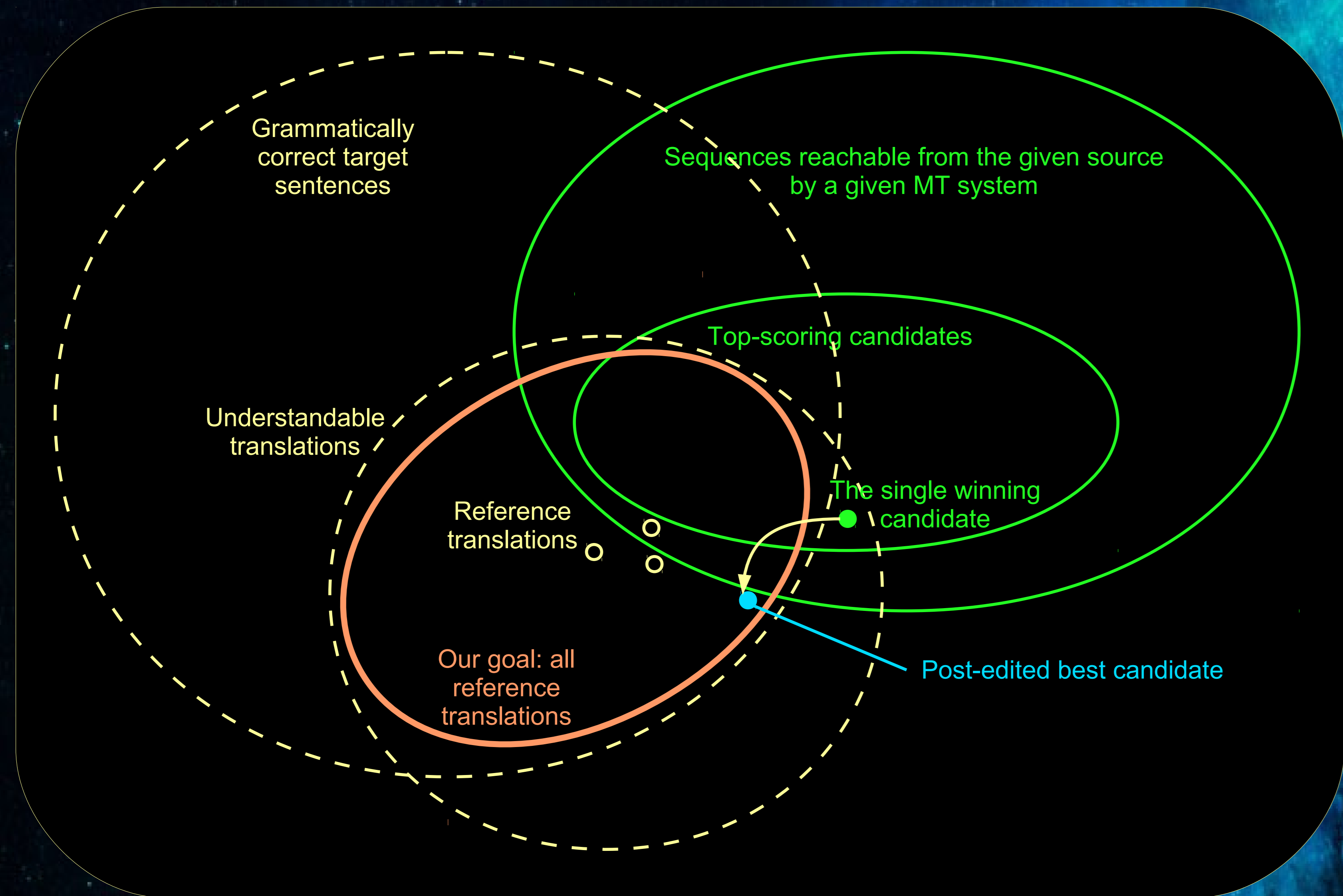
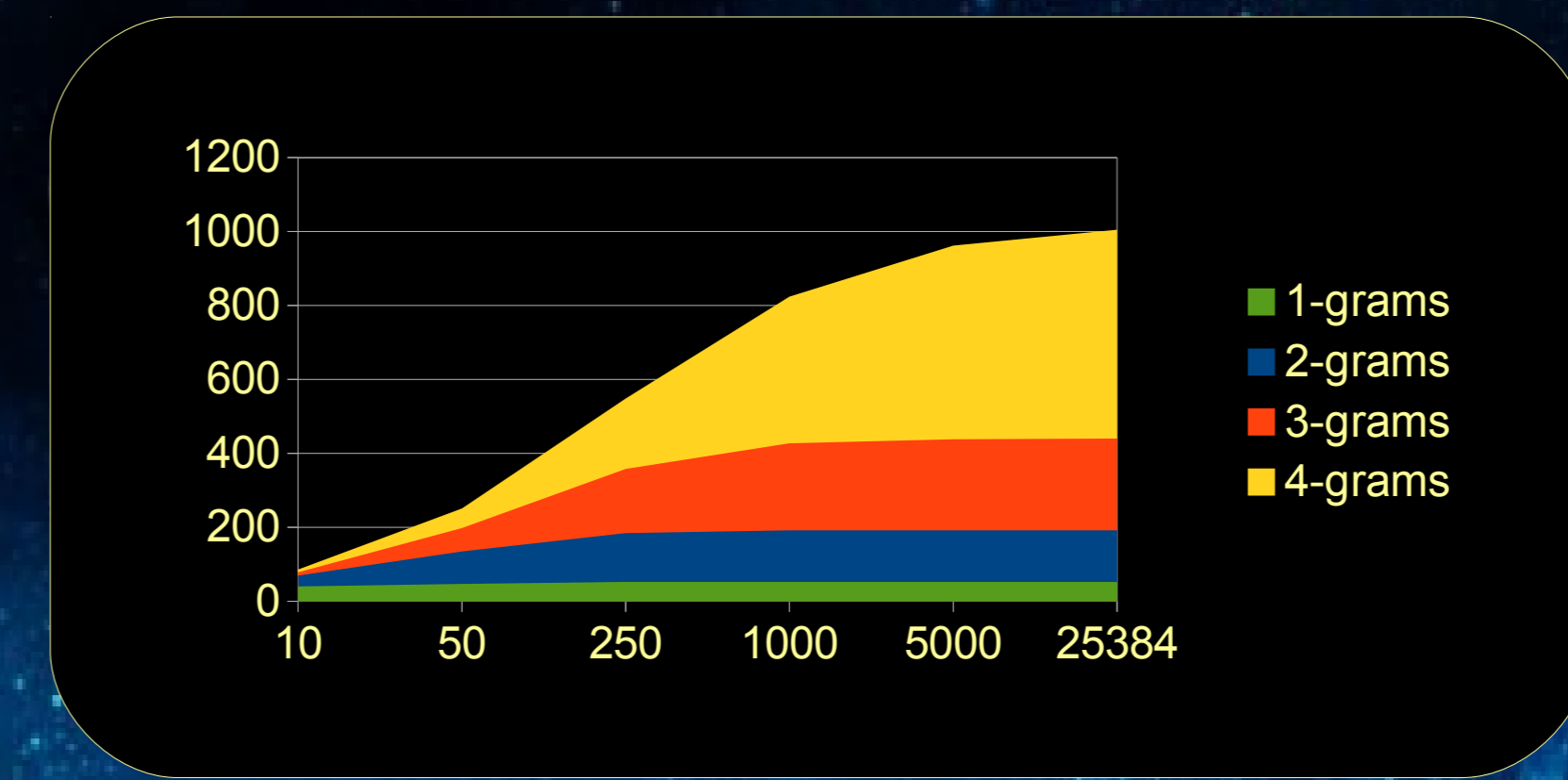




Daniel Zeman and Ondřej Bojar
 Charles University in Prague, Institute of Formal and Applied Linguistics
 Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky
 Malostranské náměstí 25, CZ-11800 Praha
 {zeman|bojar}@ufal.mff.cuni.cz



MORE REFERENCES NEEDED!



TOTAL

52	1-gram types
192	2-gram types
440	3-gram types
1005	4-gram types

Reference length: min 6, max 19 tokens.

The remaining three are single-L-159.
 Die übrigen drei Flugzeuge sind einsitzig
 Les trois appareils restants sont des vers
 Los tres aviones restantes son los mono

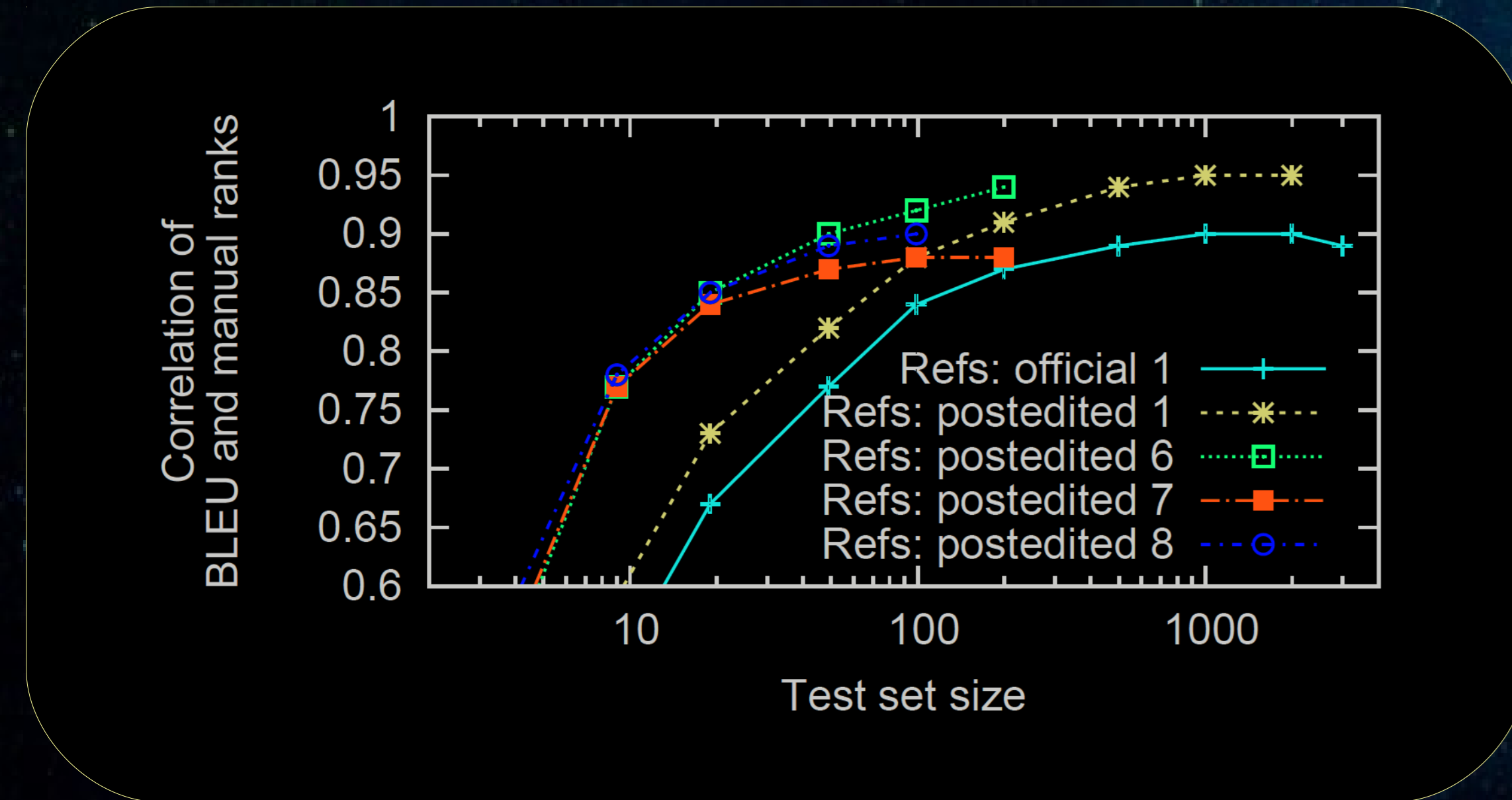
Zbývající tři stroje jsou jednomístné L-15
 Zbylé tři letouny jsou jednomístné L-159.
 Zbývající tři letouny jsou jednomístné L-1
 Ostatní tři letouny jsou jednomístné L-15

Google en-cs / de-cs:
 Zbývající tři jsou jedno-L-159.
 Zbývající tři letadla jsou jednomístný L-1

The remaining three planes...
 The remaining three airplanes...
 The remaining three aircrafts...
 The other three planes...

25,384 translations

Tři zbylé jsou L-159, do kterých se vejde
 Trojice zbyvajících jsou L-159, kam se v
 Trojice ostatních letounů jsou verze L15
 U trojice zbyvajících jde pak o L-159, ka
 U zbylých tří pak jde o verze L-159, kam
 U ostatních tří letadel pak jde o L-159, d
 U trojice ostatních se pak jedná o jedno
 Trojice letounů, která zbyvá, jsou L159,
 U zbyvajících tří letadel se pak jedná o L
 Trojice zbyvajících letounů jsou pak verze
 U zbylých tří letounů pak jde o L159, ka
 Trojice zbyvajících strojů jsou L159, kam
 Trojice letadel, která zbyvá, jsou pak ver
 U zbylých tří letounů jde o L159, kam se
 Trojice zbyvajících letounů jsou pak L15
 Trojice zbylých letounů jsou jednomístné
 Trojice zbylých jsou verze L-159, kam se
 U trojice zbylých letadel se jedná o L159
 U trojice zbylých se pak jedná o verze L1
 Zbylá trojice letounů jsou pak L159, do k
 U zbyvajících trojice strojů jde o L159, kam
 U trojice zbyvajících pak jde o verze L-1
 Trojice zbyvajících strojů jsou verze L-1
 Zbývající tři letouny jsou verze L-159 pro
 Tři zbyvajících jsou L159, do kterých se ve
 U zbyvajících trojice jde o verze L-159, d



Pilot study: we have 50 sentences from WMT 2011.
 6 translators, 77 sets of translations together.
 24 sentences translated by one translator.
 25 sentences translated by two translators.
 1 sentence translated by three translators.
 At most 2 hours per translator per sentence.
 Average: tens of thousands of references per sentence.
 Record: 19 billion references for one long sentence.

2 hours per sentence is not feasible for large-scale translations. Can we generate paraphrases automatically? (WordNet, permutations...)

Then clean it semi-manually: identify **negative n-grams**, discard references that contain them.

The more reference translations per sentence we have, the smaller test data is needed for the same level of correlation with human judgements.

