FACULTY OF ARTS AND PHILOSOPHY

LT3 & SDL

Orphée De Clercq, Joni Kruijsbergen, Fauve De Backer & Goedele Vandommele

HOW TO RELIABLY ASSESS DUTCH WRITING SKILLS: MAN VS MACHINE

2024 = LAUNCH OF CENTRALIZED TESTING IN FLANDERS

STEUNPUNT CENTRALE TOETSEN IN ONDERWIJS

Ca. 70,000 pupils
MATHEMATICS & DUTCH
Multidisciplinary team

DUTCH SKILLS?

➡ Reading Comprehension
➡ Writing (+ grammar)
} **Evaluation** {
➡ Closed
➡ Open

Explore creation of **a Dutch AES system** able **to score** young learners' **writing skills**

AUTOMATED ESSAY SCORING (AES)

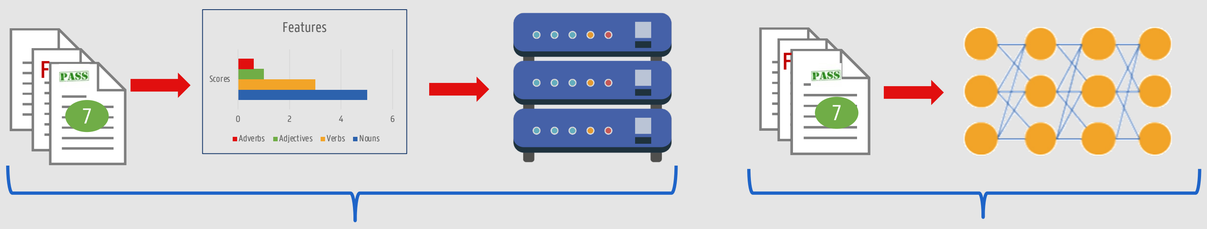
» Used for high-stakes tests → GRE & TOEFL (Richardson, 2021)

» High correlation with human raters (Allen et al., 2016)

» Mainly researched on English & essays (Strobl, 2019)

ETS e-rater

STATE OF THE ART = MACHINE LEARNING



FEATURE-BASED DEEP LEARNING

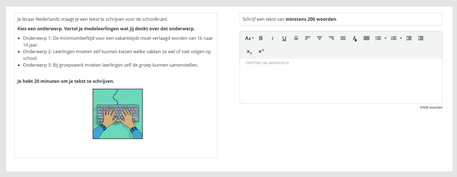
RELIABLE DATA = PREREQUISITE TO TRAIN AES

» Reliable ➔ **representative** dataset

» Flemish pupils 2nd year secondary education

» Stream A and stream B

» Different genres



Example prompt: INF

	Informative	Opinion	Avg. word lengh	Min words
STREAM A	3681	3677	843	14
STREAM B	1093	1065	622	10

1,502 texts selected

HUMAN

» Reliable ➔ **double-scoring**

Tekstkenmerken in de geselecteerde eindtermen
Eenvoudige, herkenbare tekststructuur
Redelijk herkenbare samenhang
Concrete inhoud
Gepast taalregister (afgestemd op doel en doelpubliek)
Tekstsoorten: informatief, persuasief, opiniërend, prescriptief, narratief
Bijkomend voor schriftelijke productie:
• Verzorgd taalgebruik met aandacht voor woordkeuze, spelling, zinsbouw, interpunctie, standaardnederlands, helderheid, adequaatheid, vormcorrectheid en vlotheid
• Verzorgde, gepaste lay-out

1. Analytical scoring

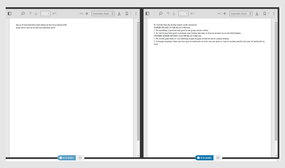
» **What?** Compare a text on certain aspects

» **How?** Designated scoring rubric +benchmarks

2. Comparative judgment

» **What?** Compare 2 texts (approx. 2 hours)

» **How?** Holistic scoring



» **Results** expected **Spring 2025**

MACHINE

» Experiments with historical data (STEP)

	Assessment	# texts
BL	instruction 1: E-mail (INF)	1254
BL	Instruction 2: Klaslokaal (OPI)	1270
A	instruction 1: Burgemeester (PERS)	684
A	instruction 2: Film A (INF)	682
B	instruction 1: Kippen (PRES)	564
B	instruction 2: Film B (INF)	547

» Scored with comparative judgment

FEATURE-BASED

» 388 features extracted with T-SCAN*

Lexical complexity Cohesion
Grammatical complexity Fluency

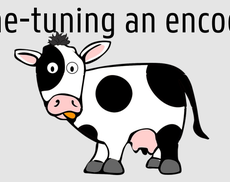
» Regression experiments

Experiment with all (388) features
Experiment with features after LASSO regression

scikit learn

DEEP LEARNING

» Fine-tuning an encoder-based LLM



» BERTje (de Vries et al.., 2019)

» Regression experiment

Hugging Face

RESULTS

10-fold cross validation | held-out evaluation split

Assessment		FEATURE-BASED				DEEP LEARNING		FEATURE-BASED				DEEP LEARNING	
		All features		LASSO		LLM		All features		LASSO		LLM	
		RMSE	QWK	RMSE	QWK	RMSE	QWK	RMSE	QWK	RMSE	QWK	RMSE	QWK
BL	instruction 1: E-mail (INF)	0.1208	0.7403	**0.1157**	0.7617	0.1177	**0.8003**	0.1085	0.7603	0.1129	0.7584	**0.1090**	**0.8064**
BL	Instruction 2: Klaslokaal (OPI)	0.1198	0.7856	0.1186	0.7923	**0.1091**	**0.8406**	0.1173	0.8012	0.1069	0.8154	**0.0962**	**0.8647**
A	instruction 1: Burgemeester (PERS)	0.1313	0.5533	**0.1243**	0.5851	0.1339	**0.6326**	0.1341	0.5463	0.1282	0.5884	**0.1174**	**0.6513**
A	instruction 2: Film A (INF)	0.1069	0.5872	**0.1021**	**0.6728**	0.1046	0.6628	0.1083	0.6276	**0.1033**	**0.7163**	0.1166	0.6414
B	instruction 1: Kippen (PRES)	0.1135	0.7369	**0.0993**	0.7890	0.1008	**0.8334**	0.1204	0.6598	0.1104	0.7525	**0.1090**	**0.7877**
B	instruction 2: Film B (INF)	0.1241	0.7220	**0.1135**	0.7830	0.1182	**0.7897**	0.1386	0.6954	**0.1319**	**0.7136**	0.1166	0.6414

FUTURE WORK

» How to operationalize formative aspects of writing in ML setting?

» More ML experiments on Steunpunt data + how to scale?

» Deep learning works well BUT focus should be on feedback and transparency (features, probing, …)

REFERENCES

- Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), Handbook of writing research (pp. 316–329). The Guilford Press.
- de Vries, W, van Cranenburgh, A, Bisazza, A., Caselli, T., van Noord, G. & Nissim, M. (2019), BERTje: A Dutch BERT Model. http://arxiv.org/abs/1912.09582..
- Richardson, M., & Clesham, R. (2021). Rise of the machines? The evolving role of AI technologies in high-stakes assessment. London Review of Education, 19(1). https://doi.org/10.14324/LRE.19.1.09
- Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. Computers & Education, 131, 33–48.

* https://github.com/UUDigitalHumanitieslab/tscan

Contact

Orphee.DeClercq@UGent.be
Joni.Kruijsbergen@UGent.be
Fauve.DeBacker@UGent.be
GoedeleVandommele@UGent.be

GHENT UNIVERSITY