GLAD: Groningen Lightweight Authorship Detection

E. van den Berg, M. Hürlimann, B. Weck, S. Šuster, M. Nissim
e.m.van.den.berg.4@student.rug.nl

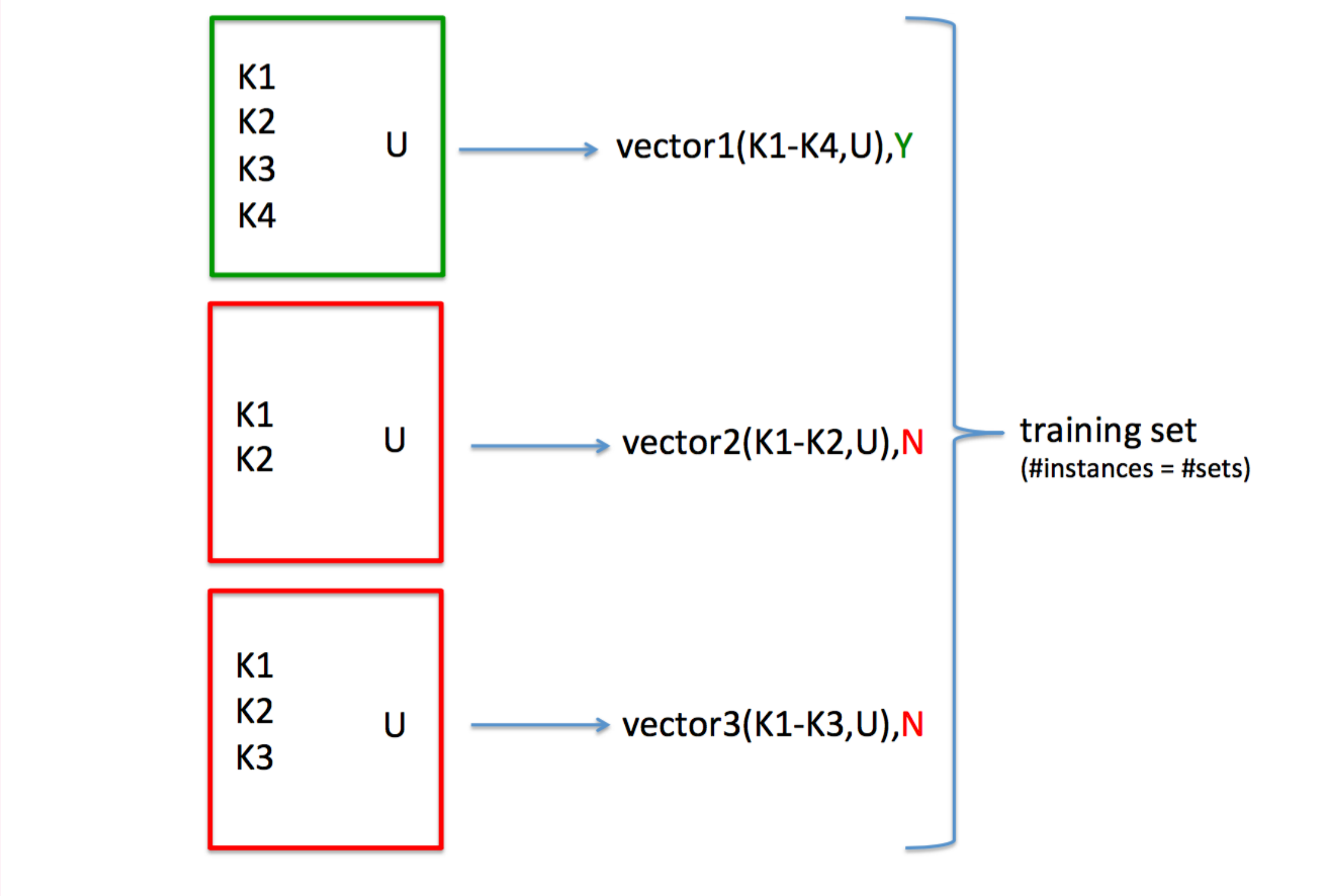
university of groningen

Authorship Verification...

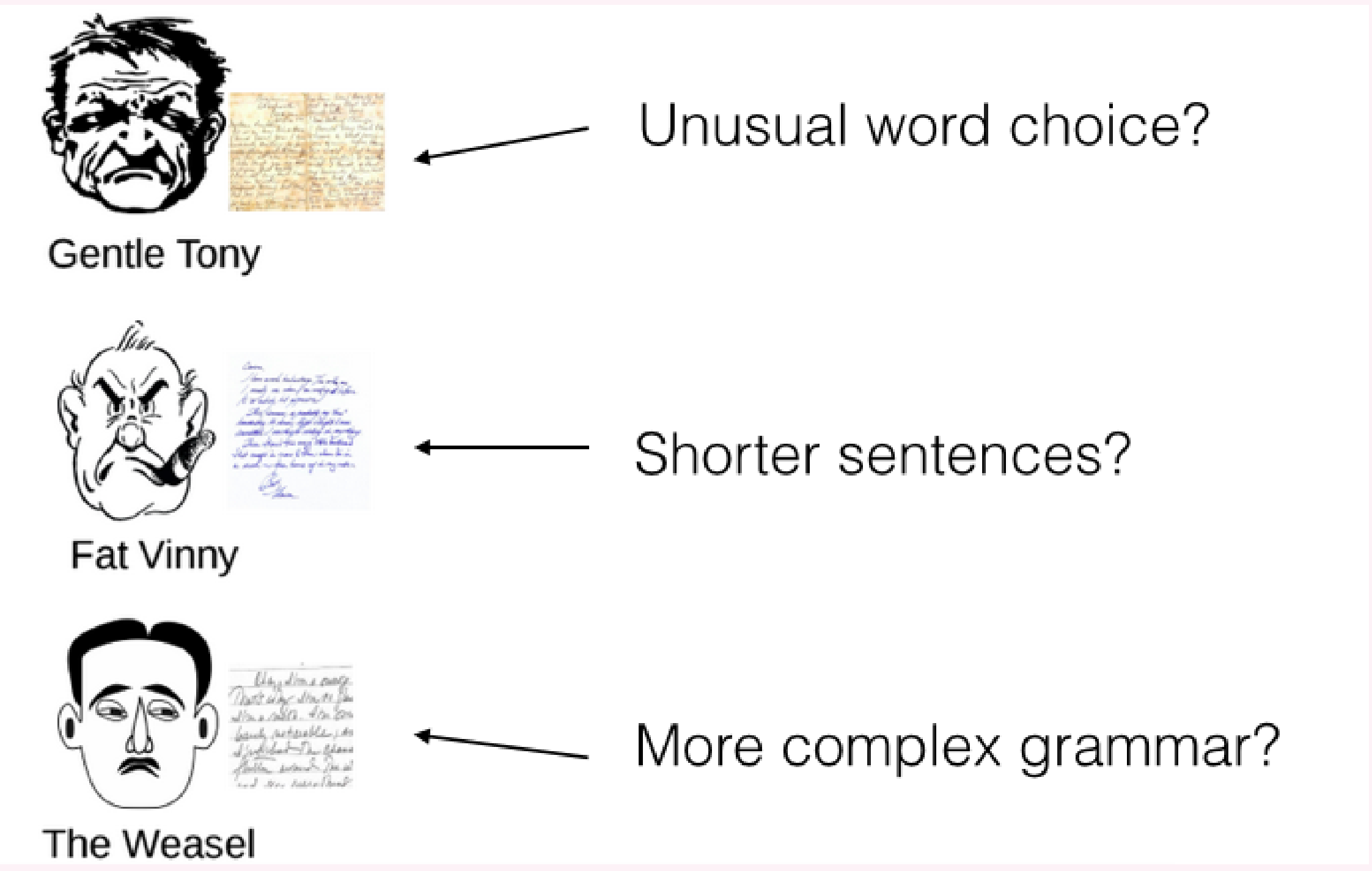
Is this an instance of the class SAME-AUTHOR or DIFFERENT-AUTHOR?

KNOWN AUTHOR	UNKNOWN AUTHOR
Jim, you won't go fightin' in the sun, with the birds all callin'? I'm not lookin' for it. Daisy, I love you. And I love you, Jim. I don't want nothin' more than you in all the world.	*You love me, don't you, David? You do, don't you, David? Tell me!* *I'm your husband, Annie, and you're my wife. Could there be aught but love between us after all these years?*

... as classification



What characterizes authors?

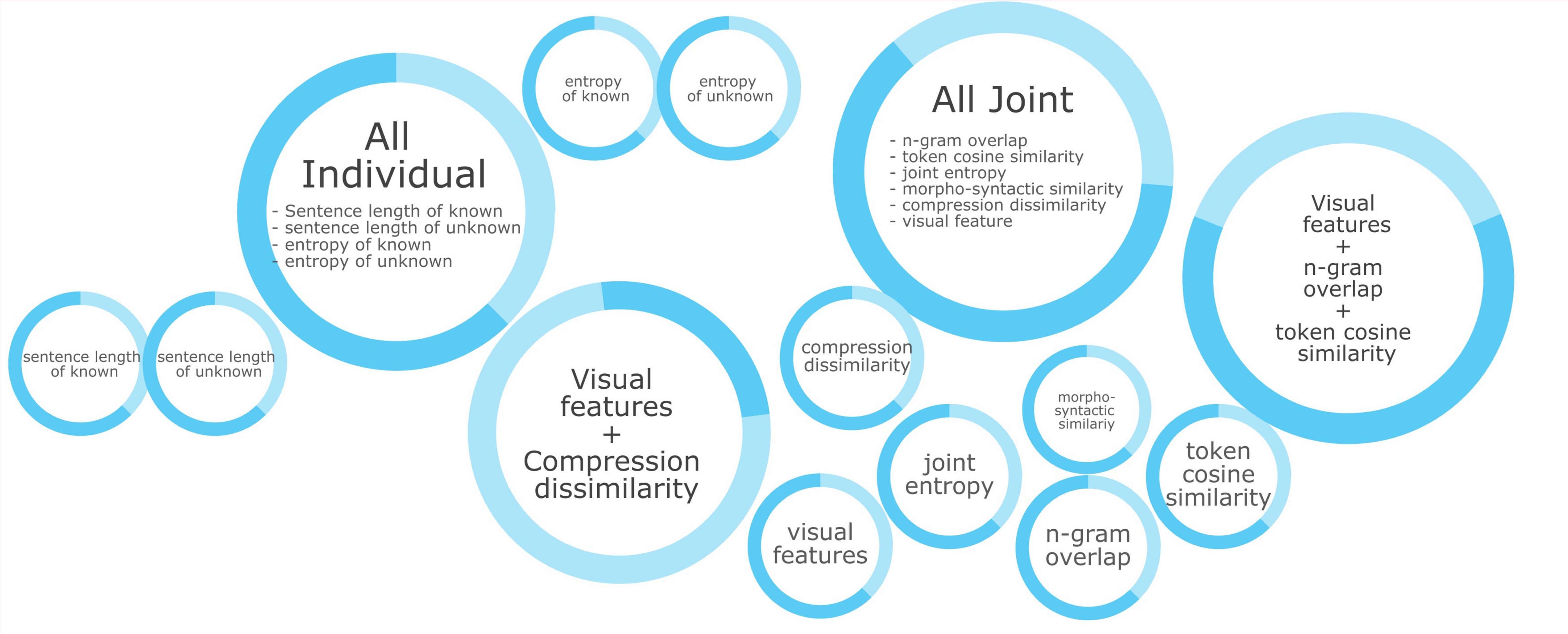


Unusual word choice?

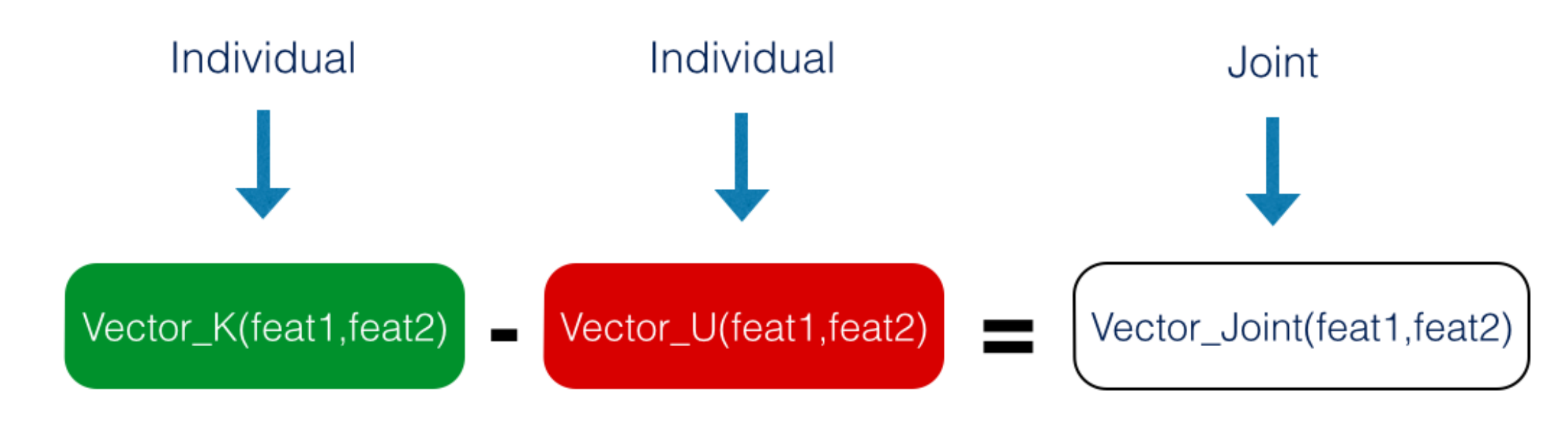
Shorter sentences?

More complex grammar?

All features and combinations of features



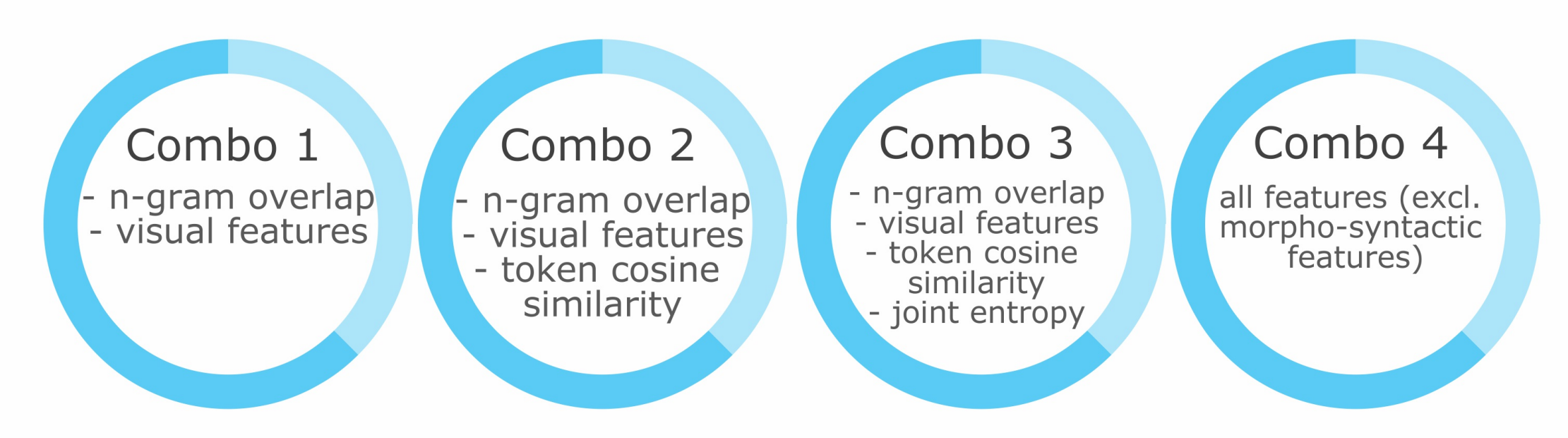
Sidenote: Individual vs Joint



Features' performance

Language	Helpful	Harmful
Dutch	All Individual	All Joint
Greek	All Joint	Just Entropy
Spanish	All Joint	Just Ngram sim
English	Vis+ngram+tok	Just POS sim

Best feature combinations



Sidenote: visual features

... incl. punctuation, letter case, line length
... to detect lay-out differences in plays / poetry

Pa-pa, pa-pa, pa-pa!

Here, stop her. She'll fall down. Here, turn around. Walk this way.

Ma-ma, ma-ma, ma-ma;

Oh, I think you are a darling.

Mer-ry Christ-mas! Mer-ry Christmas.

Results: combined scores on PAN 2015 data

Language	Training	Test	Ranking
Dutch (full set)	.55	.62	**3**/17
English (full set)	.56	.41	**8**/17
Greek (combo2)	.54	.60	**4**/15
Spanish (full set)	.90	.54	**5**/17

Average runtime: 1 minute

Discussion

- Why do complex features not outperform surface-level features?
- Are **individual** vs **joint** features valid ways of classifying texts written by the same of different authors?

Conclusion: Pros of the system

Robust

... works across genres
... works across topics
... works across languages

Simple

... easy to run
... uses simple features
... very fast runtime