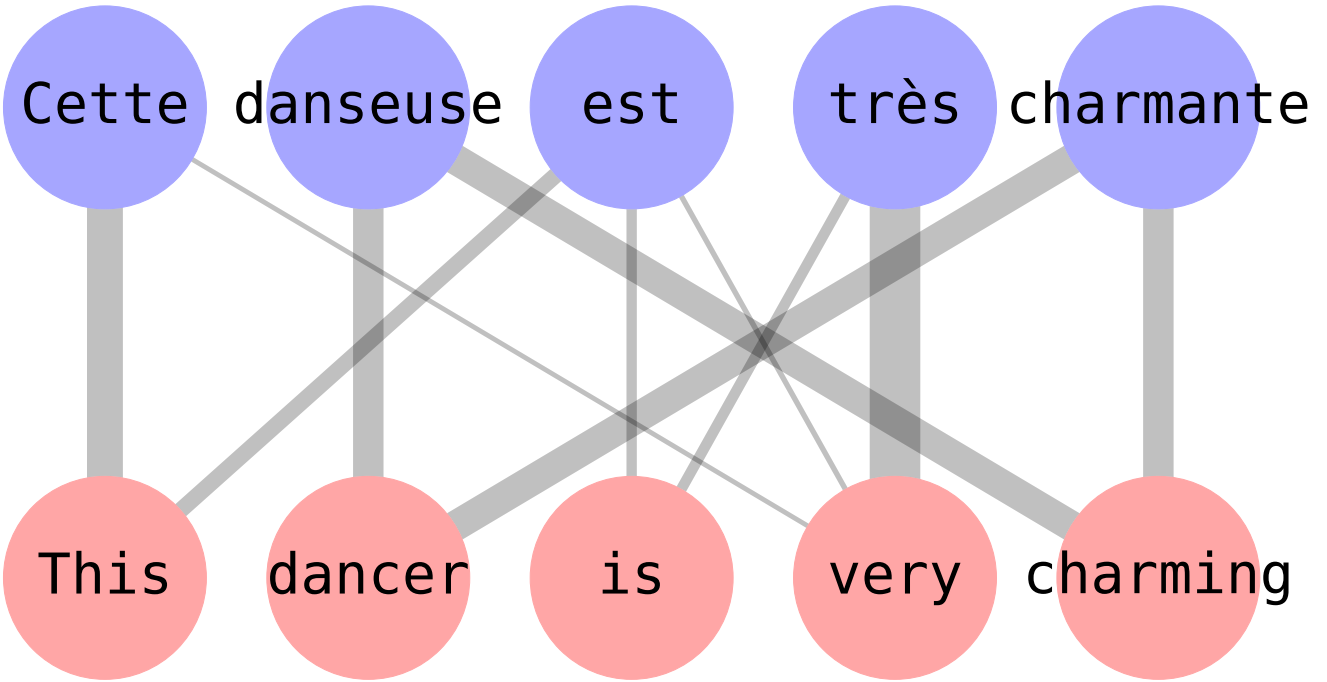
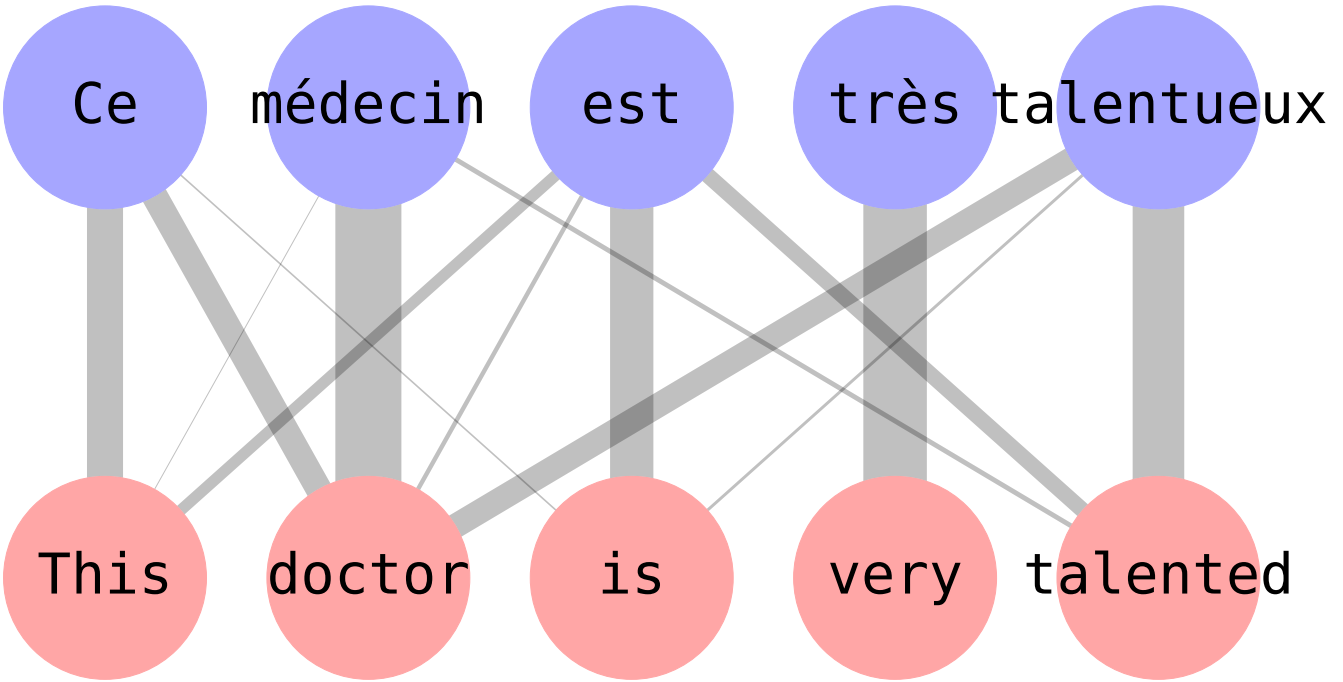
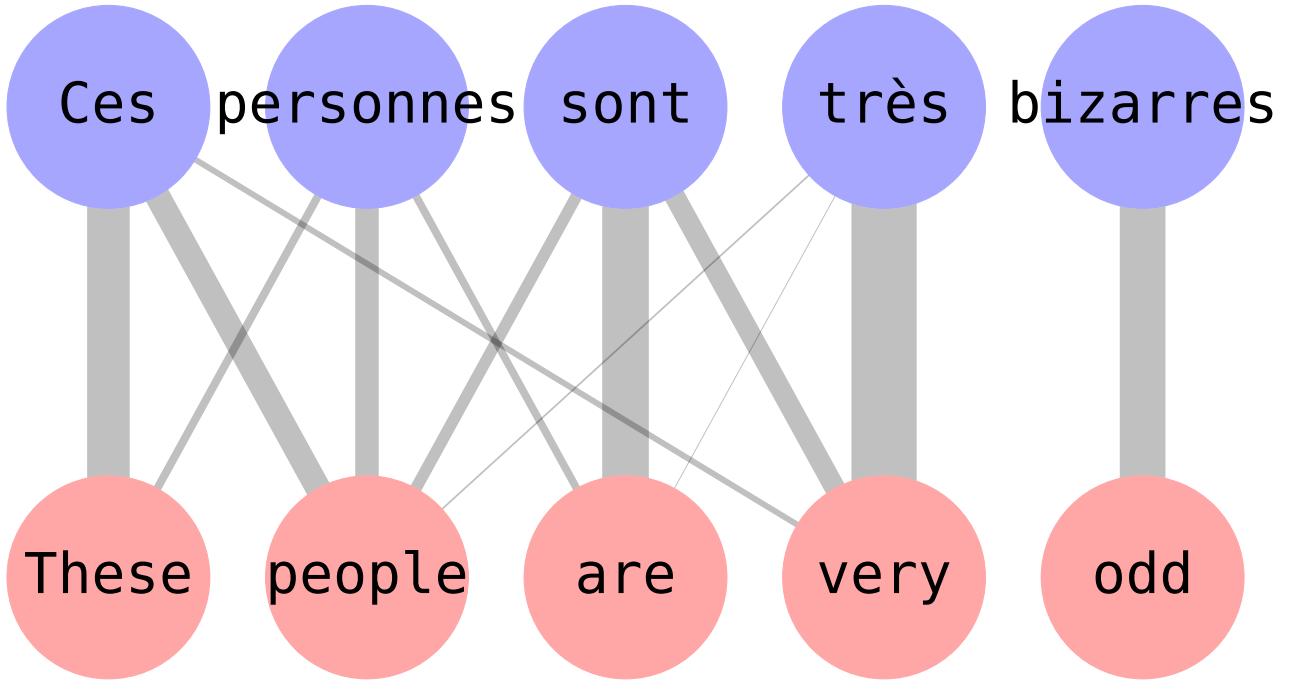
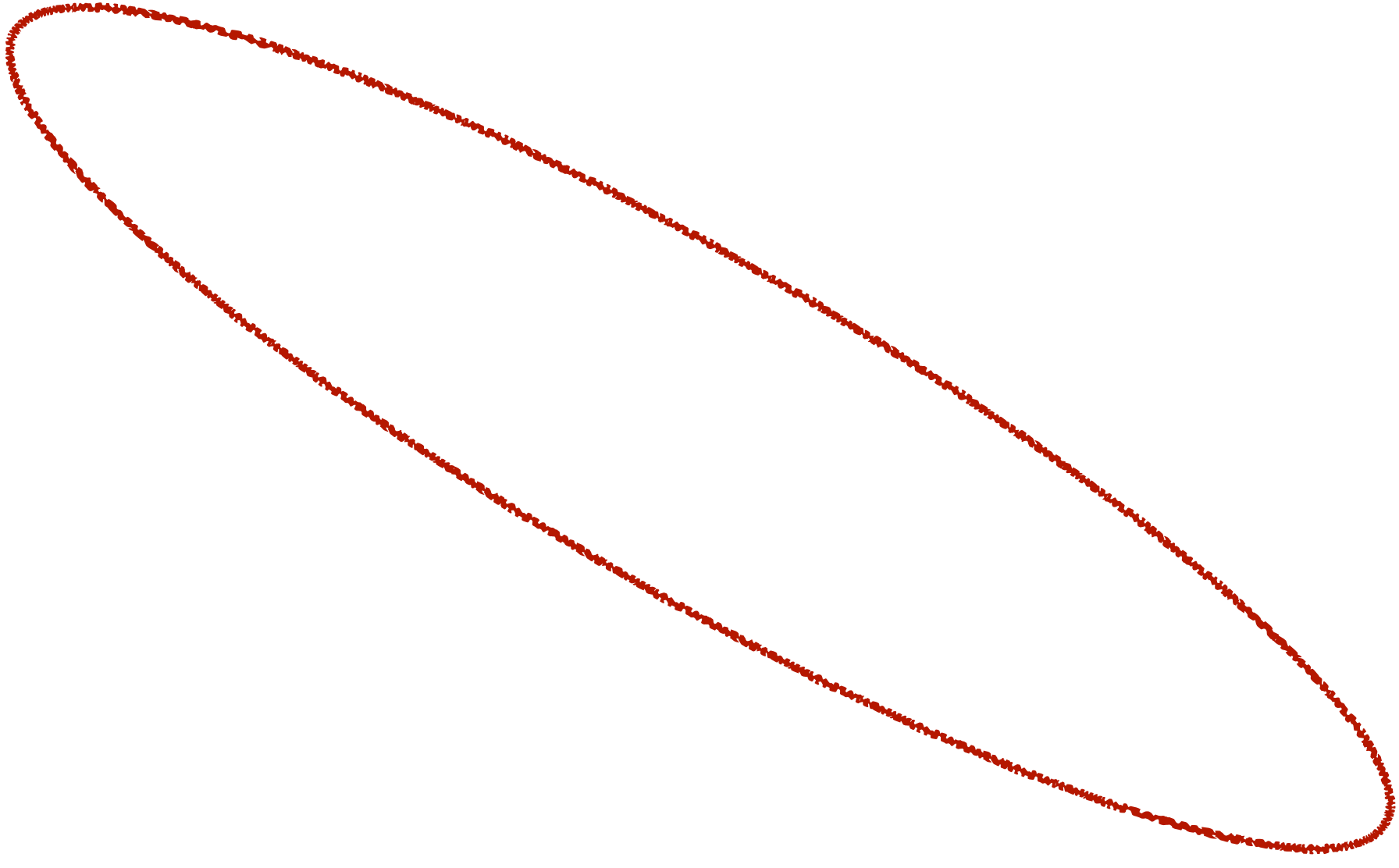
Interpretability for Bias Detection in Machine Translation

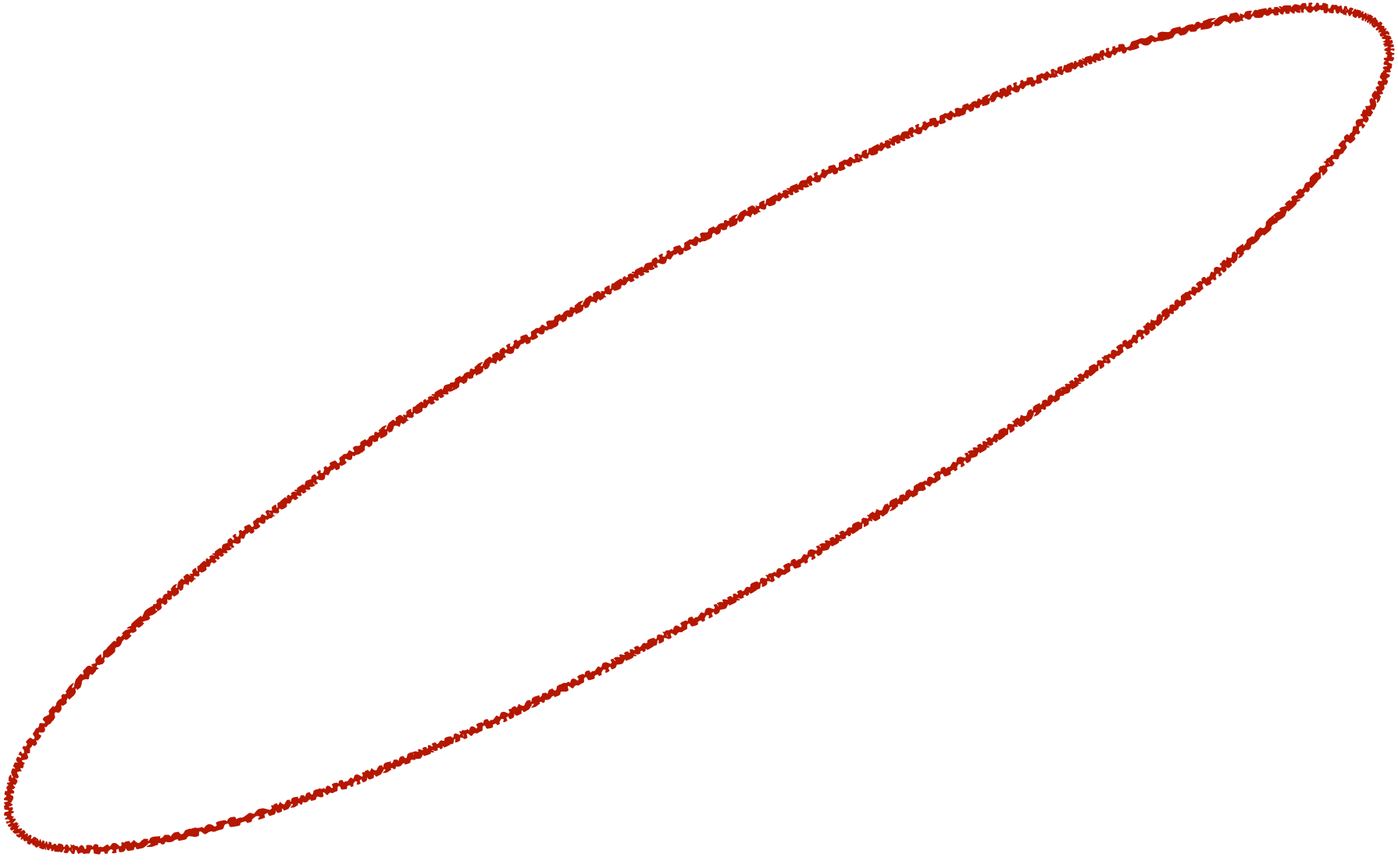
[redacted]

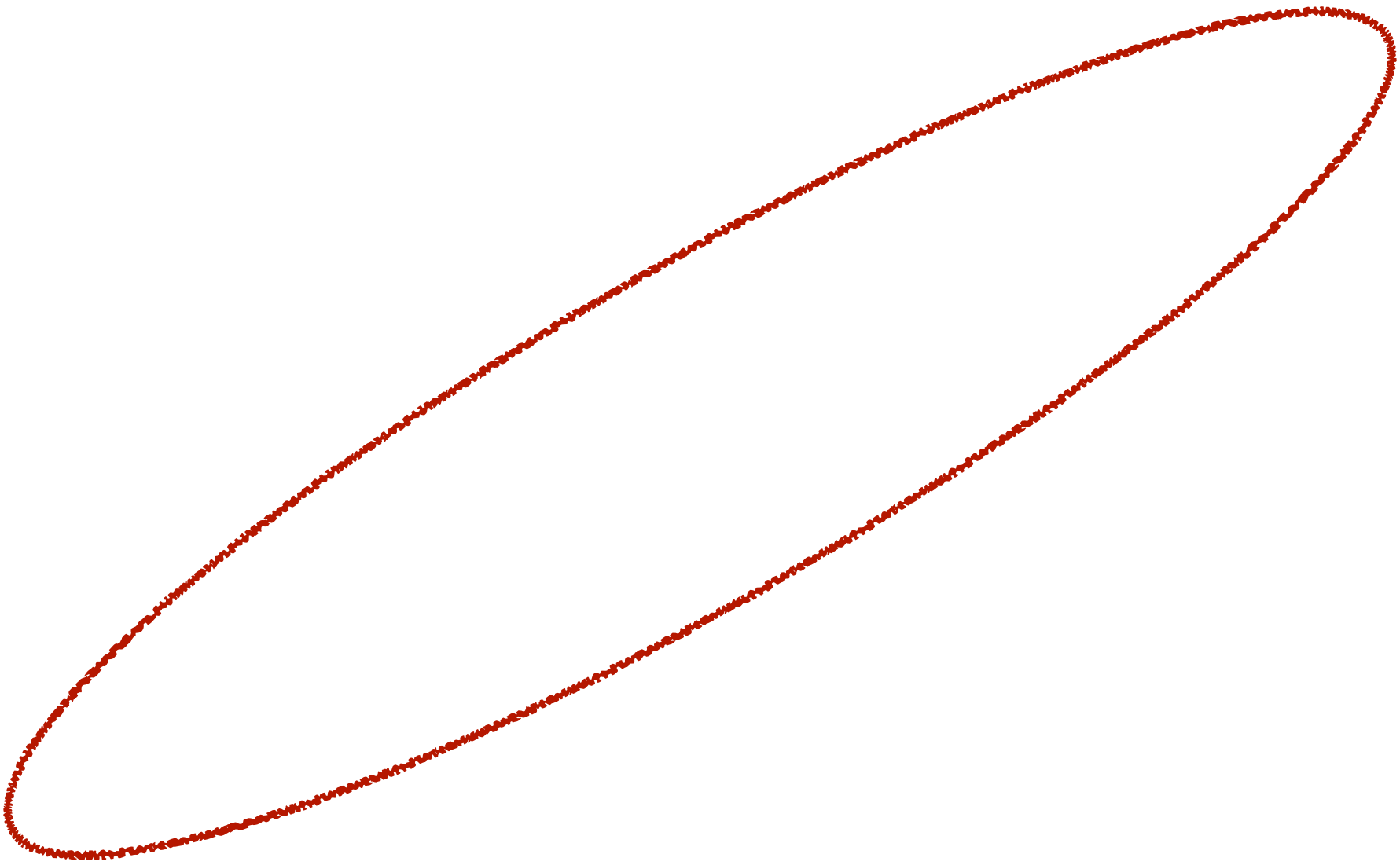

Cette
danseuse
est
très
charmante
This
dancer
is
very
charming


Ce
médecin
est
très
talentueux
This
doctor
is
very
talented


Ces
personnes
sont
très
bizarres
These
people
are
very
odd





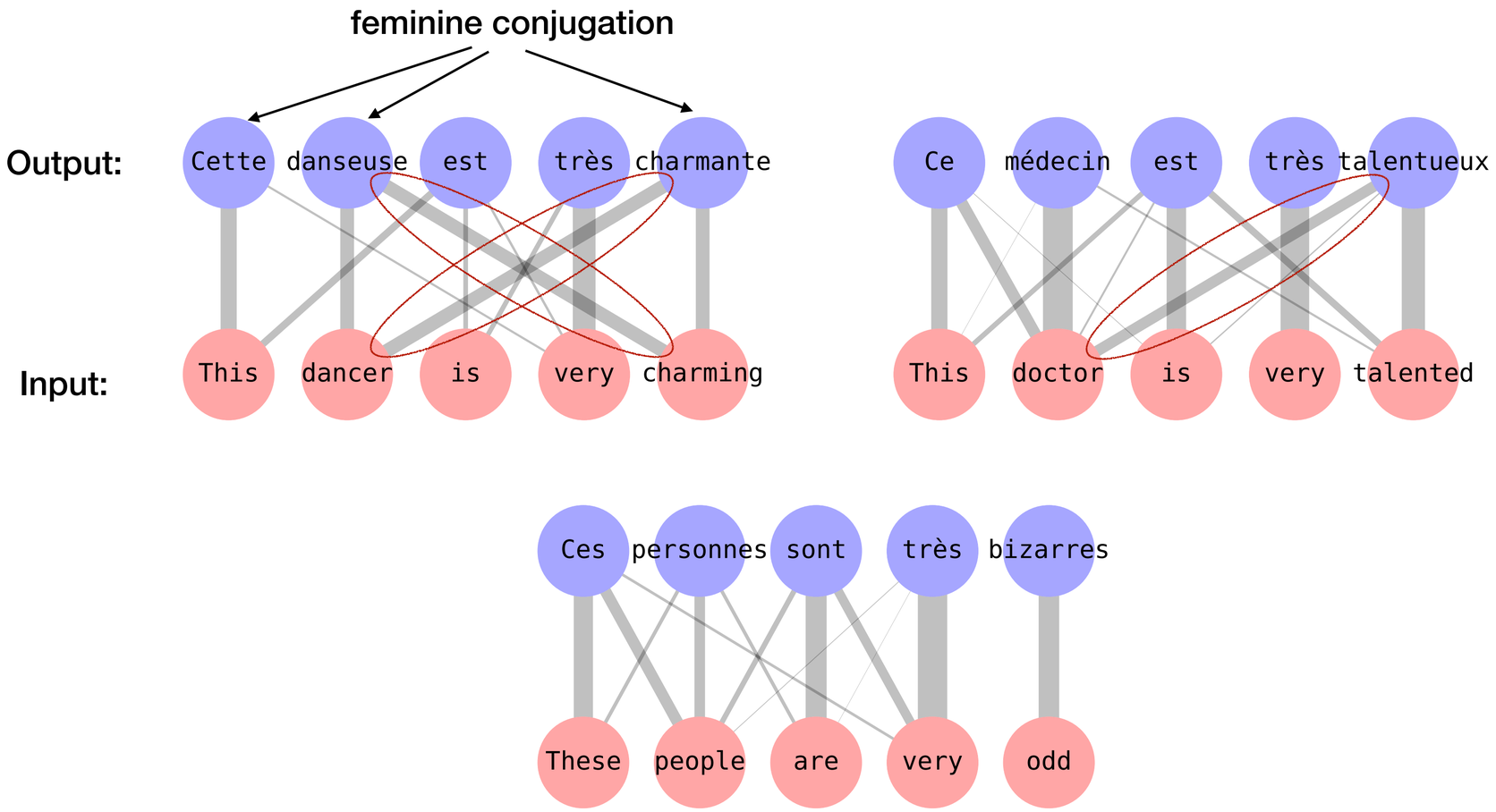


feminine conjugation

Input:

Output:

Interpretability for Bias Detection in Machine Translation



Interpretability for Flaw Detection in Dialogue Systems