CP/C CN/T CN/C Traenkle.Bailer family (prepositions & conjonctions)

Text.size.words: W
Text.size.sentences: S Coleman.Liau family (n sentences/n words) Linsear.Write
Text.variation.words: TTR C (Log TTR) R (root TTR) CTTR U S Maas lgV0 lgeV0, Dickes.Steiwer
Text.repetitions.types: Yule's K Simpson's D Herdan's Vm
Text.sophistication.wordsDaleChalList: Bormuth Bormuth.GP Bormuth, Dale.Chall family, DRP, Scrabble
Text.sophistication.wordsSpacheList: Spache Spache.old