Harry Potter and the Rosetta Stone:
uncovering differences in tense use through parallel corpora

Martijn van der Klis, Anja Goldschmidt, Jos Tellings, Bert Le Bruyn and Henriëtte de Swart

UiL OTS, Utrecht University

Although the Germanic and Romance language share most of their tense systems, there is still variation in tense use, especially in the Perfect tense. We use a parallel corpus approach to uncover these differences: translations provide us with form variation while meaning stays stable. We present a novel methodology to annotate and visualize tense variation. We apply this methodology to translations of J.K. Rowlings' *Harry Potter and the Philosopher's Stone* in six languages from two language families (Germanic (English, German and Dutch) and Romance (Spanish, French and Italian)). While we find almost no variation in tense use in the narrative parts of this novel, we do find variation in the dialogues. We there find a core Perfect shared by all languages under investigation, but also competition between Present, Perfect and Past.

Methodology

Our data stems from J.K. Rowlings' *Harry Potter and the Philosopher's Stone*, chapters 1 and 17. The novel and its translations were OCR'ed, manually checked for spelling errors and then automatically preprocessed and aligned (software used: *Uplug*, Tiedemann (2003)). We created a script to mark dialogue and narrative parts based on punctuation. Annotators then selected all finite, indicative verb phrases from the corpus and assigned a tense label to these verb phrases in a web interface (dubbed *PreSelect*). Annotators then marked the translations for each of these verb phrases in German, Spanish, French, Italian and Dutch and assigned tense labels (dubbed *TimeAlign*, van der Klis et al. (2017)). As an example:

(1) a. 'But Snape **tried** to kill me!' [en, Past]
b. 'Aber Snape **hat versucht** mich umzubringen!' [de, Perfect]
c. '¡Pero Snape **trató** de matarme!' [es, Past]
d. 'Mais Rogue **a essayé** de me tuer!' [fr, Perfect]
e. 'Ma Piton **ha tentato** di uccidermi!' [it, Perfect]
f. 'Maar Sneep **heeft geprobeerd** me te vermoorden!' [nl, Perfect]

Descriptive statistics

We see almost no variation in the narrative parts: the main difference is that the Romance languages share an Imperfect that is used instead of the Past in some cases.

	en	de	es	fr	it	nl
Present	6	6	4	3	6	5
Perfect	0	0	0	0	8	0
Past	430	461	336	313	316	465
Imperfect	0	0	126	130	112	0
Past Perfect	29	27	18	30	24	32

Tab. 1: Most frequently attributed tense categories in the **narrative** parts of the novel.

In the dialogues, the Present and Future domains are largely similar across languages. The main competition is between the Past, Perfect and Imperfect, with all languages (except for Spanish) using far more Perfects than English. We see that the French and Italian Past is almost non-existent.

	en	de	es	fr	it	nl
Present	126	154	122	129	129	162
Perfect	27	78	23	99	100	60
Past	120	64	88	0	2	84
Imperfect	0	0	38	50	39	0
Future	20	15	13	11	23	9

Tab. 2: Most frequently attributed tense categories in the **dialogue** parts of the novel.

Visualization using multidimensional scaling

We can further analyse the competition in the Perfect domain if we focus on translations of the English *simple past* cases. We use multidimensional scaling (MDS) to uncover patterns in translation (after Wälchi & Cysouw (2012)). MDS allows us to visualize the level of similarity of contexts in the dataset. We define a context as a tuple of assigned tense labels (e.g. (1) := (Past, Perfect, Past, Perfect, Perfect, Perfect)). Two contexts are considered most similar if the tense label for each language matches, and most dissimilar if none of the tense labels match. The MDS algorithm then plots each context in a n-dimensional space based on similarity (such that similar contexts cluster together).

Per language, we can then colour each context using the language's tense labelling, leading to interpretable (semantic) maps. We illustrate this colouring in Figures 1 and 2. In Figure 1, the colouring represents the assigned tenses for French, in Figure 2, for Dutch. The figures show that in a largely overlapping area, French and Dutch use Perfect rather than the (English) Past. The French Perfect area however forms a superset and covers part of the Dutch Past domain. French also has an Imperfect, which covers largely the same area in the other Romance languages (not shown here). Clicking on clusters of points yields the raw data and allows us to further investigate the differences between languages.
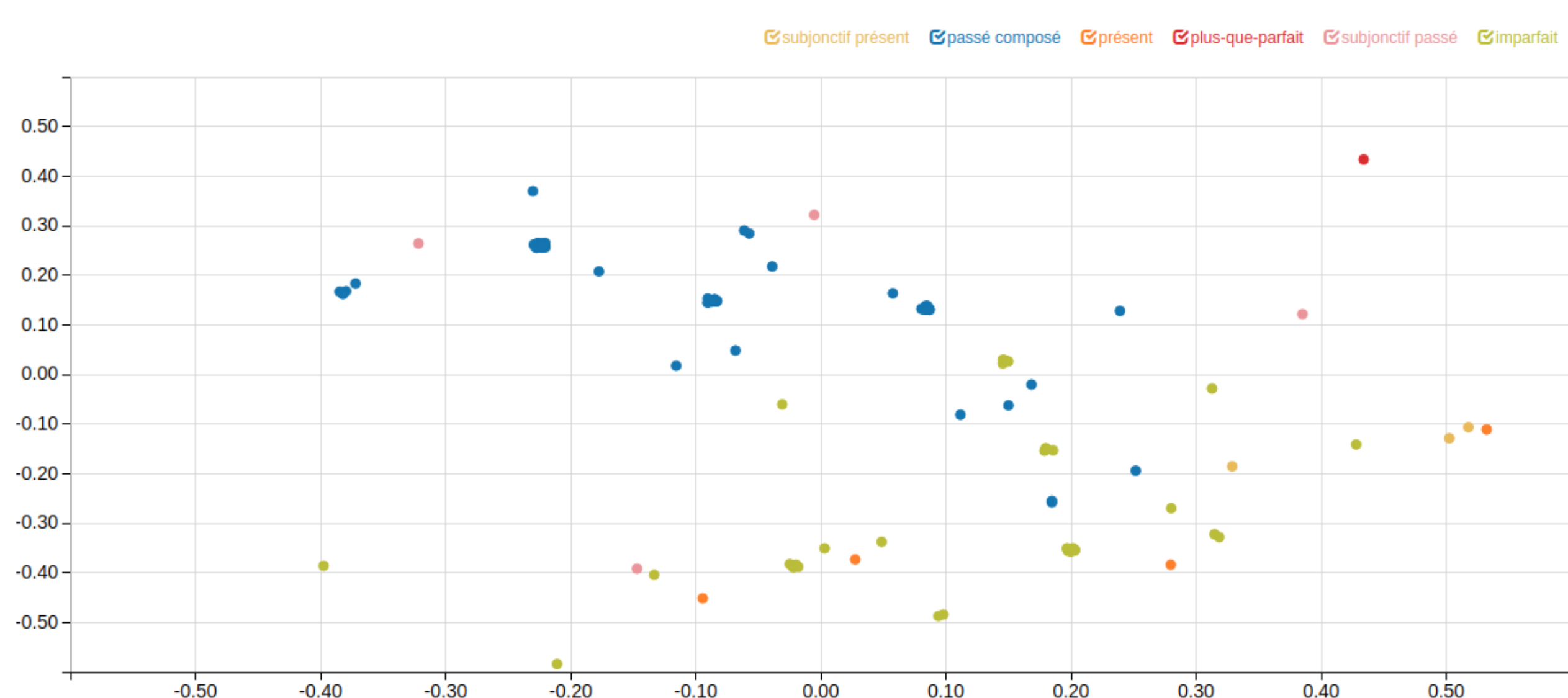

Fig. 1: Result from applying MDS, with French tense labelling.
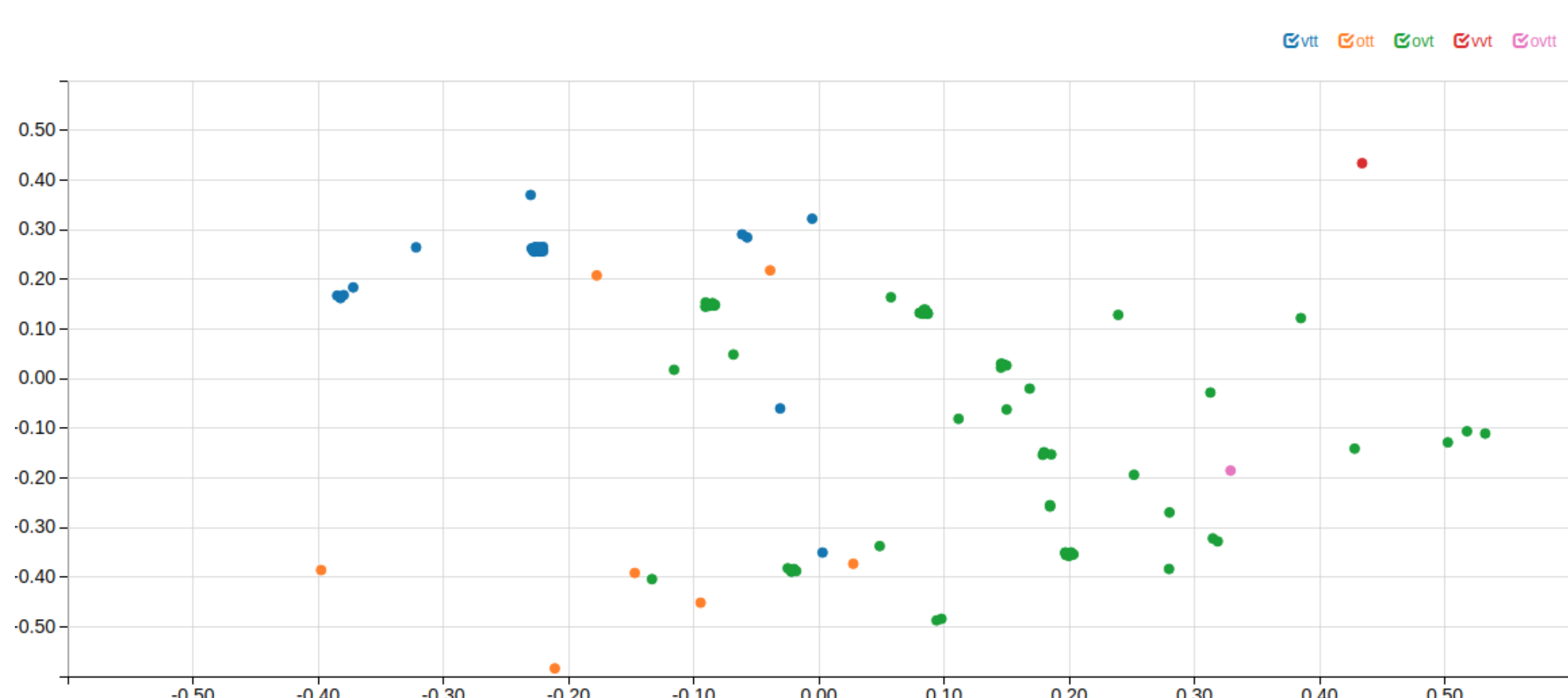

Fig. 2: Result from applying MDS, same as in Fig. 1, but now with Dutch tense labelling.

Analysis: core Perfect

We find that there is a 'core' Perfect meaning shared by the languages under investigation. We classify these cases as follows:

⚡ States that have lasted for a while, but are now over.

(2) a. 'You'**ve had** nearly fifteen minutes, now OUT', she said.
b. 'We've **had** precious little to celebrate for eleven years.'

⚡ Events that yield a result state with current relevance.

(3) a. 'See what I **have become**?', the face said.
b. 'I - I'**ve won** the House Cup for Gryffindor.'

⚡ Negative existential phrases: absence of an event that would have had current relevance.

(4) a. 'Er – Petunia, dear – you **have**n't **heard** from your sister lately, have you?'
b. 'I **have**n't **blushed** so much since Madam Pomfrey told me she liked my new earmuffs.'

Analysis: Perfect vs. Past

From the MDS analyses, we see that a Perfect is introduced frequently in English *simple past* cases. We categorize these as follows:

⚡ Narrative sequences in French, Italian and German. Note that in (5c), Dutch does allow a Perfect at the start of the narrative sequence.

(5) a. 'I killed your father first and he **put** up a courageous fight...' [en, Past]
b. 'J'ai d'abord tué ton père et il m'**a résisté** avec une grande bravoure...' [fr, Perf.]
c. 'Ik heb eerst je vader gedood en die **verweerde** zich dapper...' [nl, Past]

⚡ Perfective past events in French and Italian. German also allows a Perfect in these cases, though not with modal/stative verbs, as in (6c) below.

(6) a. 'I think he sort of **wanted** to give me a chance' [en, Past]
b. 'Penso che **abbia voluto** darmi una possibilità.' [it, Perfect]
c. 'Ich glaube, er **wollte** mir eine Chance geben.' [de, Past]

⚡ Completed events that yield a resulting state with no direct current relevance in French, Italian, German and Dutch. See also (1) for a similar context.

(7) a. 'Dumbledore **gave** me the day off yesterday ter fix it.' [en, Past]
b. 'Dumbledore me **dio** libre el día de ayer para hacerlo.' [es, Past]
c. 'Perkamentus **heb** me gisteren vrijaf **gegeven** om 't te regelen.' [nl, Perfect]

We do not find competition between Perfect and Imperfect: Germanic languages all consistently opt for a Past tense in these cases.

Analysis: Perfect vs. Present

We find that Perfect use in English is more present-oriented. In two categories of cases, other languages require a Present:

⚡ States that continue into the present.

(8) a. 'How long **have** I **been** in here?' [en, Perfect]
b. '¿Cuánto tiempo hace que **estoy** aquí?' [es, Present]
c. 'Wie lange **bin** ich schon hier?' [de, Present]

⚡ Resultatives with *to get*. In some cases, Spanish also allows a Perfect here.

(9) a. 'I'**ve got** yeh a present.' [en, Perfect]
b. 'Te **he traído** un regalo.' [es, Perfect]
c. 'Ich **hab** ein Geschenk für dich.' [de, Present]

Conclusion

We have found there is almost no variation in tense use in narrative parts of the novel. In the dialogues, the main source of variation in tense use is the Perfect. The Perfect has a core meaning shared by all languages, but is also in competition with the Past and the Present. In the languages under consideration, we find four configurations, categorized below:

narrative past	completed ev.	core Perfect	resultative	continuative
French, Italian, German				
	Dutch			
		Spanish		
		English		

The Rosetta Stone proved to be the key to deciphering Egyptian hieroglyphs. We think parallel corpora might be its contemporary in discovering variational patterns between languages. In contrast to popular black-box methods, multidimensional scaling allows for combining distant and close reading, which we think could move both linguistics and the digital humanities as a whole forward.

Mischief Managed