Ana Zwitter Vitez
Trojina, Institut de linguistique slovène appliquée*
UDK [811.163.6'322:324-051]:2-274.2

LE DÉCRYPTAGE DE L'AUTEUR ANONYME : L'AFFAIRE DES ÉLECTEURS EN SURVÊTEMENTS

1 INTRODUCTION

Durant les dernières décennies, les recherches relatives à l'attribution de l'auteur ont fait naître de nombreuses approches pluridisciplinaires. Cet épanouissement méthodologique est probablement dû à une demande croissante d'analyses textuelles dans les domaines du droit, de la criminologie, des études littéraires et des recherches mercatiques, sous l'influence des phénomènes suivants :

- plagiats (thèse de doctorat du ministre allemand Karl-Theodor zu Guttenberg, scénario du film *Avatar* signé par James Cameron),
- lettres de menace (George Bush, Nicolas Sarkozy, Jean-Paul Belmondo),
- discours d'incitation à la haine raciale (propos de Bruno Gollnisch sur la Shoah),
- pseudonymes littéraires (influence mutuelle entre Molière et Corneille, identité de l'auteur du blog *Belle de Jour: diary of a London call girl*),
- stratégies publicitaires (bases documentaires pour profiler les clients).

Dans le contexte slovène, l'opinion publique a vivement réagi à la suite de la publication d'un texte intitulé *Les électeurs en survêtements*[1] dont l'auteur n'a jamais pu être identifié. C'est pourquoi la présente étude entend comparer le texte controversé à 75 textes produits par 21 auteurs connus afin de répondre aux questions concernant :

- les similarités et la distance entre le texte examiné et les auteurs potentiels,
- les indices linguistiques à partir desquels on peut attribuer un texte à un auteur.

Pour ce faire, nous recourrons à une méthode de classification automatique se fondant sur les machines à vecteurs de support qui permettent de définir la distance ou la proximité entre différents textes selon les indices de lexique et de lisibilité textuelle.

2 ÉTAT DE L'ART DANS L'ATTRIBUTION DE L'AUTEUR

Depuis la fin du XIX^{e} siècle, le domaine de l'attribution de l'auteur n'a cessé de se développer. L'étude pionnière a été réalisée par Thomas Corwin Mendenhall (1887) qui a analysé la longueur des mots afin de différencier des langues et des auteurs

* *Adresse de l'auteur :* Trojina, Zavod za uporabno slovenistiko, Partizanska cesta 5, 4220 Škofja Loka, Slovénie. Mél : ana.zwitter@guest.arnes.si

1 Le texte est accessible sur le site suivant : http://www.delo.si/assets/media/other/20111211//Prispevek%20Toma%C5%BEa%20Majerja.pdf

particuliers. Il a démontré que les histogrammes de Shakespeare et de Marlowe étaient presque identiques, ce qui représentait une découverte captivante vu que la mort de Marlowe, quelques semaines avant la parution des premières œuvres de Shakespeare, restait inexpliquée.

De nos jours, les méthodes d'attribution de l'auteur et de détermination du profil de l'auteur sont en plein épanouissement en raison de la pertinence de leurs résultats dans les domaines du droit, de la criminologie, des études littéraires et de la mercatique. Les méthodes de classification de textes ont recours aux indices linguistiques suivants :

- **lexicaux** (Argamon/Levitan 2005),
- **syntaxiques** (Luyckx/Daelemans 2005),
- **sémantiques** (McCarthy *et al*. 2006) et
- **de caractères** (Stamatatos 2009).

Pour la langue slovène, la classification automatique a été utilisée dans deux analyses de l'attribution de l'auteur : la première s'appuie sur la longueur des phrases et des mots pour détecter les plagiats (Dović 2002), et la deuxième prend les mots fonctionnels comme indice linguistique pour déterminer l'auteur d'un texte littéraire (Limbek 2008).

3 L'AFFAIRE DES ÉLECTEURS EN SURVÊTEMENTS

La motivation de la présente étude réside dans les réactions de l'opinion publique suite à la publication d'un texte sur le site officiel du Parti Démocrate Slovène (SDS), quelques jours après sa défaite aux élections législatives. Le prétendu auteur de ce texte, signé « Tomaž Majer », explique son interprétation de la victoire du parti La Slovénie Positive (PS).

Les éléments du texte les plus cités par les médias concernaient l'interprétation que cet auteur donnait de la victoire du PS : elle serait le fait des électeurs étrangers qui se rendent dans les bureaux de vote habillés en survêtements et qui obéiraient aux consignes de vote en inscrivant dans le creux de leurs mains le numéro du candidat qu'ils doivent choisir. Quelques jours après cette publication controversée, la Commissaire à l'information a dénoncé l'auteur du texte pour incitation à la haine raciale, mais la cour a rejeté la plainte et Tomaž Majer n'a jamais été poursuivi. Malgré cette décision, l'opinion publique n'a cessé de spéculer sur le véritable auteur du texte.

4 HYPOTHÈSE ET RÉSULTATS ATTENDUS

Pour pouvoir comparer le texte d'origine inconnue aux auteurs potentiels, nous avons formulé notre hypothèse dans les termes suivants :

- si l'auteur du texte sur les électeurs en survêtements a publié le texte controversé immédiatement après les élections législatives, il semble fort possible que le même auteur ait publié d'autres messages sur le site du parti SDS sous son vrai/autre nom.

Pour cette raison, nous avons décidé d'intégrer dans une analyse quantitative les messages publiés sur le site du SDS trois mois avant et trois mois après la publication du texte controversé.

Au cas où aucun des auteurs examinés ne montrerait suffisamment de similarités avec l'auteur Tomaž Majer, les résultats de l'étude apporteraient de nouvelles connaissances sur les indices linguistiques à partir desquels il est possible d'attribuer un texte en slovène à un auteur particulier.

5 MÉTHODOLOGIE

5.1 Compilation du corpus

La méthodologie de l'attribution de l'auteur demande une préparation minutieuse des textes dans le cadre de laquelle il faut :

- anonymiser les auteurs et enregistrer chaque texte dans un fichier séparé,
- nettoyer les traces de mise en forme et convertir les fichiers en format .txt,
- former les en-têtes de documents composés du code de l'auteur anonymisé et du titre du texte,[2]
- réaliser l'annotation morphologique automatique des textes.[3]

L'image 1 montre un extrait du texte préparé selon les étapes décrites. Les annotations du champ noir représentent les annotations morphologiques pour tous les lemmes[4] du corpus.

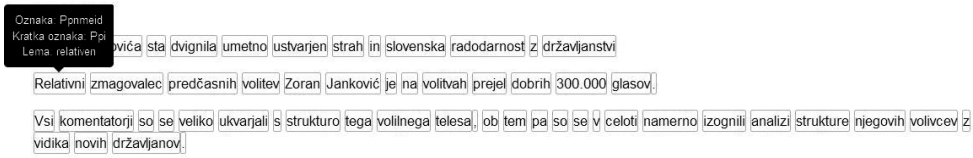


Image 1 : Exemple de texte préparé pour l'analyse[5]

Une fois le corpus préparé, nous pouvons procéder à l'analyse statistique à partir des indices linguistiques qui permettent d'attribuer un texte d'origine inconnue à l'un des auteurs potentiels.

---

2 Exemple d'un en-tête de document: « J_2 : La démocratie enlevée »

3 L'annotation morphologique des textes en slovène est accessible sur le site suivant : http://oznacevalnik.slovenscina.eu/Vsebine/Sl/SpletniServis/SpletniServis.aspx

4 Nous prenons pour lemmes tous les mots rattachés à leurs entrées de dictionnaire.

5 Traduction en français : *La victoire de Janković est due à une peur artificiellement créée et aux naturalisations trop généreusement octroyées. Le vainqueur relatif des élections anticipées, Zoran Janković, a récolté 300.000 voix. Alors que les commentateurs s'occupaient de la structure globale du corps électoral, ils n'ont rien dit sur l'origine des électeurs de Janković.*

5.2 Analyse

Pour pouvoir comparer le texte d'origine inconnue aux textes des 21 auteurs potentiels, nous avons procédé à une technique d'apprentissage automatique basée sur les machines à vecteurs de support qui aident à résoudre les problèmes de discrimination entre les éléments analysés.

Dans une première étape, il faut choisir les indices linguistiques[6] selon lesquels nous voulons classifier les textes et ensuite observer les différences entre ces indices auprès des différents auteurs. Pour notre analyse, les principaux paramètres pour le calcul concernent le lexique et la lisibilité, chacun comportant ses propres formules pour atteindre le résultat optimal.

Les indices lexicaux :

a) la diversité du vocabulaire,
b) la formule Brunet (Brunet 1988), qui quantifie la richesse du vocabulaire indépendamment de la longueur du texte,
c) la proportion des hapax dans le texte (Holmes 1992), qui donne les résultats concernant les lemmes apparaissant une seule fois dans le texte,
d) la formule Honoré (Honoré 1979), basée sur l'hypothèse que le nombre de hapax est proportionnel à la richesse du vocabulaire d'un auteur.

Les indices de lisibilité :

a) la proportion entre le nombre de mots et le nombre de phrases,
b) la proportion entre le nombre de signes et le nombre de mots,
c) la formule ARI (Automated Readability Index), qui mesure le niveau de formation nécessaire pour pouvoir facilement comprendre un texte après une première lecture,
d) la formule Gunning Fog (Gunning 1952), qui propose une autre formule pour calculer le nombre d'années de formation pour pouvoir comprendre un texte après une première lecture.

À partir de la distance et de la parenté entre les textes fournies par les indices linguistiques présentés ci-dessus, il est possible de formuler des hypothèses sur l'auteur du texte d'origine inconnue.

6 RÉSULTATS

Nous avons d'abord effectué le calcul des indices du lexique et de lisibilité pour chacun des textes analysés. Les tableaux 1 et 2 présentent les résultats pour le texte de Majer.

6 Les paramètres que nous avons pris en compte pour le calcul ont été élaborés pour l'anglais et le français, mais ont également été testés sur d'autres langues. Afin de vérifier le fonctionnement de ces formules pour l'attribution de l'auteur en slovène, nous allons calculer, dans une étape suivante, la puissance prédictive pour chacun des indices proposés.

Tableau 1 : Les indices lexicaux dans le texte de Majer

Indice	Valeur
Variabilité du vocabulaire	0,38
Index Brunet	12,96
Statistique Honoré	1998,79
Hapax legomena	0,24

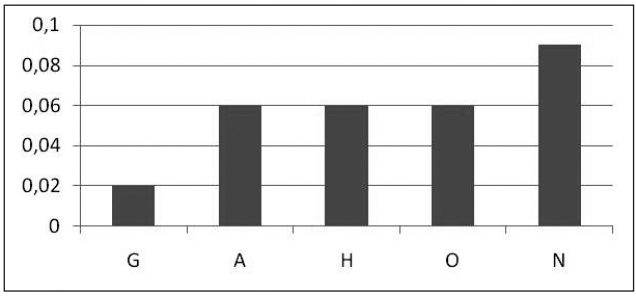
Tableau 2 : Les indices de lisibilité dans le texte de Majer

Indice	Valeur
Proportion mots/phrases	21,24
Proportion signes/mots	5,14
ARI	13,38
Gunning Fog	21,81

Si on compare l'extrait du texte présenté par l'image 1 aux résultats des tableaux 1 et 2, nous pouvons conclure que les résultats des calculs ne peuvent être dégagés du texte par les procédés traditionnels de l'analyse linguistique. De plus, les valeurs absolues des indices de lexique et de lisibilité ne peuvent fournir de résultats sur les particularités stylistiques d'un texte qu'en étant comparées à d'autres textes du corpus.

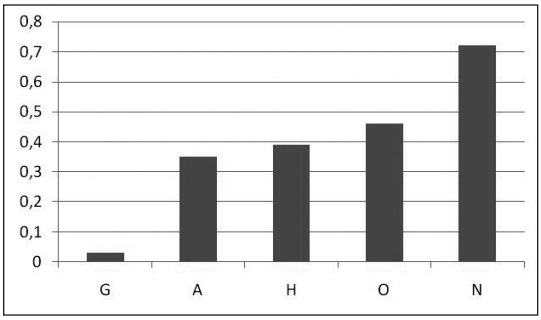
Pour cette raison, nous proposons dans la suite une comparaison entre le texte de Majer et les 75 textes des auteurs connus. Les résultats présentés par les graphiques 1 à 8 concernent les cinq auteurs dont les particularités stylistiques rejoignent le plus celles de l'auteur inconnu.

a) Indices lexicaux



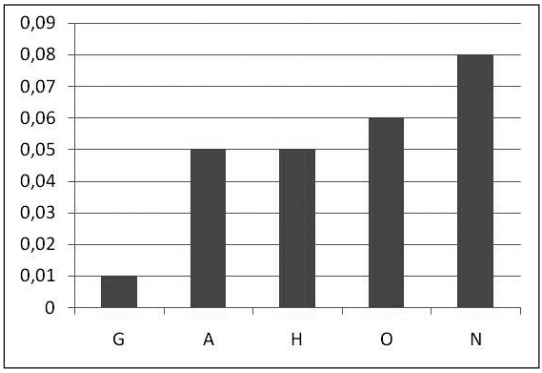
Graphique 1 : Variabilité du vocabulaire

Le calcul du graphique 1 est fait à partir du nombre de lemmes différents sur la totalité du texte. Les résultats montrent les cinq auteurs qui se rapprochent le plus de l'auteur inconnu par la variabilité du vocabulaire.



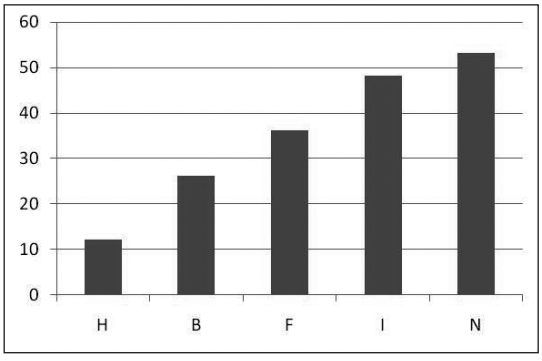
Graphique 2 : Formule Brunet

Le graphique 2 représente l'échelle de distance selon la formule Brunet qui s'appuie sur la richesse du vocabulaire relativisée par la longueur du texte.



Graphique 3 : Hapax legomena

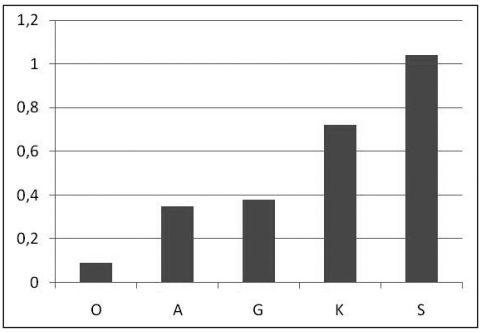
Les résultats proposés par le graphique 3 concernent la proportion des hapax dans le texte (les lemmes qui apparaissent une seule fois dans le texte).



Graphique 4 : Statistique Honoré

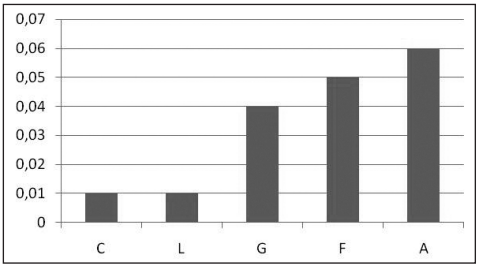
Le graphique 4 visualise les auteurs selon la formule Honoré (Honoré 1979), calculée à partir de la proportion des hapax et la richesse du vocabulaire d'un auteur.

b) Indices de lisibilité textuelle



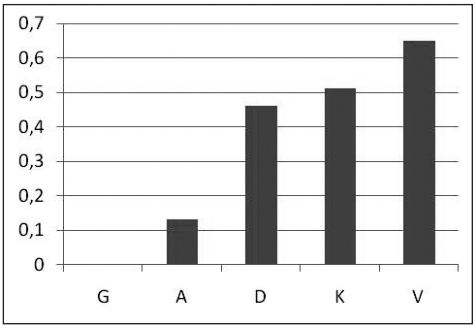
Graphique 5 : Proportion entre le nombre de mots et le nombre de phrases

La proportion entre le nombre de mots et le nombre de phrases est connue comme étant un indice fiable pour évaluer la lisibilité d'un texte.



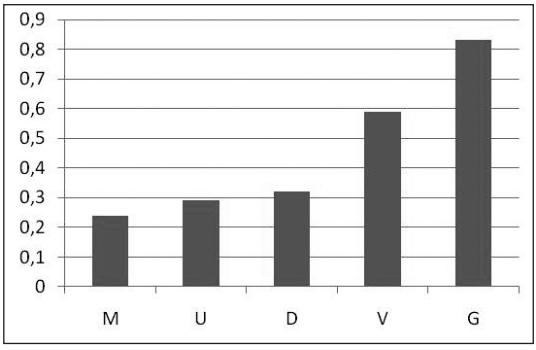
Graphique 6 : Proportion signes/mots

Le graphique 6 met en valeur la proportion entre le nombre de signes et le nombre de mots des textes analysés.



Graphique 7 : Formule ARI

La formule ARI (Automated Readability Index) mesure le niveau de formation nécessaire pour pouvoir facilement comprendre un texte lors d'une première lecture. L'ordre des auteurs qui rejoignent les indices linguistiques de Majer ressemble à celui fourni par les indices lexicaux.



Graphique 8 : Formule Gunning Fog

Le calcul présenté par la formule Gunning Fog distingue les auteurs selon une autre formule pour calculer le nombre d'années de formation nécessaire pour pouvoir comprendre un texte après une première lecture.

6. 1 Synthèse

La synthèse de la classification de textes à l'aide de différents indices linguistiques mène à l'établissement de l'échelon des auteurs dont les traits stylistiques rejoignent le texte de Tomaž Majer par la majorité de critères:

G	54
A	42
H	28
O	22
D	12

Tableau 3 : Les cinq auteurs les plus proches du texte de Majer

En ce qui concerne la puissance prédictive des indices à partir desquels il est possible d'attribuer un texte en slovène à un auteur, les résultats les plus fiables sont fournis par la variabilité du vocabulaire, la formule Brunet et la fréquence relative des hapax. Vu la limitation du corpus analysé, une étude sur un corpus élargi aurait sans doute donné des résultats plus fiables.

7 CONCLUSION

Dans la présente communication, nous avons essayé d'identifier l'auteur du texte intitulé *Les électeurs en survêtements* qui a suscité de vives réactions lors de sa publication. Pour l'analyse, nous avons eu recours à la méthodologie de machines à vecteurs de support dans le cadre de laquelle nous avons observé les indices lexicaux et les indices de lisibilité. Les résultats de l'analyse montrent que :

- la puissance prédictive pour le slovène repose sur la variabilité du vocabulaire, la formule Brunet et la fréquence relative des hapax,
- parmi les 21 auteurs potentiels, il semble fort probable que l'auteur G a écrit le texte controversé.

Comme il s'agit d'une étude de cas authentique, 75 textes produits par 21 auteurs représentent un corpus plutôt limité. Cette étude aurait beaucoup gagné si on avait pu préalablement réaliser une analyse plus élaborée des indices linguistiques fiables pour la langue slovène tels que ceux que l'on connaît pour le français ou l'anglais.[7]

Cette analyse témoigne de l'immense complexité d'un texte écrit ou oral : non seulement l'auteur ne peut pas entièrement contrôler sa production linguistique, de même, au niveau de l'analyse, il est impossible d'appréhender toutes les structures qu'il a utilisées par une seule approche linguistique ou statistique. C'est pourquoi il est grand temps que la linguistique de corpus et l'apprentissage automatique conjuguent leurs efforts au profit d'une approche pluridisciplinaire à fort potentiel.

Sources primaires

Texte sur *Les électeurs en survêtements* :
http://www.delo.si/assets/media/other/20111211//Prispevek%20Toma%C5%BEa%20Majerja.pdf

Archives du site SDS :
http://www.sds.si/arhiv?id=12

Outil d'annotation morphologique automatique pour le slovène :
http://oznacevalnik.slovenscina.eu/Vsebine/Sl/SpletniServis/SpletniServis.aspx

Bibliographie

ARGAMON, Shlomo/Shlomo LEVITAN (2005) « Measuring the usefulness of function words for authorship attribution. » In: A. Bia (éd), P*roceedings of ACH/ALLC 2005*. Victoria BC : Association for Computing and the Humanities, 1–3.

7 Un projet de recherche est actuellement consacré à la description des indices linguistiques à partir desquels il est possible d'attribuer un texte slovène à un auteur potentiel : pour plus d'informations, lire la présentation sur le site http://www.trojina.si/vsebine/14/Current%20projects?b=11

BRUNET, Étienne (1988) « Une mesure de la distance intertextuelle : la connexion lexicale. » *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines* 24, 81–116.
BUCKS, Romola/Sameer SINGH/Joanne M. CUERDEN/Gordon K. WILCOCK (2000) « Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. » *Aphasiology* 4/1, 71–91.
DOVIĆ, Marijan (2002) « Podbevšek in Cvelbar: Poskus empirične preverbe namigov o plagiatorstvu. » *Slavistična revija* 50, 233–249.
HONORÉ, Antony (1979) « Some Simple Measures of Richness of Vocabulary. » *Association of Literary and Linguistic Computing Bulletin* 7, 172–177.
LABBÉ, Cyril/Dominique LABBÉ (2003) « La distance intertextuelle. » *Corpus* 2, 1–16.
LIMBEK, Marko (2008) « Usage of Multivariate Analysis in Authorship Attribution: Did Janez Mencinger Write the Story »Poštena Bohinčeka«? » *Metodološki zvezki* 5/1, 81–93.
LUYCKX, Kim/Walter, DAELEMANS (2005) « Shallow text analysis and machine learning for authorship attribution. » In: T. Van der Wouden (éd), *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands*. Leiden : Leiden University, 149–160.
MENDENHALL, Thomas Corwin (1887) « The characteristic curves of composition. » *Science* IX, 237–249.
STAMATATOS, Efstathios (2009) « A Survey of Modern Authorship Attribution Methods. » *Journal of the American Society for Information Science and Technology* 60/3, 538–556.
ZWITTER VITEZ, Ana (2012) « Authorship Attribution: Specifics for Slovene. » *Slavia Centralis* 5/1, 75–85.

Résumé

LE DÉCRYPTAGE DE L'AUTEUR ANONYME :
L'AFFAIRE DES ÉLECTEURS EN SURVÊTEMENTS

Durant les dernières décennies, les recherches sur l'attribution de l'auteur ont fait un progrès remarquable en raison d'une demande croissante dans les domaines du droit (plagiats), de la criminologie (lettres de menace), des études littéraires (pseudonymes) et des recherches mercatiques (stratégies publicitaires).

L'article propose l'analyse d'un texte intitulé *Les électeurs en survêtements* et sa comparaison avec 75 textes produits par 21 auteurs connus. La méthodologie se fonde sur les machines à vecteurs de support qui permettent de définir la distance ou la proximité entre différents textes selon les indices de lexique et de lisibilité textuelle. Les résultats démontrent que les particularités de l'un des auteurs rejoignent celles du texte controversé selon la variabilité du vocabulaire, la formule Brunet et la fréquence relative des hapax.

Les résultats de l'analyse soulignent l'importance du lien entre la linguistique de corpus et l'apprentissage automatique, ce qui permettra aux deux disciplines de profiter des grandes potentialités de cette approche pluridisciplinaire.

Mots-clés : attribution de l'auteur, linguistique de corpus, électeurs en survêtements

Povzetek

NA SLEDI ANONIMNEMU AVTORJU: AFERA VOLIVCEV V TRENIRKAH

V zadnjih desetletjih je ugotavljanje avtorstva besedil doživelo velik razmah, saj prinaša izrazito aplikativne rezultate na področju prava (plagiatorstvo), kriminologije (grozilna pisma), literarnih študij (psevdonimi) in tržnih analiz (strategije oglaševanja).

V prispevku analiziramo besedilo, imenovano »Volivci v trenirkah«, in ga primerjamo s 75 besedili 21 znanih avtorjev. Analiza temelji na metodi podpornih vektorjev (SVM), ki omogočajo določanje razlik in podobnosti med primerjanimi besedili na podlagi značilk besedišča in berljivosti. Rezultati kažejo, da so specifike enega izmed opazovanih avtorjev precej podobne besedilu neznanega izvora glede na raznolikost besedišča, Brunetovo formulo in relativno frekvenco hapaksov v besedilu.

V sklepu poudarimo pomembnost povezovanja korpusnega jezikoslovja in strojnega učenja, s katerim lahko obe področji doživita nov razmah in izkoristita izjemno moč tega interdisciplinarnega pristopa.

Ključne besede: ugotavljanje avtorstva besedil, korpusno jezikoslovje, volivci v trenirkah