

Translation and Target-language Norms

DeepL Translator

DeepL Translator

Translator

Linguee

DeepL Pro

Blog

Info

Translate from **English** (detected)

Translate into **French**

Neural Machine Translation (NMT) has obtained state-of-the-art performance for several language pairs, while only using parallel data for training. Target-side monolingual data plays an important role in boosting fluency for phrase-based statistical machine translation, and we investigate the use of monolingual data for NMT. In contrast to previous work, which combines NMT models with separately trained language models, we note that encoder-decoder NMT architectures already have the capacity to learn the same information as a language model, and we explore strategies to train with monolingual data without changing the neural network architecture. By pairing monolingual training data with an automatic back-translation, we can treat it as additional parallel training

La traduction automatique neuronale (NMT) a obtenu des performances de pointe pour plusieurs paires de langues, tout en utilisant uniquement des données parallèles pour la formation. Les données monolingues côté cible jouent un rôle important dans l'amélioration de la fluidité de la traduction automatique statistique basée sur les phrases, et nous étudions l'utilisation de données monolingues pour les NMT. Contrairement aux travaux précédents, qui combinent des modèles NMT avec des modèles de langage formés séparément, nous constatons que les architectures NMT codeur-décodeur ont déjà la capacité d'apprendre les mêmes informations qu'un modèle de langage, et nous explorons des stratégies pour s'entraîner avec des données monolingues sans modifier

Daniel HENKEL

TransCrit EA1569

UNIVERSITÉ
PARIS 8 *Neural Machine Translation Workshop*
VINCENNES-SAINT-DENIS

Paris 8-Paris Diderot, 15/03/19

TransCrit
Transferts critiques anglophones

- **1. Introduction**

- Machine Translation & training corpora
- The problems with bilingual corpora
 - *Translationese (a.k.a. “third code”)*
 - *Interference (a.k.a. “shining through”)*

- **2. Methods**

- Corpus “Henkel 108” (4 × 27 authors/translators)
- Quantitative/statistical methods

- **3. Results & Analysis**

- Grammatical indicator: En. *Would, Could, Should* vs. Fr. Conditional (*il viendrait*)
- Lexical indicator: Verbs of perception (En. *Hear* vs. Fr. *Entendre*)

- **4. Conclusions**

- **5. Références**

1. Introduction

- **Machine translation (including NMT) relies on parallel corpora :**
 - From a probabilistic perspective, translation is equivalent to finding a target sentence y that maximizes the conditional probability of y given a source sentence x , i.e., $\arg \max_y p(y | x)$. In neural machine translation, we fit a parameterized model to maximize the conditional **probability of sentence pairs using a parallel training corpus**. (Bahdanau, Cho, & Bengio, 2014)
 - Neural machine translation has recently achieved impressive results (...), while **learning from raw, sentence-aligned parallel text** and using little in the way of external linguistic information (Sennrich & Haddow, 2016)
 - Recent years have shown a rapid shift from phrase-based (PBMT) to neural machine translation (NMT) (...) as the most common machine translation paradigm. With **large quantities of parallel data**, NMT outperforms PBMT for an increasing number of language pairs (Van der Wees, Bisazza & Monz, 2017)

1. Introduction

- The problem with parallel corpora:
 - Target-text is not *French* but *French-translated-from-English* (FtrE), i.e. *translationese*

Source	Target
--THOMAS PARKE D'INVILLIERS Chapter 1	
In my younger and more vulnerable years my father gave me some advice that I've been turning over in my mind ever since.	Quand j'étais plus jeune, ce qui veut dire plus vulnérable, mon père me donna un conseil que je ne cesse de retourner dans mon esprit :
"Whenever you feel like criticizing any one," he told me, "just remember that all the people in this world haven't had the advantages that you've had."	- Quand tu auras envie de critiquer quelqu'un, songe que tout le monde n'a pas joui des mêmes avantages que toi.
He didn't say any more but we've always been unusually communicative in a reserved way, and I understood that he meant a great deal more than that.	Il n'en dit pas davantage, mais comme lui et moi avons toujours été exceptionnellement communicatifs tout en y mettant beaucoup de réserve, je compris que la phrase impliquait beaucoup plus de choses qu'elle n'en exprimait.
In consequence I'm inclined to reserve all judgments, a habit that has opened up many curious natures to me and also made me the victim of not a few veteran bores.	En conséquence, je suis porté à réserver mes jugements, habitude qui m'a ouvert bien des natures curieuses, non sans me rendre victime de pas mal de raseurs invétérés.
The abnormal mind is quick to detect and attach itself to this quality when it appears in a normal person, and so it came about that in college I was unjustly accused of being a politician, because I was privy to the secret griefs of wild, unknown men.	Un esprit anormal est prompt à découvrir cette qualité et à s'y attacher, quand elle se montre chez quelqu'un de normal ; voilà pourquoi, à l'Université, on m'a injustement accusé de politicailler parce que j'étais le confident des chagrins secrets de garçons déréglés et inconnus.
Most of the confidences were unsought--frequently I have feigned sleep, preoccupation, or a hostile levity when I realized by some unmistakable sign that an intimate revelation was quivering on the horizon--for the intimate revelations of young men or at least the terms in which they express them are usually plagiaristic and marred by obvious suppressions.	La plupart de ces confidences, je ne les avais pas recherchées - j'ai souvent feint le sommeil, la préoccupation ou une hostile légèreté quand, à un de ces signes qui ne trompent jamais, je reconnaissais qu'une révélation d'ordre intime pointait à l'horizon ; car d'habitude les révélations intimes des jeunes hommes, ou tout au moins les termes dans lesquels ils les expriment, sont entachées de plagiat et gâtées par de manifestes suppressions.
Reserving judgments is a matter of infinite hope.	Réserver son jugement implique un espoir infini.
I am still a little afraid of missing something if I forget that, as my father snobbishly suggested, and I snobbishly repeat, a sense of the fundamental decencies is parcelled out unequally at birth.	J'aurais encore un peu peur de rater quelque chose si j'oubliais, comme le suggérait mon père avec snobisme et comme avec snobisme je le répète ici, que le sentiment des décences fondamentales nous est réparti en naissant d'une manière inégale.
And, after boasting this way of my tolerance, I come to the admission that it has a limit.	Or, ayant fait ainsi étalage de tolérance, j'en viens à l'aveu que la mienne a ses limites.
Conduct may be founded on the hard rock or the wet marshes but after a certain point I don't care what it's founded on.	Notre conduite peut avoir pour fondation un roc dur ou de fluides marécages, mais passé un certain point, peu me chaut sur quoi elle est fondée.
When I came back from the East last autumn I felt that I wanted the world to be in uniform and at a sort of moral attention forever; I wanted no more riotous excursions with privileged glimpses into the human heart.	Quand je rentrai de New-York, l'automne dernier, j'aurais voulu que le monde entier portât un uniforme et se tînt figé dans une sorte de garde à vous moral ; je ne souhaitais plus d'excursions tumultueuses avec coups d'œil privilégiés dans le cœur humain.
Only Gatsby, the man who gives his name to this book, was exempt from my reaction--Gatsby who represented everything for which I have an unaffected scorn.	De cette réaction, je n'excluais que Gatsby, l'homme qui donne son nom à ce livre. Gatsby représentait pourtant tout ce à quoi je porte un mépris dénué d'affectation.
If personality is an unbroken series of successful gestures, then there was something gorgeous about him, some heightened sensitivity to the promises of life, as if he were	S'il est vrai que la personnalité est une suite ininterrompue de gestes réussis, il y avait en cet homme quelque chose de magnifique, je ne sais quelle sensibilité exacerbée aux

1. Introduction

- The problem with parallel corpora:

- Parallel corpora ...

- *are not comparable samples of English and French*

- *consist of:*

- English + French-translated-from-English (Fr. Translationese)

- or

- French + English-translated-from-French (En. Translationese)

- As a result, MT systems learn to imitate and produce *Translationese*

- (cf. Loock, 2019)

1. Introduction

- The problem with parallel corpora:
 - Two ways in which target-texts are different from target-language norms
 - 1) Translationese, a.k.a. Third code
 - “**Inappropriate SL metaphors and syntax, unnatural word order** and a high concentration of **unnatural-sounding terminology** are the sort of features which are typical of translationese.” (Shuttleworth, 2014)
 - “It is common, when reading translations, to feel that they are written in their own **peculiar style**. Translation scholars even speak of the language of translation as **a separate ‘dialect’ within a language**, which they call **third code or translationese**” (Korzen & Gylling, 2017)

1. Introduction

- The problem with parallel corpora:
 - Two ways in which target-texts are different from target-language norms
 - 2) Interference / Shining-through
 - We identify “genuine” shining through of properties of the source language into translations as a general **tendency of translators to introduce feature patterns that are typical of the source language** into the target texts, quantified in terms of the relative frequencies of comparable lexico-grammatical features. (Evert & Neumann, 2017)
 - “On a more general level, Hansen-Schirra and Steiner (2012: 272) describe the relationship between different types of translation-related behavior towards source and target language norms (...) as a continuum ranging from **shining through, i.e. orientation towards source language norms**, to normalization, orientation towards target language norms. (Ibid.)

1. Introduction

- Terminology

- Different classes of English/French
 - *En0* = *English ex nihilo* (“original English”, no prior influence)
 - *Fr0* = *French ex nihilo* (“original French”, no prior influence)
 - ≠
 - *EtrF* = *English-translated-from-French*
 - *FtrE* = *French-translated-from-English*
- Two types of inadequacies
 - *Translationese*/'third code' = *indirect interference*
 - *Interference*/'shining-through' = *direct interference*
- Two types of bilingual corpora: Comparable vs. Parallel (cf. McEnery & Xiao, 2007)
 - *Comparable corpora*
 - *ex nihilo* texts sharing common characteristics (same period, same discourse type, same size etc.)
 - *Parallel corpora*
 - source-texts (original) + target-texts (translations)

1. Introduction

- **Objectives (3 questions):**
 - 1. Do target-texts (i.e. translations) have the same characteristics as the natural (*ex nihilo*) target-language?
 - *How can this be demonstrated quantitatively?*
 - 2. Can inter-linguistic influences/interferences be demonstrated between source- and target-texts?
 - *Can they be measured?*
 - 3. Do we evaluate translation “quality” in terms of:
 - *bare semantic equivalence?*
 - *similarity to human translation? (BLEU score)*
 - *conformity to the norms of the target-language?*

Methods

2. Methods

- **Corpus design**

- Public domain original works+translations available in electronic format (.txt, .epub, .html)
- Inclusion criteria:
 - *Publication after 1850*
 - *Narrative prose*
 - *1 work per author/translator*
 - When several works+translations available
 - Newer is better
 - Bigger is better
- In total: 4 × 2.8m words = 10+ million words
 - *27 authors in En0 (1868-1928, median 1901)*
 - *+27 translators into FtrE*
 - *27 authors in Fr0 (1869-1921, median 1901)*
 - *+27 translators into EtrF*

2. Methods

- Corpus design

- Public domain original works+translations available in electronic format:

- *Bellamy, E.* *Looking Back* (1888), 78414 words
- *Benson, R.* *Lord of the World* (1907), 101303 words
- *Buchan, J.* *Greenmantle* (1916), 100621 words
- *Burnett, F.H.* *The Secret Garden* (1911), 83065 words
- *Collins, W.* *The Moonstone* (1868), 199978 words
- *Conan Doyle, A.* *The Lost World* (1912), 77152 words
- *Cox, E.* *Out of the Silence* (1925), 114354 words
- *Eliot, G.* *Middlemarch* (1871), 164459 words
- *Fitzgerald, F.S.* *Great Gatsby* (1925), 50091 words
- *Hardy, T.* *Tess of d'Urbervilles* (1891), 153078 words
- *Hope, A.* *Rupert of Hentzau* (1898), 85029 words
- *James, H.* *Washington Square* (1881), 65220 words
- *Joyce, J.* *Dubliners* (1914), 69204 words
- *Kipling, R.* *Kim* (1901), 108400 words
- *Lewis, S.* *Free Air* (1919), 83803 words
- *London, J.* *Martin Eden* (1909), 142463 words
- *Mansfield, K.* *The Garden Party* (1922), 59773 words
- *Morrow, W.* *The Ape, the Idiot and Other People* (1897), 57062 words
- *Reid, M.* *The Finger of Fate* (1872), 94265 words
- *Stevenson, R.L.* *Kidnapped+David Balfour* (1889), 186177 words
- *Stoker, B.* *Dracula* (1897), 163104 words
- *Twain, M.* *The Prince and Pauper* (1881), 70647 words
- *Wallace, E.* *Green Archer* (1923), 122339 words
- *Wallace, L.* *Ben Hur* (1880), 201313 words
- *Wells, H.G.* *The War in the Air* (1908), 99499 words
- *Wilde, O.* *The Portrait of Dorian Gray* (1890), 80410 words
- *Woolf, V.* *Orlando* (1928), 80310 words

2. Methods

- Corpus design

- Public domain original works+translations available in electronic format:

- *Audoux, M.* *Marie Claire* (1910), 38233 words
- *Barbusse, H.* *Le Feu* (1916), 119442 words
- *Bazin, R.* *Les Oberlé* (1901), 78366 words
- *Benoît, P.* *L'Atlantide* (1919), 67638 words
- *Bourget, P.* *La Duchesse Bleue* (1898), 89231 words
- *Daudet, A.* *L'Immortel* (1888), 65954 words
- *Flaubert, G.* *L'Éducation sentimentale* (1869), 149768 words
- *France, A.* *Les dieux ont soif* (1912), 71967 words
- *Gaboriau, E.* *L'argent des autres* (1874), 173224 words
- *Gourmont, R.* *Sixtine* (1890), 72967 words
- *Gyp* *Bijou* (1896), 58636 words
- *Hémon, L.* *Maria Chapdelaine* (1913), 49352 words
- *Hugo, V.* *Quatrevingt-treize* (1874), 122395 words
- *Huysmans, J.K.* *La Cathédrale* (1898), 132784 words
- *Leblanc, M.* *Les Dents du Tigre* (1921), 130095 words
- *Leroux, G.* *Rouilletable chez le Tsar* (1913), 102905 words
- *Loti, P.* *Les Derniers Jours de Pékin* (1902), 72326 words
- *Louÿs, P.* *Aphrodite* (1896), 55481 words
- *Malot, H.* *Conscience* (1888), 105655 words
- *Massenet, J.* *Mes Souvenirs* (1912), 76634 words
- *Mirbeau, O.* *Le journal d'une femme de chambre* (1915), 116755 words
- *Proust, M.* *Du côté de chez Swann* (1913), 181464 words
- *Rolland, R.* *Jean-Christophe 1-4* (1905), 239132 words
- *Souvestre, P.* *Fantômas* (1911), 91924 words
- *Vandérem, F.* *Les deux rives de la Seine* (1897), 93832 words
- *Verne, J.* *L'île mystérieuse* (1874), 209872 words
- *Zola, E.* *Les trois villes Paris* (1898), 192462 words

2. Methods

- 3 comparisons possible:
 - between **En0** and **Fr0**, *to obtain benchmark data for each language*

English *ex nihilo* (En0)

- 27 works/27 authors
- 2 891 533 words
- dates 1868-1928, median 1901

French *ex nihilo* (Fr0)

- 27 works/27 authors
- 2 958 494 mots
- dates 1869-1921, médiane 1901

2. Methods

- 3 comparisons possible:

- between En0 and Fr0, *to obtain benchmark data for each language*
- between **EtrF** and **En0**, *to ascertain whether the distribution of the conditional perfect in EtrF is similar to En0, and likewise for FtrE/Fr0*

<p style="text-align: center;">English <i>ex nihilo</i> (En0)</p> <ul style="list-style-type: none">• 27 works/27 authors• 2 891 533 words• dates 1868-1928, median 1901	<p style="text-align: center;">French-translated-from-English (FtrE)</p> <ul style="list-style-type: none">• 27 works/27 translators• 2 908 865 mots
<p style="text-align: center;">English-translated-from-French (EtrF)</p> <ul style="list-style-type: none">• 27 works/27 translators• 2 859 167 words	<p style="text-align: center;">French <i>ex nihilo</i> (Fr0)</p> <ul style="list-style-type: none">• 27 works/27 authors• 2 958 494 mots• dates 1869-1921, médiane 1901

2. Methods

- **3 comparisons possible:**

- between En0 and Fr0, *to obtain benchmark data for each language*
- between EtrF and En0, *to ascertain whether the distribution of the conditional perfect in EtrF is similar to En0, and likewise for FtrE/Fr0*
- between **source-** and **target-texts**, *to detect inter-linguistic influences/interference*

<p>English <i>ex nihilo</i> (En0)</p> <ul style="list-style-type: none">• 27 works/27 authors• 2 891 533 words• dates 1868-1928, median 1901	<p>French-translated-from-English (FtrE)</p> <ul style="list-style-type: none">• 27 works/27 translators• 2 908 865 mots
<p>English-translated-from-French (EtrF)</p> <ul style="list-style-type: none">• 27 works/27 translators• 2 859 167 words	<p>French <i>ex nihilo</i> (Fr0)</p> <ul style="list-style-type: none">• 27 works/27 authors• 2 958 494 mots• dates 1869-1921, médiane 1901

2. Methods

- Corpus preparation & Data collection

- Texts cleaned, punctuation homogenized, encoded in UTF-8 etc.

- Tagged for POS and Lemma with *TreeTagger*:

- *The*_{/DT/the} *sun*_{/NN/sun} *was*_{/VBD/be} *not*_{/RB/not} *yet*_{/RB/yet} *risen*_{/VNN/rise} ,/, *and*_{/CC/and} *the*_{/DT/the} *whole*_{/NN/whole} *of*_{/IN/of} *Crescent*_{/NP/Crescent} *Bay*_{/NP/Bay} *was*_{/VBD/be} *hidden*_{/VNN/hide} *under*_{/IN/under} *a*_{/DT/a} *white*_{/JJ/white} *sea-mist*_{/NN/sea-mist} ./SENT/.

- Queries targeting

- *Syntactic indicators, e.g.:*

- Would, Could, Should vs. Conditional, etc.

- *Lexical indicators, e.g.:*

- Hear (En) / Entendre (Fr)

- All data converted to frequencies per 1000 words (f/1k)

- Statistical analyses performed in R

Results & Analysis

3. Results & Analysis

1st indicator: Would, Could vs. Fr. Conditional

3. Results & Analysis

- Frequency of En. Would, Could, etc. Fr. Conditional:

- En0

- *Total n=17754 (frequency = 8.32/1k words)*

- max. Eliot 12.24/1k
 - median Woolf 8.14/1k
 - min. London 5.67/1k

- EtrF

- *Total n=14946 (frequency = 6.32/1k words)*

- max. Tr.Proust 13.77/1k
 - median Tr.Bazin 5.71/1k
 - min. Tr.Loti 1.88/1k

- FtrE

- *Total n=10746 (frequency = 4.85/1k words)*

- max. Tr.Cox 7.17/1k
 - median Tr.Stoker 4.92/1k
 - min. Tr.Woolf 2.63/1k

- Fr0

- *Total n=7116 (frequency = 3.15/1k words)*

- max. Proust 5.6/1k
 - median Mirbeau 2.96/1k
 - min. Loti 1.82/1k

3. Results & Analysis

1st indicator: Would, Could vs. Fr. Conditional

Question 1:

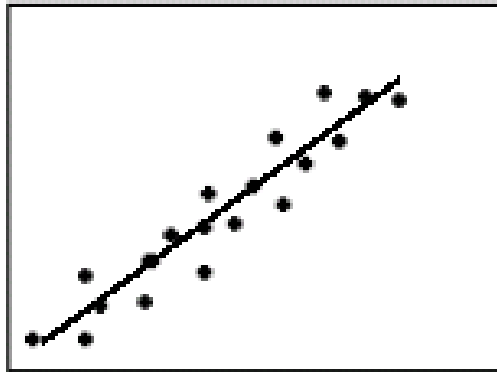
Does $En0$ exert an influence on $FtrE$?

Does $Fr0$ exert an influence on $EtrF$?

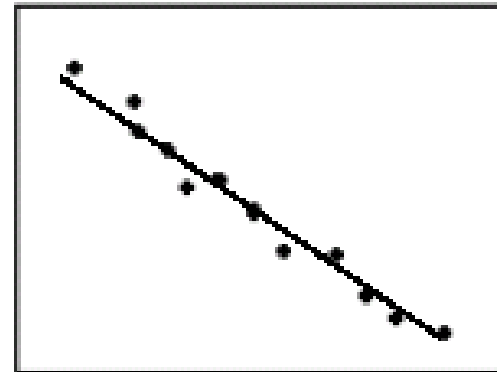
3. Results & Analysis

- Interpreting scatterplots:

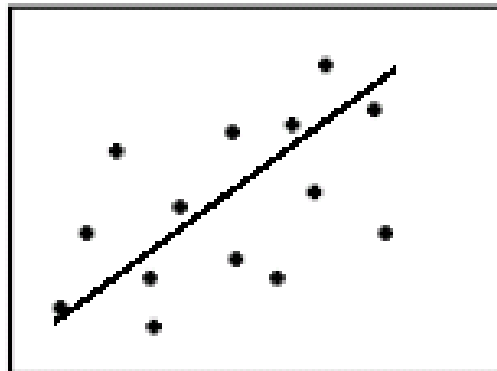
Degree of Correlation



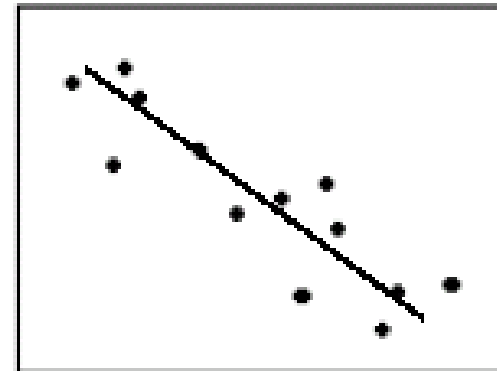
Strong Positive



Strong Negative



Weak Positive



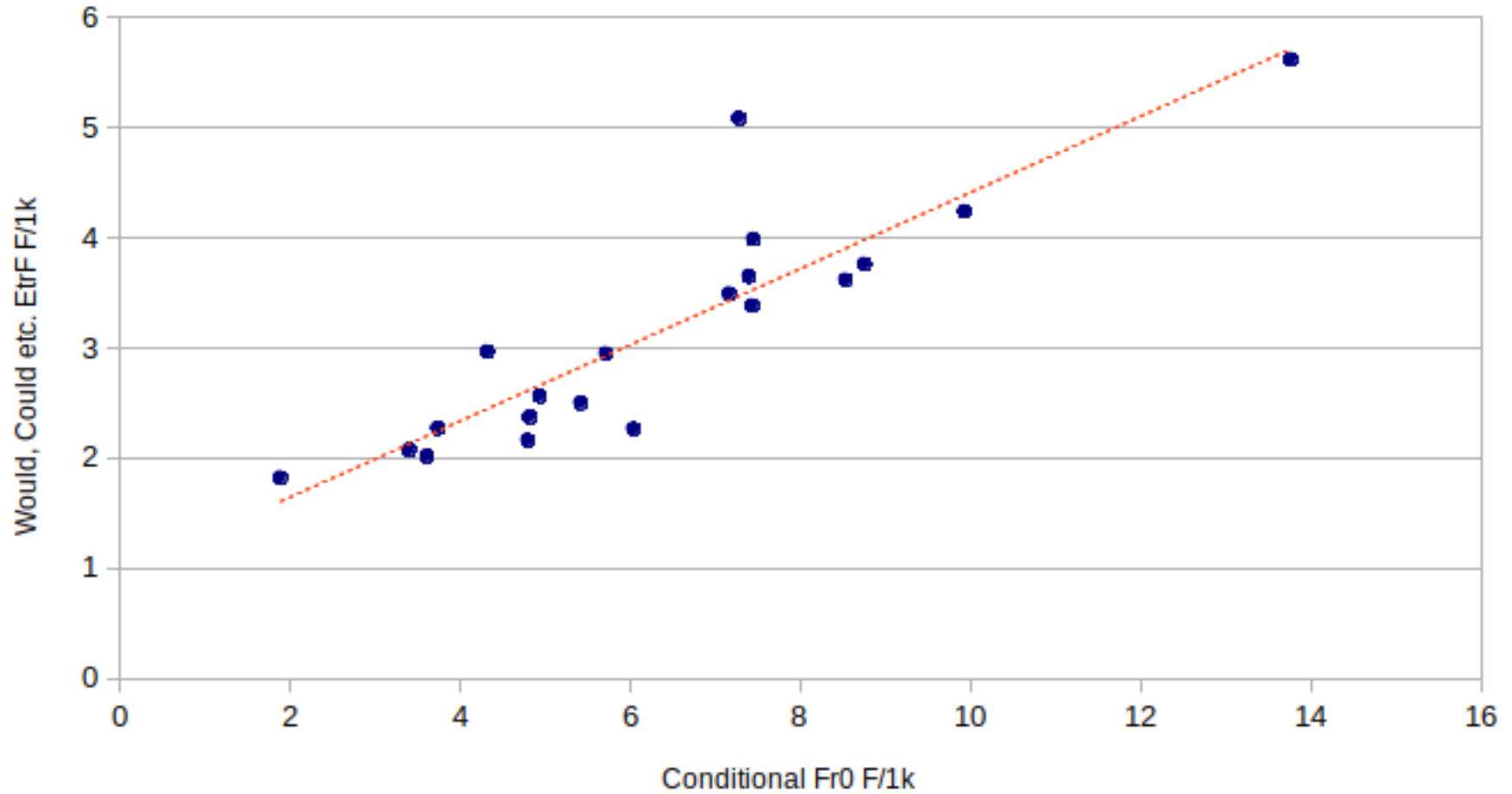
Moderate Negative

**adapted from: <https://financetrain.com/statistical-foundations-understanding-correlations/>*

3. Results & Analysis

Would, Could etc. EtrF ~ Conditional Fr0

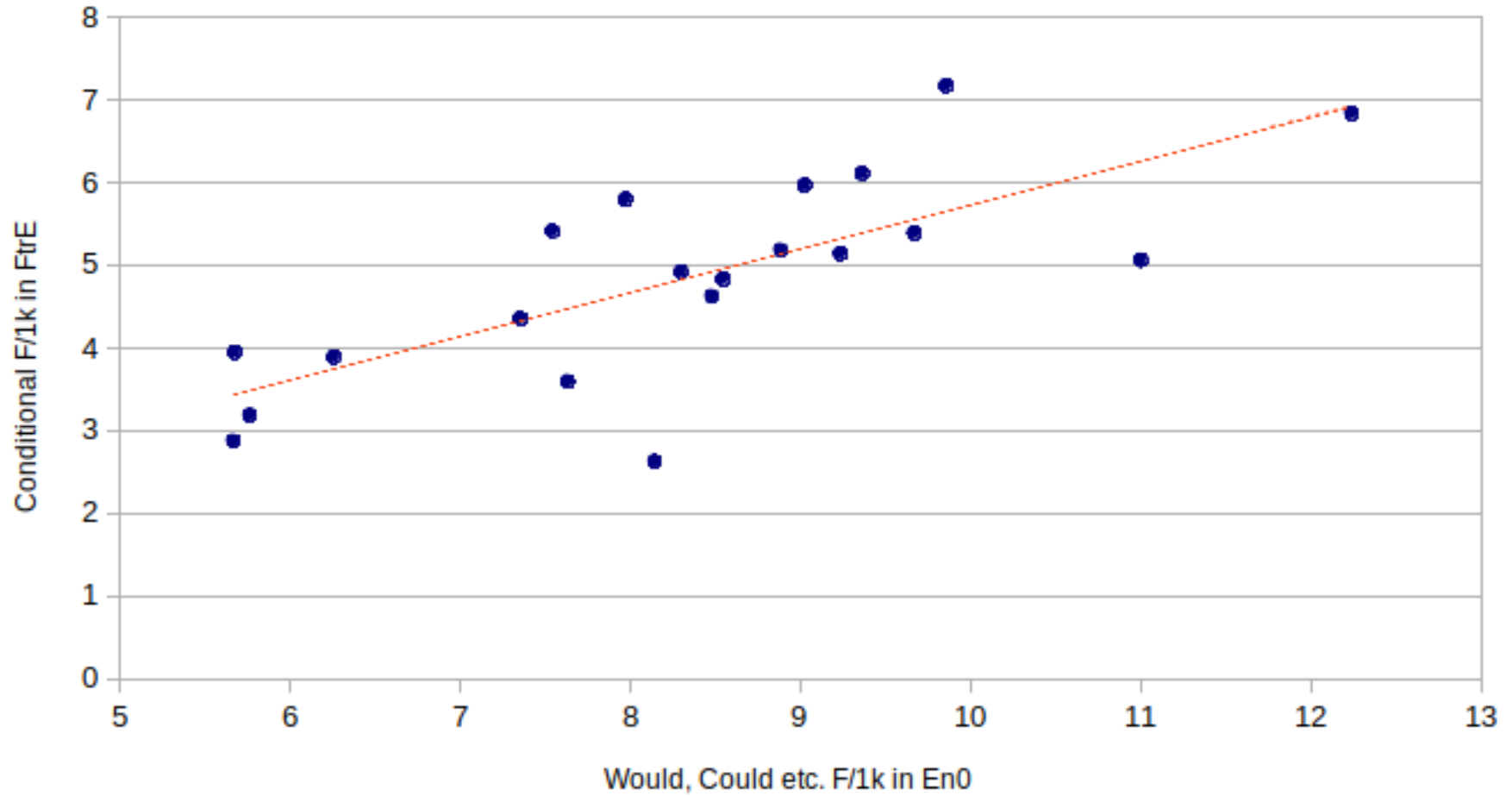
(rho=0.89, p<0.001)



3. Results & Analysis

FtrE Conditional ~ Would, Could etc. in En0

(rho=0.73, p<0.001)



3. Results & Analysis

- **In summary**

- *Would* exerts an influence on translators into FtrE
- Fr. Conditional exerts an influence on translators into EtrF

3. Results & Analysis

- But, isn't it normal for translators to follow/imitate the source-text?
- What does it mean to “follow” the source text? at what level? (*semantic? syntactic? stylistic? ...*)
 - *If the source-text follows the stylistic norms of the source-language ...*
 - *i.e. the intentio auctoris (or intentio operis) is not to create an effect of strangeness, surprise ...*
 - *Then shouldn't the target-text also follow the norms of the target-language?*
 - But do translations follow target-language norms?
 - And how do we define “norms”?

3. Results & Analysis

- Frequency

- What can frequencies show us?

- *For example,*

- Would, Could, etc. average frequency = 8.14/1k words

- Conditionnel average frequency = 2.96/1k words

- *The frequency of a word, expression, structure varies considerably, depending on:*

- person (style)

- context (subject, register)

- discourse type

- etc.

- *BUT ...*

- we inherit a language from our ancestors and use it with our peers

- we can innovate ... to a certain extent

- we make linguistic choices ... from what's available

- our individual choices tend to gravitate around a common center

3. Results & Analysis

1st indicator: Would, Could vs. Fr. Conditional

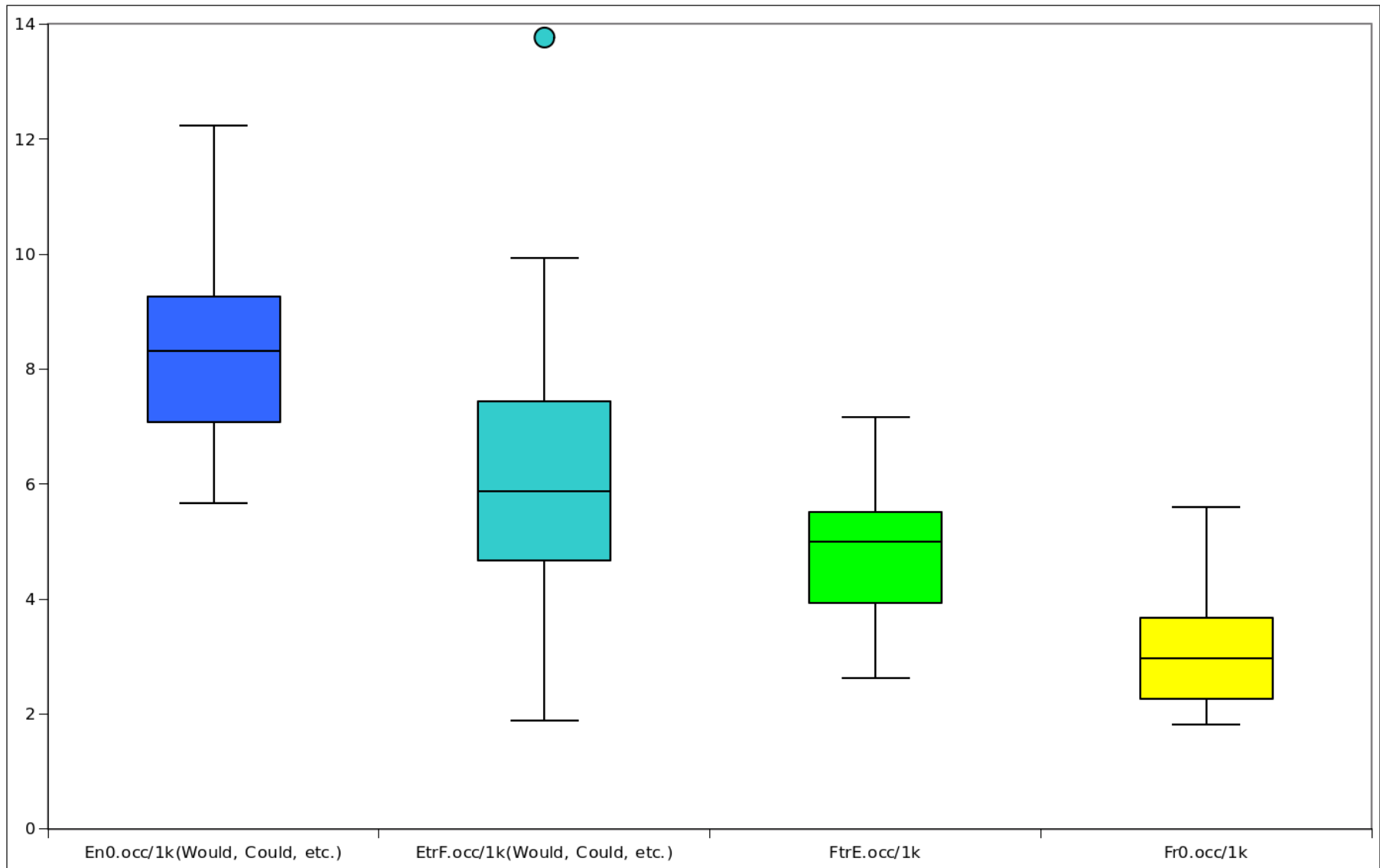
Question 2:

Does EtrF have the same characteristics as En0?

Does FtrE have the same characteristics as Fr0?

3. Results & Analysis

Fig. 1 Frequency of conditional forms in En0, EtrF, FtrE and Fr0.



3. Results & Analysis

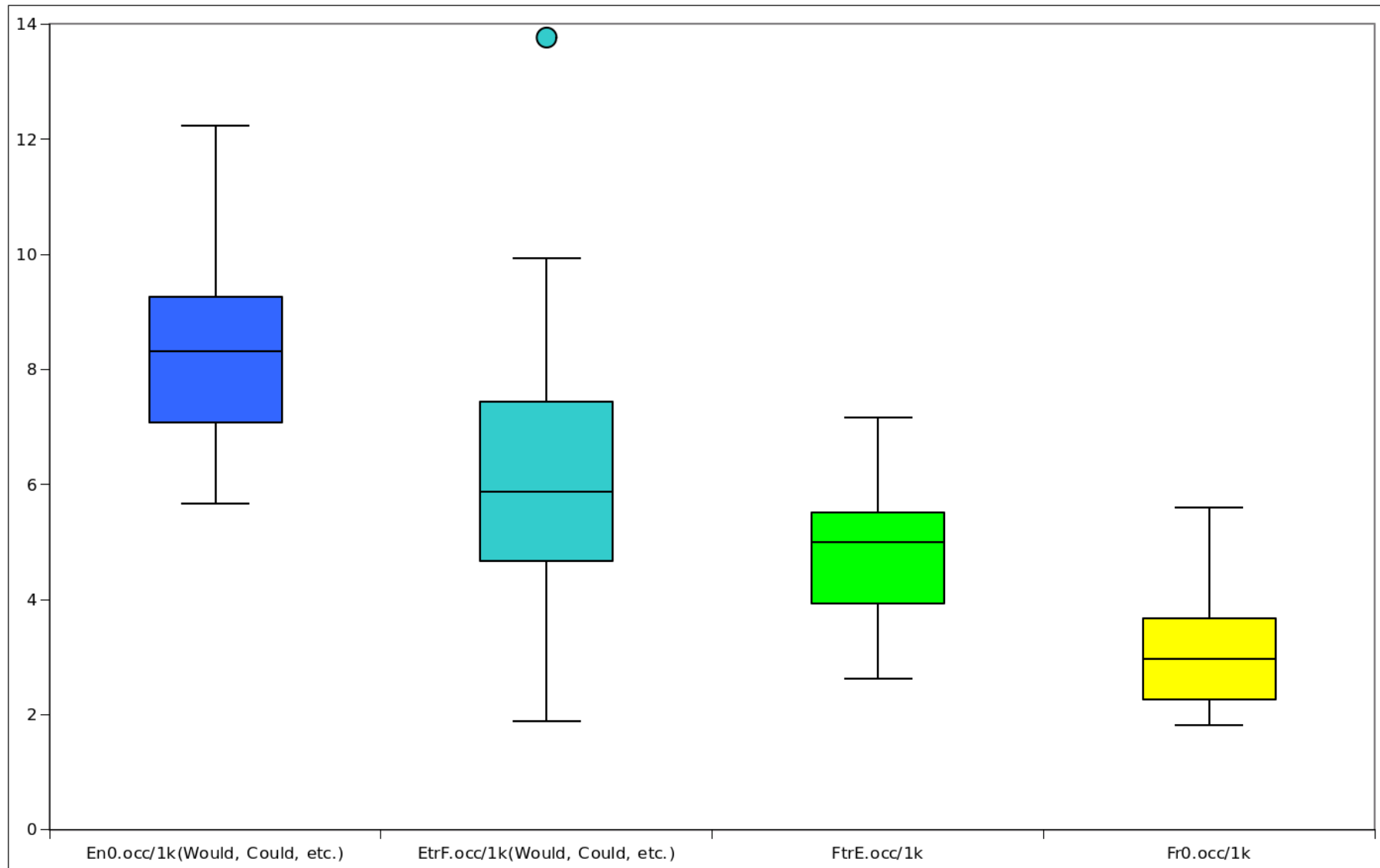
- **Wilcoxon-Mann-Whitney test**

- A “contest” between Group A vs. Group B
 - *Out of all possible combinations, how many times does a member of Group A beat a member of Group B (and vice versa)?*
 - *How likely is such an outcome, by chance, if the two groups were equal?*
- H_0 = Sample A and Sample B come from populations with equal distributions (or from the same population)
- Answers questions like:
 - *Do men tend to be taller than women?*
 - *Are Boeing 737s more reliable than Airbus A320s?*
 - *Is Drug X more effective than a placebo?*

 - *Do Translators use Would/Conditional more (or less) than Authors?*

3. Results & Analysis

Fig. 1 Frequency of conditional forms in En0, EtrF, FtrE and Fr0.



- Is FtrE significantly different from Fr0? YES (U=340, $p < 0.001$)
- Is EtrF significantly different from En0? YES (U=95, $p = 0.004$)

3. Results & Analysis

2nd indicator: En. Hear vs. Fr. Entendre

Question 1:

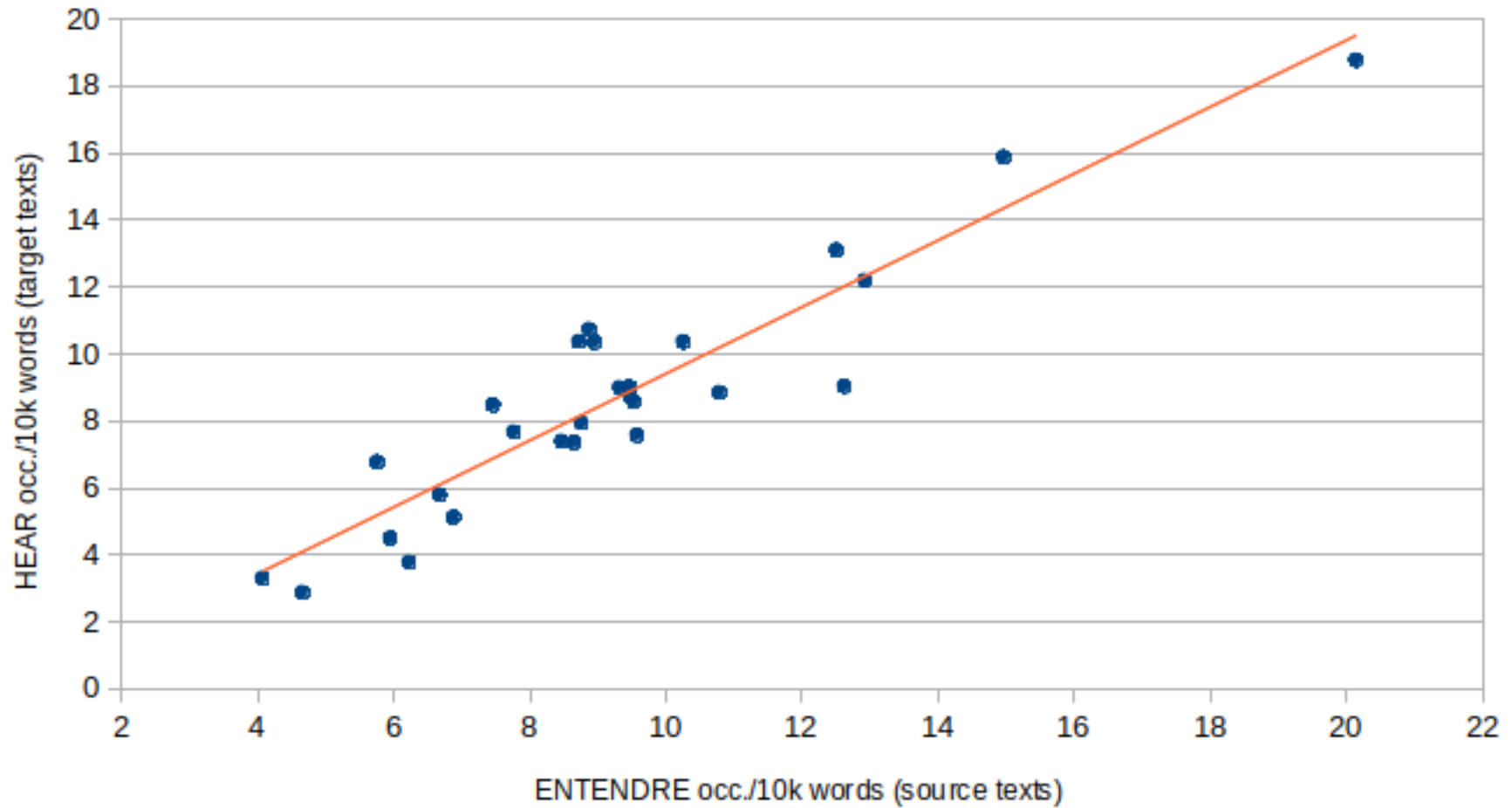
Does En0 exert an influence on FtrE?

Does Fr0 exert an influence on EtrF?

3. Results & Analysis

ENTENDRE (Fr0) ~ HEAR (EtrF)

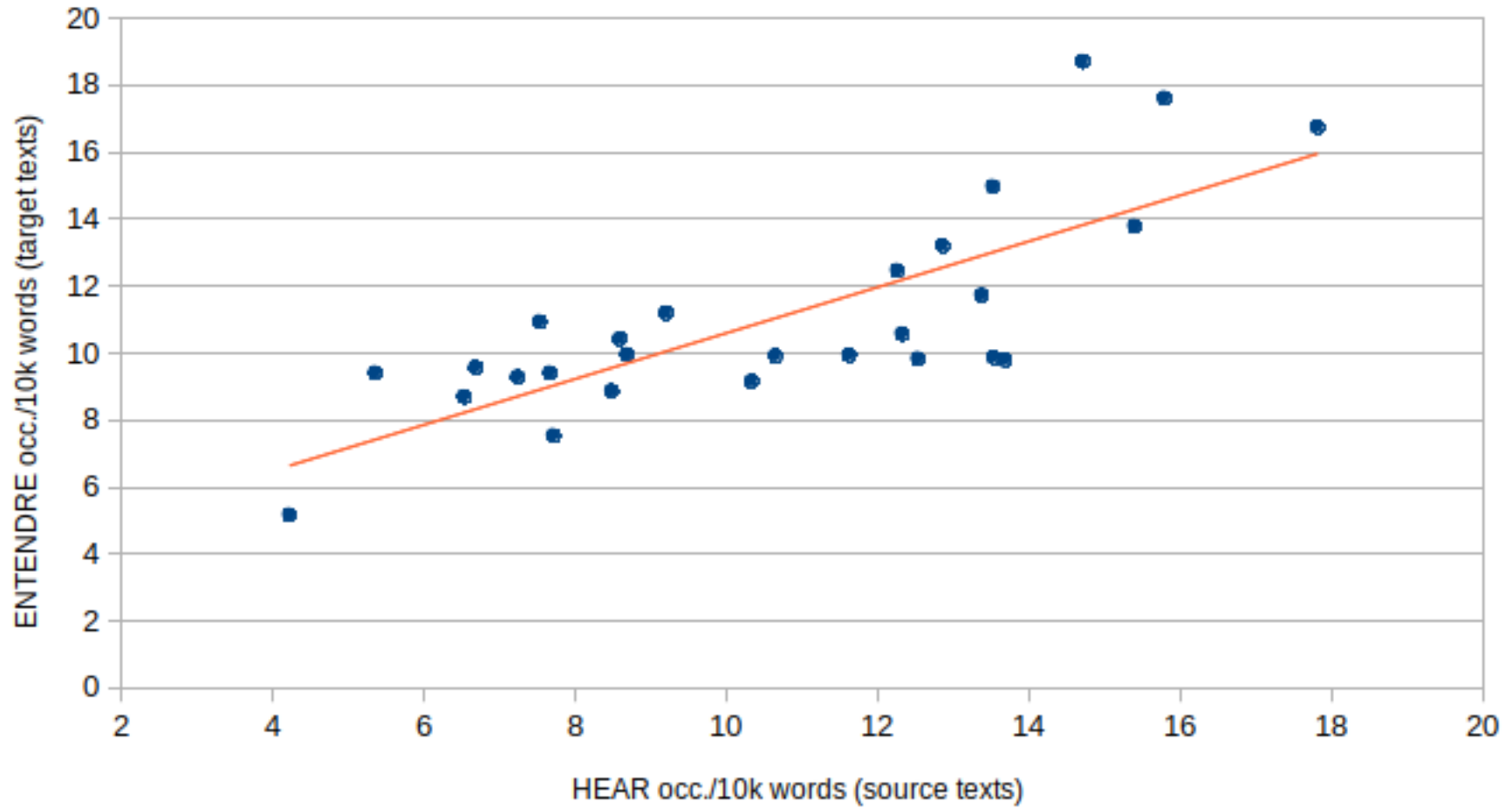
(rho=0.86, p<0.001)



3. Results & Analysis

HEAR (En0) ~ ENTENDRE (FtrE)

(rho=0.75, p<0.001)



3. Results & Analysis

- **In summary**
 - *Hear* exerts an influence on translators into FtrE
 - *Entendre* exerts an influence on translators into EtrF

3. Results & Analysis

2nd indicator: En. Hear vs. Fr. Entendre

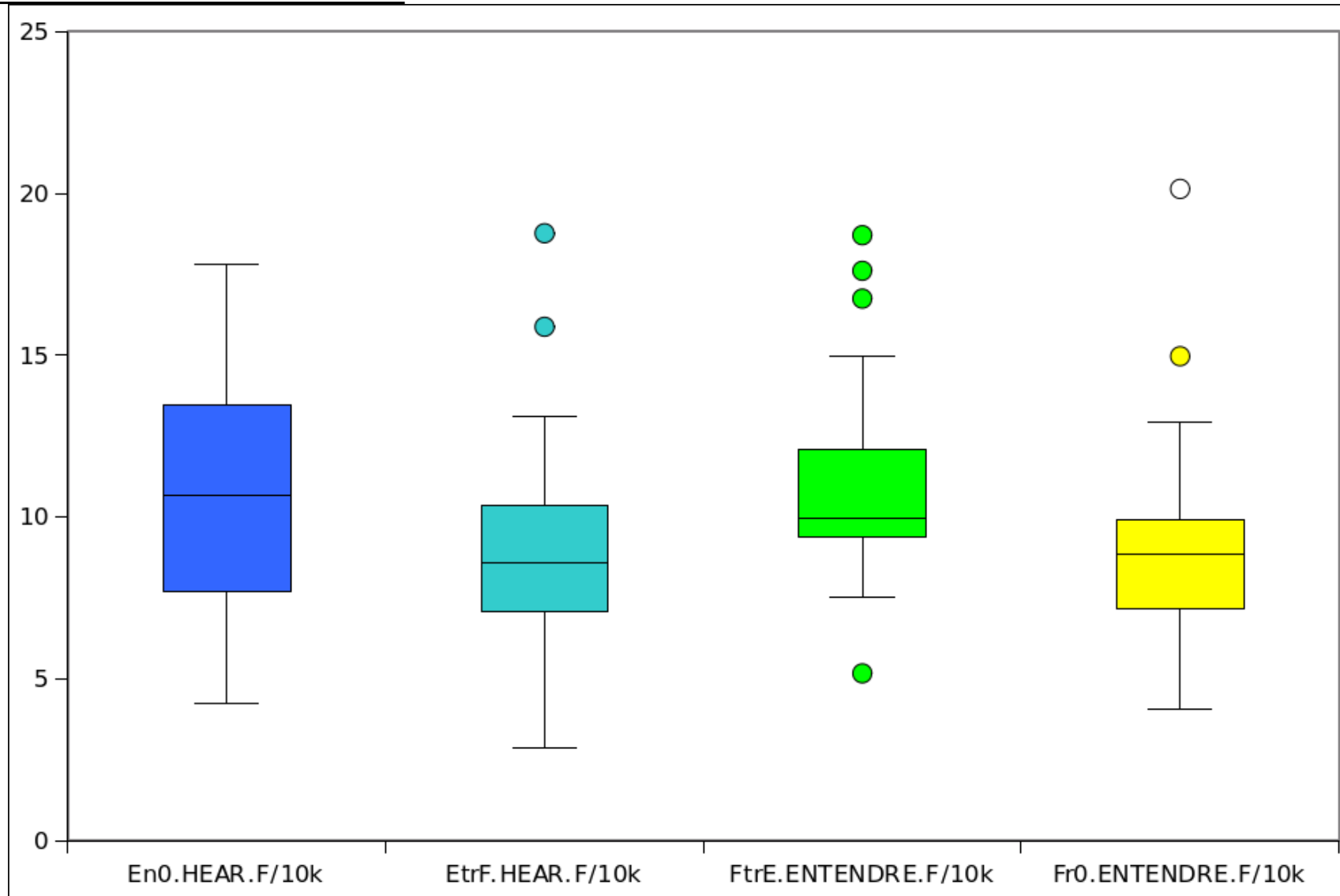
Question 2:

Does EtrF have the same characteristics as En0?

Does FtrE have the same characteristics as Fr0?

3. Results & Analysis

- HEAR & ENTENDRE**



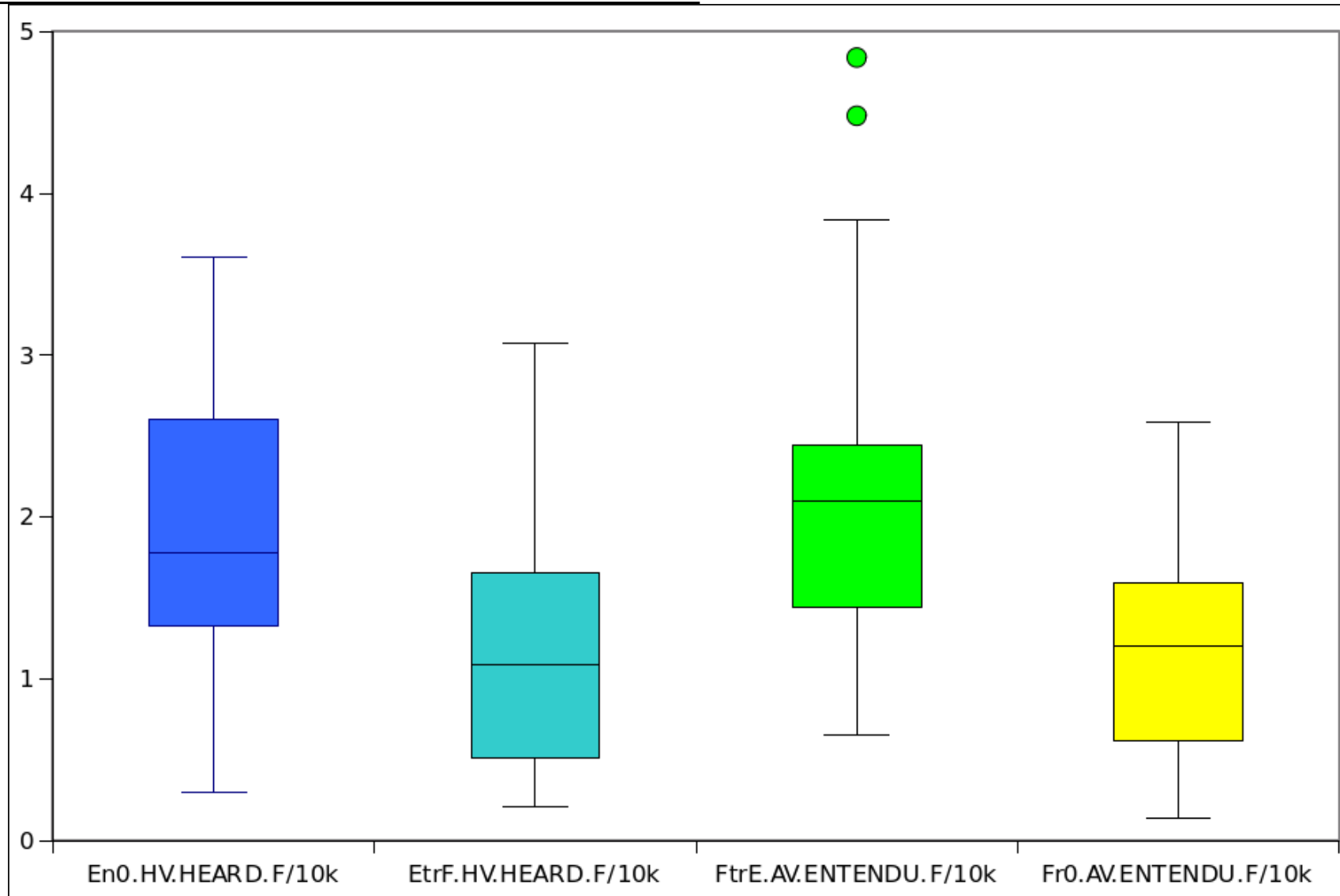
- Is EtrF a distinct sub-species from En0? U=251, p=0.05, Yes(?)
- Is FtrE a distinct sub-species from Fr0? U=519.5, p<0.01, YES

3. Results & Analysis

- **En. Hear vs. Fr. Entendre = Shining-through**
 - Translators tend to ...
 - *neglect 'Hear' in En.*
 - *overuse 'Entendre' in Fr.*
 - When? What contextual factors come into play?
 - *perfect aspect: Have+Heard, Avoir+Entendu*
 - +1st person: “I have heard”
 - *passive voice: Be+Heard, Être+Entendu*

3. Results & Analysis

- HAVE+HEARD & AVOIR+ENTENDU**



- Is EtrF a distinct sub-species from En0? U=183, p=0.002, YES
- Is FtrE a distinct sub-species from Fr0? U=582, p<0.001, YES

3. Results & Analysis

- **HAVE+HEARD (En0)**

- Most frequent n-gram of “hear” in En0: “I have hear”

- *recurs in 25/27 authors (92.5%), 151 occurrences*

The screenshot shows a software interface for analyzing text. At the top, there is a menu bar with 'File', 'Global Settings', 'Tool Preferences', and 'Help'. Below the menu bar, there are several tabs: 'Concordance', 'Concordance Plot', 'File View', 'Clusters/N-Grams', 'Collocates', 'Word List', and 'Keyword List'. The 'Concordance' tab is active, displaying a table of results for the search term 'hear'. The table has columns for 'Rank', 'Freq', 'Range', and 'Cluster'. The first row shows a rank of 1, a frequency of 151, a range of 25, and the cluster 'i have hear'. Below the table, there are search options: 'Search Term' (set to 'hear'), 'Cluster Size' (Min. 3, Max. 5), 'Sort by' (set to 'Range'), and 'Search Term Position' (set to 'On Left'). A 'Clone Results' button is located at the bottom right of the interface.

Rank	Freq	Range	Cluster
1	151	25	i have hear
2	71	23	have hear of
3	36	22	have never hear
4	75	22	he have hear
5	57	21	he hear the
6	35	20	hear it .
7	31	19	never hear of
8	28	19	to hear .
9	44	18	" i hear
10	74	18	. i hear
11	43	18	have hear the
12	32	17	do not hear
13	32	17	hear of it
14	44	17	to hear the
15	46	16	, and hear
16	61	16	could hear the
17	35	16	have not hear
18	30	16	he hear a

3. Results & Analysis

- **HAVE+HEARD (EtrF)**

- n-gram “I have hear” in EtrF

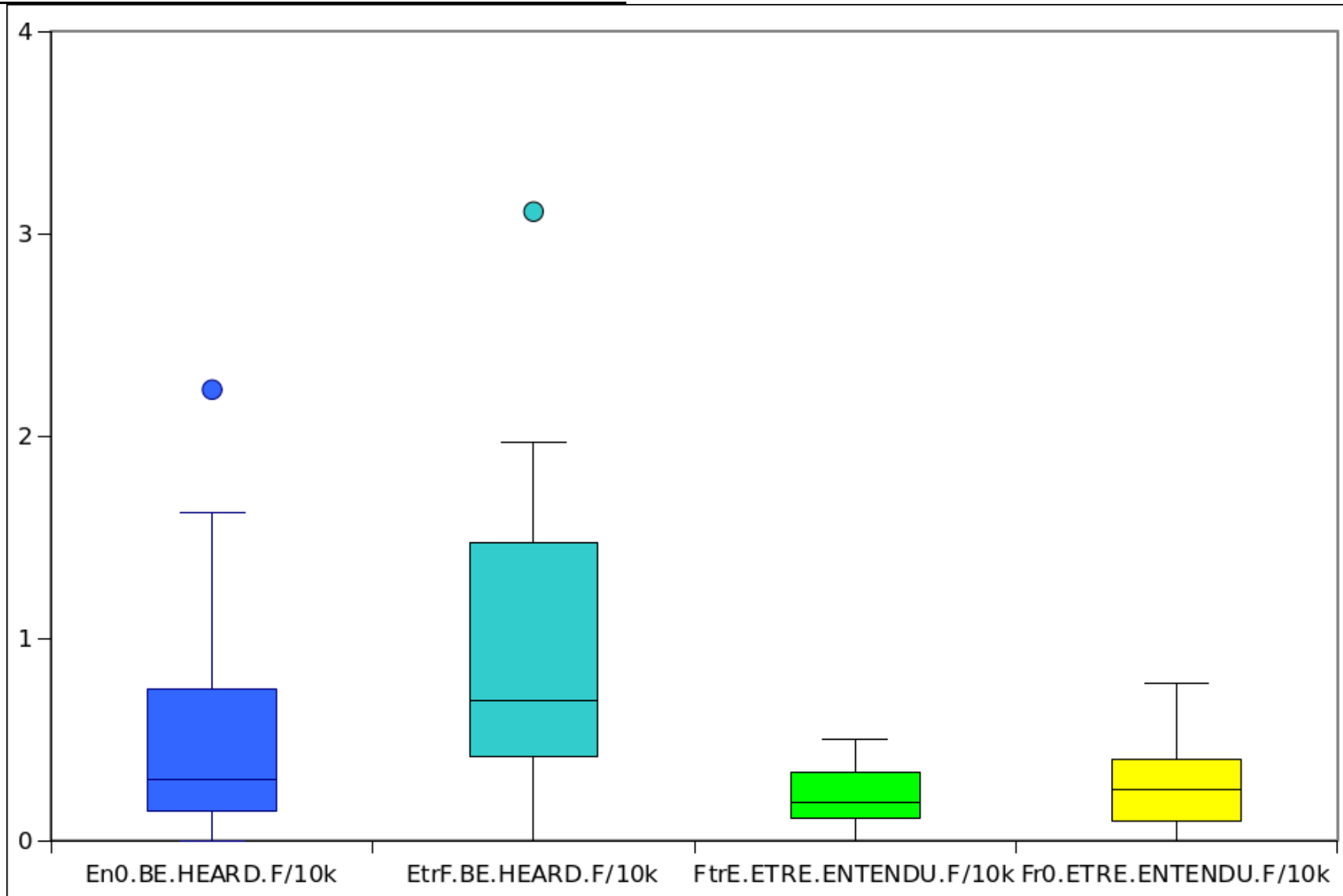
- *recurs in 14/27 authors (52% vs. 92.5%), 40 occurrences (vs. 151)*

The screenshot shows a software interface with a menu bar (File, Global Settings, Tool Preferences, Help) and a toolbar with tabs for Concordance, Concordance Plot, File View, Clusters/N-Grams, Collocates, Word List, and Keyword List. The main window displays a table of search results for the term 'hear'. The table has columns for Rank, Freq, Range, and Cluster. The search term is 'hear', and the cluster size is set to 3-5. The search term position is set to 13. The results list 18 clusters, with the top ones being 'to hear the', 'do not hear', 'do you hear', 'have hear the', 'he hear the', 'hear the sound', 'hear the sound of', '. i hear', 'have not hear', 'he have hear', '" i hear', 'when he hear', 'be hear .', 'could hear the', 'hear in the', 'i have hear', 'to be hear', and 'vou hear ?'. The interface also shows a list of corpus files on the left and a 'Clone Results' button at the bottom right.

Rank	Freq	Range	Cluster
1	45	21	to hear the
2	42	19	do not hear
3	49	19	do you hear
4	33	19	have hear the
5	55	17	he hear the
6	27	17	hear the sound
7	27	17	hear the sound of
8	40	16	. i hear
9	25	16	have not hear
10	53	16	he have hear
11	20	15	" i hear
12	36	15	when he hear
13	52	14	be hear .
14	27	14	could hear the
15	26	14	hear in the
16	40	14	i have hear
17	31	14	to be hear
18	27	14	vou hear ?

3. Results & Analysis

- BE+HEARD & ÊTRE+ENTENDU**



- Is EtrF a distinct sub-species from En0? U=497.5, p=0.02, Yes
- Is FtrE a distinct sub-species from Fr0? U=342.5, p=0.71, NO

3. Results & Analysis

- BE+HEARD En0 authors

- sequence “voice+be+heard” in En0: 2 authors/27, 2 occurrences

The screenshot shows a software interface for searching through a corpus. The top menu includes 'File', 'Global Settings', 'Tool Preferences', and 'Help'. Below the menu is a toolbar with options: 'Concordance', 'Concordance Plot', 'File View', 'Clusters/N-Grams', 'Collocates', 'Word List', and 'Keyword List'. The main window displays 'Concordance Hits 2' with two results. The first result (Hit 1) is from 'ReidM_Fi' and contains the text 'from declaring it in speech . Before she could reply , a voice was heard outside the door , accompanied by'. The second result (Hit 2) is from 'CollinsW_' and contains 'ething did happen . The landlord was called for . Angry voices were heard up-stairs . The mechanic suddenl'. The search term is '\S+/voice(\S+/(VH\S?|MD)\S+(\S+/RB[RS])'. The search window size is set to 200. The interface also shows a list of corpus files on the left, a search term input field, and a 'Clone Results' button at the bottom right.

Hit	KWIC	File
1	from declaring it in speech . Before she could reply , a voice was heard outside the door , accompanied by	ReidM_Fi
2	ething did happen . The landlord was called for . Angry voices were heard up-stairs . The mechanic suddenl	CollinsW_

Search Term: Words Case Regex
Search Window Size: 200
Show Every Nth Row: 1
Kwic Sort: Level 1 1R, Level 2 0, Level 3 0

3. Results & Analysis

- BE+HEARD EtrF translators

- sequence “voice+be+heard” in EtrF: 17 authors/27 (63%), 28 occurrences

The screenshot shows a software interface for text analysis. At the top, there are menu items: File, Global Settings, Tool Preferences, and Help. Below the menu is a toolbar with buttons for Concordance, Concordance Plot, File View, Clusters/N-Grams, Collocates, Word List, and Keyword List. The main window is titled 'Concordance Hits 28'. It features a table with three columns: 'Hit', 'KWIC', and 'File'. The 'Hit' column contains line numbers from 0 to 24. The 'KWIC' column contains snippets of text with the search term 'voice+be+heard' highlighted in red and blue. The 'File' column lists the source files, such as 'HugoV_Q' and 'Vernej_Ile'. At the bottom of the window, there is a search control panel with options for 'Words', 'Case', and 'Regex'. The search term is '\S+/voice(\S+/(VH\S?|MD)/AS+(\S+/RB[RS]):'. There are also buttons for 'Start', 'Stop', 'Sort', and 'Show Every Nth Row' (set to 1). A 'Search Window Size' is set to 200. At the bottom left, there are statistics: 'Total No. 27' and 'Files Processed'. At the bottom right, there is a 'Clone Results' button.

Why do EtrF. translators overuse “Be+Heard”?

3. Results & Analysis

- Entendre+Voix in Fr0 authors

The screenshot displays a concordance tool interface with the following components:

- Menu Bar:** File, Global Settings, Tool Preferences, Help
- Corpus Files:** A list of author names including AudouxM_Marie, Barbusse_LeFev, BazinR_LesOber, BenoitP_Atlantic, BourgetP_LaDuc, DaudetA_Limmo, FlaubertG_Educ, FranceA_LesDie, GaboriauE_LArg, GourmontR_Sixt, Gyp_BijouFR_pg, HemonL_MariaC, HugoV_Quatre, HuysmansJK_La, LeblancM_LesDe, LerouxG_Rouille, LotiP_LesDernie, LouysP_Aphrodi, MalotConscienc, MassenetJ_MesS, MirbeauO_Journ, ProustM_DuCote, RollandR_JeanCl, SouvestreP_Fan, VanderemF_Les, Vernej_IleMyste, ZolaE_LesTroisV.
- Search Term:** `+) {0,3} \S+ /entendre(\w\S+) {0,3} \S+ /voix` (Advanced)
- Search Window Size:** 200
- Search Options:** Start, Stop, Sort, Show Every Nth Row: 1
- Kwic Sort:** Level 1: 1R, Level 2: 0, Level 3: 0
- Concordance Hits:** 21
- Table:** A table with columns 'Hit', 'KWIC', and 'File'. It lists 19 hits with corresponding text snippets and author names.
- Buttons:** Clone Results

Hit	KWIC	File
1	... t aussi la cuisine bien entendu , était éclairée , on entendait des voix . « J' ai passé en tendant l'	Barbusse
2	ère filtrer dans les interstices des planches , et on entendait des voix bruisser . -- C' est de l' autre	Barbusse
3	e bruit eut cessé de sonner sous les branches , on entendit la voix de M. Ulrich , déjà loin , qui	BazinR_L
4	leur dit un garçon de bureau taillé en force . On entendait sa voix brutale , car il était , non pas	Gaboriau
5	imer . On n' attendait plus que la catastrophe . On entendait les petites voix répéter : Maman ! maman ! On	HugoV_Q
6	sant au préfet , sortirent pour écarter la foule , on entendit la voix éraillée d' un camelot qui hurlait : «	LeblancM
7	, s' agitaient dans le brouillard . Et tout à coup on entendit la voix de Koupriane qui donnait des ordres , ex	LerouxG_
8	aux bruissait , sifflait , chantait dans le jardin . On entendait des voix de femmes qui parlaient et qui riaient	LouysP_A
9	mes , le silence se rétablit immédiatement , et l' on entendit la voix du plus ancien des pensionnaires , Henne	Massenet
10	, à la cuisine , au jardin , des heures entières , on entend sa voix qui glapit ... Ah ! qu' elle est	MirbeauC
11	était malade ... On s' agitait derrière le rideau , on entendait un bruit de voix et de pas précipités . C'	RollandR_
12	é à la porte l' arracha à son engourdissement . On entendait un murmure de voix indistinctes . Il se rappela	RollandR_
13	t cela ne troublait point son caquet inlassable . On entendait sans cesse sa voix aiguë qui racontait des histo	RollandR_
14	« Chut ! dit - elle , on m' appelle . » On entendait la voix d' Amalia . Rosa demanda : « Veux	RollandR_
15	' est tout ce que vous méritez ! ... » De la rue on entendait sa voix . Les gens s' arrêtaient pour écouter .	RollandR_
16	d' une voiture écrasèrent le gravier du jardin . On entendait des appels de voix , des rires . Georges arrivai	Vanderer
17	développa , et , des profondeurs de ce nuage , on entendit une voix délirante qui répétait : « Du tabac !	Vernej_Ile
18	t' aide à le refaire ! » Enfin , le 25 août , on entendit la voix de Nab qui appelait ses compagnons .	Vernej_Ile
19	et d'ailleurs , la porte resta grande ouverte on entendait leurs grosses voix confuses . Le général de Ro	ZolaF_Le

3. Results & Analysis

- Entendre+Voix in FtrE translators**

The screenshot shows a software interface for analyzing text. At the top, there is a menu bar with 'File', 'Global Settings', 'Tool Preferences', and 'Help'. Below the menu is a toolbar with buttons for 'Concordance', 'Concordance Plot', 'File View', 'Clusters/N-Grams', 'Collocates', 'Word List', and 'Keyword List'. The main window is titled 'Concordance Hits 7' and displays a list of search results. The results are organized into three columns: 'Hit', 'KWIC', and 'File'. The 'Hit' column contains line numbers 1 through 7. The 'KWIC' column contains the search term 'entendre' in various contexts, with some words highlighted in blue. The 'File' column lists the source files for each hit, such as 'BellamyE', 'MansfieldK', 'StokerB', 'TwainM', 'WallaceL', and 'WellsHG'. At the bottom of the window, there is a search control panel. It includes a 'Search Term' field with the text '\b[Oo]n\S+(\ \w\S+){0,3} \S+/entendre(\ \w\S+)', a 'Search Window Size' dropdown set to '200', and buttons for 'Start', 'Stop', and 'Sort'. There is also a 'Show Every Nth Row' dropdown set to '1' and a 'Kwic Sort' section with checkboxes for 'Level 1', 'Level 2', and 'Level 3', all set to '0'. A 'Clone Results' button is located in the bottom right corner. On the left side of the window, there is a 'Corpus Files' list with various author names and file names.

Hit	KWIC	File
1	une cloche sonna , et quelques minutes après ,	BellamyE
2	uement du fouet . Elle approcha assez pour qu'	MansfieldK
3	r frapper ; et il n' était pas vraisemblable que l'	StokerB
4	èrent sur ceux du redoutable monarque ; puis ,	TwainM
5	r l' arrière-pont , redoublait d' éclat ; à l' avant ,	WallaceL
6	rchaient devant lui s' arrêchèrent brusquement .	WallaceL
7	eux qui n' ont rien perdu d' un beau spectacle .	WellsHG

So how do En. authors/translators describe the perception of a person's voice?

3. Results & Analysis

- How do En. authors describe the perception of a “voice”?

The screenshot shows a concordance search tool interface. The search term is `\S+/voice(\w\S+){0,3} \S+/come`. The results are displayed in a table with columns for Hit, KWIC, and File. The KWIC column highlights the search term and its context in various colors. The search window size is set to 200. The interface includes a menu bar (File, Global Settings, Tool Preferences, Help), a toolbar (Concordance, Concordance Plot, File View, Clusters/N-Grams, Collocates, Word List, Keyword List), and a sidebar with a list of corpus files. The search window size is set to 200. The interface includes a menu bar (File, Global Settings, Tool Preferences, Help), a toolbar (Concordance, Concordance Plot, File View, Clusters/N-Grams, Collocates, Word List, Keyword List), and a sidebar with a list of corpus files.

Hit	KWIC	File
1	" Were you at the Bells ' last week ? " the voice came again . It sounded tired . Leila wondered whether sh	Mansfield
2	ce a fine cold blooded pair . " The grating voice came back . " I hold a theory that all of the ill	CoxE_Ou
3	ll you , " he roared . Only the echo of his voice came back . " Elaine ! " he shouted . Had she followed I	WallaceE
4	All hail , Thou Mighty One ! " A velvety bass voice came booming over the water . Great Scott ! Damnatio	Mansfield
5	orry . " Before Dundas could reply Dick 's voice came booming and echoing down the gallery . " Dundas ,	CoxE_Ou
6	has brought me home . " The owner of the voice came forward into the light . " Oh ! Dundas , come in ,	CoxE_Ou
7	not speak to me ? ' ' It is an order . ' The voice came from behind him and he started . ' Very good . Bu	KiplingR_
8	nd in a more compact group . Suddenly a voice came from inside the hall : it was distinct and loud , yet no	HopeA_R
9	entials meant the least work . Presently a voice came from the kitchen . " Hector , old man , there 's a	CoxE_Ou
10	a standstill at the gate , and as he did so a voice came from the shadow of the avenue beyond . " Is that yo	CoxE_Ou
11	at 's the matter with me , anyhow ? " The voice came from the sofa across the room , and its tones broug	StokerB_
12	ant had got out and opened the door and a voice came from within — a clear soft voice speaking in some to	BuchanJ_
13	uld no longer endure the waiting . Sapt 's voice came harsh and grating . " Well ? " he cried . " Which	HopeA_R
14	her hands , and commenced to speak . His voice came in a low , monotonous monotone , but very distinctly	CoxE_Ou
15	dened her . Then she spoke again , and her voice came in a little more than a whisper . " Our people knew c	CoxE_Ou
16	hymn ceased . Then after a long interval a voice came out of heaven . " Vat id diss blace here galled itself	WellsHG_
17	hard to find in any other way . Usually her voice came over the wire as something fresh and cool as if a div	Fitzgeral
18	dusk she could see him straighten , and his voice came sharp as he ignored the ever-present parental back	LewisS_F
19	them facing the Prime Minister . Then his voice came so naturally and distinctly that again only the restr	CoxE_Ou

Search Term: Words Case Regex
Search Window Size: 200
Advanced
Start Stop Sort Show Every Nth Row 1
Kwic Sort
Level 1 2R Level 2 3R Level 3 4R
Clone Results

3. Results & Analysis

- And En. translators?

File Global Settings Tool Preferences Help

Corpus Files

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Concordance Hits 16

Hit	KWIC	File
1	n the darkness . The group from which the	HugoV_Q
2	itself . It made me tremble , and when the	AudouxM
3	whether it was really she who spoke . The	RollandR
4	ld be an end of morality on earth ... The	RollandR
5	and the situation — I understand . " His	BazinR_L
6	of the wine-cellar is shaken , and a man 's	Barbusse
7	e wittiest of all shouted " Encore ! " A bass	RollandR
8	fear of hearing himself addressed by other	RollandR
9	I used to tremble when I heard it . The full	AudouxM
10	ugh in her prayers . Even at a distance the	HemonL
11	ong the dark walls on the sides whence the	BenoitP_
12	Silence fell , terrible silence . Guillaume 's	ZolaE_Le
13	udden departure . " " Me ! " she said in a	Gaboriau
14	n a stick and towards him , drinking in the	Barbusse
15	shouts — except for the curiously piercing	Barbusse
16	l as though she were overcome , and , in a	AudouxM

Search Term Words Case Regex Search Window Size 200

Show Every Nth Row 1

Kwic Sort

Level 1 2R Level 2 3R Level 3 4R

And how do Google and DeepL translate the sequence
“On+entendre+voix”?

Le café était servi dans le petit salon bleu et argent, où fleurissait une admirable corbeille de roses jaunes, cette passion que la baronne avait pour les fleurs, et qui changeait l'hôtel en un continuel printemps. Tout de suite, leurs tasses fumantes à la main, Duvillard emmena Fonsègue dans son cabinet, pour fumer un cigare, en causant librement ; et, d'ailleurs, la porte resta grande ouverte, **on entendait leurs grosses voix confuses**. Le général de Bozonnet, ravi d'avoir trouvé en madame Fonsègue une personne sérieuse et résignée, écoutant sans jamais interrompre, lui racontait la très longue histoire de la femme d'un officier qui avait suivi son mari dans toutes les batailles, en 1870. Hyacinthe ne prenait pas de café, qu' il appelait avec mépris un breuvage de concierge.

(E. Zola)

Coffee was served in the small blue and silver lounge, where an admirable basket of yellow roses blossomed, that passion the baroness had for flowers, and which changed the hotel into a continuous spring. Immediately, with their steaming cups in hand, Duvillard took Fonsègue to his office to smoke a cigar, chatting freely; and, moreover, the door remained wide open, **their big confused voices were heard**. General de Bozonnet, delighted to have found in Madame Fonsègue a serious and resigned person, listening without ever interrupting, told her the very long story of the wife of an officer who had followed her husband in all battles in 1870. Hyacinth did not take coffee, which he contemptuously called a janitor's drink.

(DeepL)

Le café était servi dans le petit salon bleu et argent, où fleurissait une admirable corbeille de roses jaunes, cette passion que la baronne avait pour les fleurs, et qui changeait l'hôtel en un continuel printemps. Tout de suite, leurs tasses fumantes à la main, Duvillard emmena Fonsègue dans son cabinet, pour fumer un cigare, en causant librement ; et, d'ailleurs, la porte resta grande ouverte, **on entendait leurs grosses voix confuses**. Le général de Bozonnet, ravi d'avoir trouvé en madame Fonsègue une personne sérieuse et résignée, écoutant sans jamais interrompre, lui racontait la très longue histoire de la femme d'un officier qui avait suivi son mari dans toutes les batailles, en 1870. Hyacinthe ne prenait pas de café, qu' il appelait avec mépris un breuvage de concierge.
(E. Zola)

The coffee was served in the little blue-and-silver room, where an admirable basket of yellow roses was blooming, that passion which the baroness had for the flowers, and which changed the hotel into a continual spring. Immediately, with their steaming cups in their hands, Duvillard took Fonseca to his cabinet to smoke a cigar, talking freely; and besides, the door was wide open, and **their thick, confused voices were heard**. General de Bozonnet, delighted to have found in Madame Fonseca a serious and resigned person, listening without ever interrupting, told him the very long story of the wife of an officer who had followed her husband in all battles, in 1870. Hyacinthe did not take coffee, which he scornfully called a concierge's drink.

(Google)

Conclusions

- **1. NMT systems trained on parallel corpora learn to**
 - imitate translators
 - produce Translationese
- **2. Target-texts, on the whole, do not conform to target-language norms**
 - grammatical norms (e.g. Would, Conditional)
 - lexical norms (e.g. Hear, Entendre)
 - grammatical+lexical = stylistic
- **3. EtrF and FtrE are distinct sub-species or hybrids due to**
 - compensation strategies (= Frawley's “third code”, Baker's “universals”)
 - interference/shining-through
- **4. To improve translation quality (human or MT)**
 - *linguists* have to identify disparities in Translationese
 - *translators* need to draw inspiration from original, *ex nihilo* texts
- **5. As long as MT learns from human translators, human translators will continue to define the upper limit (≠ Chess, Go, finite number of rules)**

Références

- Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*.
- Evert, S., & Neumann, S. (2017). *The impact of translation direction on characteristics of translated texts*
- Korzen, I., & Gylling, M. (2017). *Text structure in a contrastive and translational perspective*. in *Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation*
- Loock, R. (2019). *Rencontres du 'troisième code': Analyse de traductions automatiques réunies en corpus*. JE Corpus comparables, parallèles, spécialisés: quelle(s) approche(s) du corpus en traduction?
- Sennrich, R. & Haddow, B. (2016). *Linguistic Input Features Improve Neural Machine Translation*
- Shuttleworth, M. (2014). *Dictionary of translation studies*. Routledge.
- Van der Wees, M., Bisazza, A., & Monz, C. (2017). *Dynamic data selection for neural machine translation*.
- Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/software>
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Schmid, H. TreeTagger, Universitaet Stuttgart, <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

Thank you!

File Global Settings Tool Preferences Help

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Corpus Files

- BellamyE_Loo
- BensonR_Lord
- BuchanJ_Gree
- BurnettFH_Ser
- CollinsW_Moo
- ConanDoyleA
- CoxE_OutOfTr
- EliotG_Middler
- FitzgeraldFS_C
- HardyT_TessC
- HopeA_Ruperl
- JamesH_Wash
- JoyceJ_Dubline
- KiplingR_KimE
- LewisS_FreeAi
- LondonJ_Marti
- MansfieldK_Ga
- MorrowW_Ape
- ReidM_FingerC
- StevensonRL
- StokerB_Dra

Concordance Hits 142

Hit	KWIC	File
53	out after him: "Well, thank you , sir." He waited outside the d	JoyceJ_
54	eresting.' 'But I can't thank you enough.' 'There's one thing y	KiplingF
55	e till I see it in print.' ' Thank you . That goes straight to an ethn	KiplingF
56	ek, adding in English: ' Thank you verree much, my dear.' Kissi	KiplingF
57	red, "Good night, and thank you —oh, thank you." If Milt had b	LewisS_
58	t, and thank you—oh, thank you ." If Milt had been driving at tl	LewisS_
59	uld—" "Thanks, uh, thank you , sir, but I wouldn't like to do it	LewisS_
60	some water myself!" " Thank you , but I won't trouble you. Good	LewisS_
61	"Nothing else I want, thank you . D-don't bother me." "Why, I'	LewisS_
62	r your car?" "Oh yes, thank you . So nice of you to bother with	LewisS_

Search Term Words Case Regex Advanced Search Window Size

Show Every Nth Row

Kwic Sort

Level 1 Level 2 Level 3

Total No. 27
Files Processed