YOUR NAME: REGISTRATION #

C. Interstellar First Contact (1/2)

C-1 The questions in this assignment are based on examples in Knight (1997). In fact, both Centauri and Arcturan have underlying real world languages, as it turns out Centauri is English and Arcturan is Spanish. The languages are obfuscated to Centauri and Arcturan in order to illustrate how a Statistical Machine Translation (SMT) system must start from scratch, since it has no prior knowledge of how the languages work.

CENTAURI	**ARCTURAN**
Ok-voon ororok sprok. Garcia and associates.	At-voon bichat dat. Garcia y asociados.
Ok-drubel ok-voon anok plok sprok. Carlos Garcia has three associates.	At-drubel at-voon pippat rrat dat. Carlos Garcia tiene tres asociados.
Erok sprok izok hihok ghirok. His associates are not strong.	Totat dat arrat vat hilat. Sus asociados no son fuertes.
Ok-voon anok drok brok jok. Garcia has a company also.	At-voon krat pippat sat lat. Garcia tambien tiene una empresa.
Wiwok farok izok stok. Its clients are angry.	Totat jjat quat cat. Sus clientes están enfadodos.
Lalok sprok izok jok stok. The associates are also angry.	Wat dat krat quat cat. Los asociados tambien están enfadados.
Lalok farok ororok lalok sprok izok enemok. The clients and the associates are enemies.	Wat jjat bichat wat dat vat eneat. Los clientes y los asociados son enemigos.
Lalok brok anok plok nok. The company has three groups.	Iat lat pippat rrat nnat. La empresa tiene tres grupos.
Wiwok nok izok kantok ok-yurp. Its groups are in Europe.	Totat nnat quat oloat at-yurp. Sus grupos están en Europa.
Lalok mok nok yorok ghirok clok. The modern groups sell strong pharmaceuticals.	Wat nnat gat mat bat hilat. Los grupos modernos venden medicinas fuertes.
Lalok nok crrrok hihok yorok zanzanok. The groups do not sell zanzanine.	Wat nnat arrat mat zanzanat. Los grupos no venden zanzania.
Lalok rarok nok izok hihok mok. The small groups are not modern.	Wat nnat forat arrat vat gat. Los grupos pequeños no son modernos.

YOUR NAME: REGISTRATION #

C. Interstellar First Contact (2/2)

The novel sentence which was offered for translation in English is: "clients do not sell pharmaceuticals in Europe."

Answers
C-1 jjat

C-2 hihok = arrat, yorok = mat

C-3 We need to use the process of elimination, when mapping all the words between the two sentences two words are unaligned, we assume these are translations of each other. Thus, clok = bat.

C-4 Here are the new matches:

crrok	(empty)
kantok	oloat
ok-yurp	at-yurp

"crrok" does not seem to have a Arcturan equivalent, like in English the word "do" is not translated in "do not sell" which simply becomes "not sells" in Spanish. (Or to put it another way, the Centauri word *crrrok* **has** a translation, but it's the "empty" word.)

C-5 *jjat arrat mat bat oloat at-yurp*
Since you cannot deduce with certainty the exact order of the Arcturan sentence, various orders of these words will be accepted.

C-6 Immediately, you are faced with a dilemma: should you translate *totat* as *erok* or wiwok? Because *wiwok* occurs more frequently and because you've never seen *erok* followed by any of the other words you're considering, *wiwok* seems more likely. (However, admittedly, this is only a best guess, and *erok* will also be accepted.) Next, you consider various word orders. There appears to be no grammatical path through these words. Suddenly, you remember that curious Centauri word *crrrok*, which had no translation. *Crrok*, however, turns out to be a natural bridge between *nok* and *hihok*, giving you the translation:

wiwok rarok nok crrrok hihok yorok clok.