Which Languages are “Hot”, and which are “Cool”? Using Universal Dependencies for Large-scale Comparisons of Subject Expression

Natalia Levshina

Radboud University
Nijmegen, The Netherlands
natalia.levshina@ru.nl

Abstract

This study uses Universal Dependencies to investigate subject omission across fifty-six news corpora and twenty geographic varieties of English. Building on McLuhan’s “hot–cool” distinction, Hall’s LC–HC continuum, and Bisang’s notion of overt vs. hidden complexity, it tests whether subject omission rates reflect degrees of contextual reliance. The results broadly support these theories: low-context, “hot” languages, such as German, Dutch and Swedish, show low omission rates, while high-context, “cool” languages, such as Japanese, Korean and Chinese, show higher rates. English behaves as a “hot” language but exhibits internal variation across varieties, with Southeast Asian varieties exhibiting more omission than African ones. The study provides large-scale quantitative evidence while highlighting the need for further theoretical and methodological refinement, particularly regarding the role of word order and agreement.

Keywords: “hot” and “cool” communication, high- and low-context cultures, subject, varieties of English

1. “Hot” and “Cool” Languages

The well-known distinction between "hot" and "cool" media, introduced by Marshall McLuhan (1964), captures differences in the degree of sensory detail provided by a medium and the level of audience participation required to complete the message. For example, films can be considered “hot” in comparison to books, as the former supply more sensory input and place fewer interpretative demands on the recipient.

A related and influential distinction is that by between high-context (HC) and low-context (LC) communication, proposed by E. T. Hall (1976). In HC communication, much of the information is inferred from context rather than stated explicitly, whereas LC communication relies on the explicit encoding of information with less shared background assumed. Although different versions of the scale exist (see discussion of this and other problems in Cardon 2008; Kittler et al., 2011), countries such as Germany and the USA are typically classified as LC cultures, while China and Japan are often situated close to the HC end of the spectrum (see Figure 1).

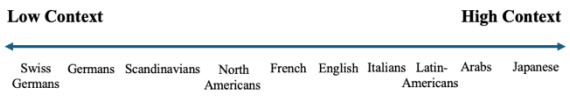

Figure 1: A continuum of high- and low-context cultures in Rösch & Segler (1987), based on Hall (1976).

Similar conceptual distinctions have also been proposed in linguistic typology. For examp[le, Bisang (2009) differentiates between overt and hidden complexity, contrasting obligatory explicit encoding of grammatical meanings with their inference from context and pragmatic principles. Likewise, Tsao (1977) distinguishes between “sentence-oriented” languages (e.g., English) and “discourse-oriented” ones (e.g., Chinese). Similar to McLuhan’s distinction, Ross (1982) suggested that languages can also be placed along a "hot–cool" continuum, particularly with regard to how anaphoric referents are expressed. For example, English, in which the speaker provides information about the referents with the help of anaphoric pronouns, as in (1), would be close to the "hot" pole of the continuum.

(1) ***She*** *is writing an article.*

In contrast, Japanese would be close to the "cool" pole because the recovery of the anaphoric referents in zero anaphora contexts requires more effort on the addressee’s part. Consider (2):

(2) *Kinoo* *ryoori* *o* *shita.*
Yesterday cooking OBJ do.PST
‘Yesterday, (someone) cooked’.

Other languages can be located in-between. For instance, in Russian, pronominal subjects can be either overt or implicit:

(3) *(Ona)* *pišet* *statju.*
3SG.NOM.F write.IPFV.PRES.3SG article.SG.ACC
‘(She) is writing an article.’

The use and omission of referents are guided by pragmatic principles (Horn, 1984; Huang, 1984), whereby more accessible (i.e., recoverable, non-contrastive, topic-continuous) referents tend to be omitted more often than less accessible ones

(Levshina, 2022). Sometimes the information about the referent can be restored from the verb agreement, as in (3), but this is not always the case. For example, Japanese verbs, as in (2), have no person, number or gender agreement markers, but the use of personal pronouns is very restricted (McGloin et al., 2014: Ch. 8).

This study exploits the Universal Dependencies framework for large-scale cross-linguistic and cross-varietal comparison. The first case study examines subject expression across 56 news corpora in different languages. The second focuses on blogs in twenty varieties of English, a language typically characterized as "hot." The study addresses two main questions:

1) Do cross-linguistic patterns of subject omission support existing theoretical accounts?

2) Are all English varieties equally "hot," and if not, do differences correlate with sociolinguistic or cultural factors?

2. Subjects in 56 News Corpora

The data for this case study are drawn from 56 news corpora in the Leipzig Corpus Collection (Goldhahn et al., 2012). The corpora consist of sentences from diverse online news sources. Where possible, I selected corpora representing specific varieties:

- German: a general ("unspecified") corpus (deu) and data from Switzerland (deu-ch).
- Portuguese: a general corpus (por), alongside data from Portugal (por-pt), Brazil (por-br) and Macao (por-mo).
- Arabic: a general corpus (ara) and data from Saudi Arabia (ara-sa).

The sentences were annotated with UD information using the default language models in Stanza (Qi et al., 2020). Using information about the dependencies and morphological features, I extracted potentially finite verbal predicates (UPOS = VERB or AUX) with the following dependency labels: *acl*, *acl:relcl*, *advcl*, *ccomp*, *csubj*, *parataxis* and *root*. Imperative forms were excluded because they normally do not allow for overt subjects. Every predicate was checked for having an overt subject as its dependent (*nsubj*, *nsubj:pass* or *csubj*).

The proportions of predicates without overt subjects are displayed in Figure 2. The distribution reflects the general outline of the LC–HC continuum (cf. Rösch and Segler, 1987; Mordon, 1999; Meyer, 2014) and the "hot–cool" distinction. Typical low-context (LC) varieties of German cluster at the lower end of the scale, alongside Afrikaans, Dutch, and the North Germanic languages, all showing low rates of subject omission (with Icelandic closer to the high-context pole). At the opposite end, prototypical high-context (HC) languages, such as Korean, Japanese, and Chinese, occupy the top positions.

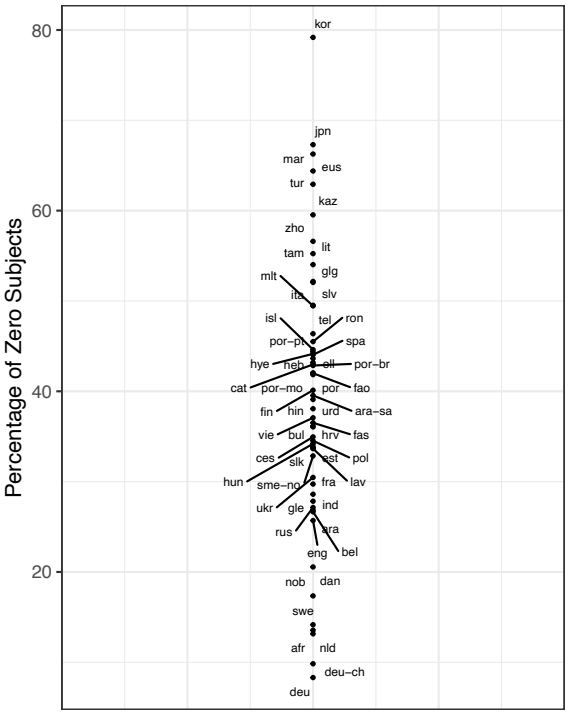

Figure 2: Percentages of predicates without overt subject in 56 online news corpora.

Typical LC German varieties are located at the bottom followed by Afrikaans, Dutch and some North Germanic languages with low subject omission rates (although Icelandic and Faroese are closer to the HC pole). In contrast, prototypical HC languages, such as Korean, Japanese and Chinese are located at the top.

Romance languages display variation: French patterns more closely with English toward the LC end, while the other Romance languages lie nearer the HC pole. Among Slavic languages, East Slavic varieties (Belarusian, Russian, Ukrainian) are relatively LC-oriented, Slovene is closest to the HC end, and Bulgarian, Croatian, Czech, Polish, and Slovak fall in between.

Regional varieties of German and Portuguese show little internal divergence, whereas Arabic exhibits greater variation, with Saudi Arabic positioning closer to the HC end than the general Arabic corpus.

Although the results broadly align with the LC–HC continuum, other factors are likely involved. In particular, verb-final languages tend to realize fewer arguments than VO languages, often by omitting the subject (Levshina, 2025). This behaviour may reflect pressures to manage processing costs and avoid long dependencies. The study also does not address subject indexing on the verb (e.g., subject clitics), which may be in complementary distribution with full forms. Because agreement systems and their representation in UD vary across languages, this issue is left for future research.

3. Subjects in Geographic Varieties of English

The second case study compares the rates of explicit and implicit subjects in varieties of English spoken in twenty countries: Australia, Bangladesh, Canada, Ghana, Great Britain, Hong Kong, India, Ireland, Jamaica, Kenya, Malaysia, New Zealand, Nigeria, Pakistan, the Philippines, Singapore, South Africa, Sri Lanka, Tanzania and the USA. The texts are blogs from the Global Corpus of Web-based English (Davies, 2013).

There is substantial research in variational linguistics on the use and omission of pronominal subjects in English. For example, the electronic World Atlas of Varieties of English, or eWAVE (Kortmann et al., 2020), describes the omission of referential and dummy pronouns as subjects (Features 43 and 44, respectively). The data relevant for the countries in this case study is shown in Table 1. According to eWAVE, traditional L1 varieties (e.g., most British dialects), normally have overt subjects, although exceptions are found in Newfoundland and Scottish Englishes, in which referential subject pro-drop is possible, being neither pervasive nor extremely rare. Subject omission is common in indigenized L2 varieties that originated in Asia (e.g., Indian, Pakistani, Sri Lankan and Malaysian Englishes), but not in those with African origin (e.g., spoken in Nigeria and Jamaica). A potential explanation may have to do with language contact: West Africa is an area in which languages requiring overt subject pronouns are particularly common (Dryer, 2013).

Omission of subject pronouns	**Referential pronouns**	**Dummy pronouns**
Absent	Australia, US (colloquial), Ghana, Jamaica, Nigeria	Ghana, Hong Kong, Jamaica, New Zealand, Nigeria
Extremely rare	Philippines	Ireland, Pakistan, Philippines
Neither pervasive nor extremely rare	Ireland, Pakistan, Sri Lanka, (white) South Africa	Sri Lanka
Pervasive or obligatory	Singapore (colloquial), Hong Kong, India, Malaysia	Singapore (colloquial), India, Malaysia

Table 1: Omission of subject pronouns in English varieties according to eWAVE.

There is no consensus about the position of the twenty English-speaking countries on the LC–HC continuum. Also, not all the countries have been investigated. As an illustration, Figure 3 displays the classifications from Meyer (2014) and Morden (1999). Note that the positions are only approximate, representing rankings, rather than absolute values. According to Morden (1999), New Zealanders and (white) South Africans represent the most LC cultures in this list, whereas Meyer (2014) regards the USA as the most LC culture. At the same time, there is some convergence. For example, African cultures are the most HC in both classifications, followed by the Indian subcontinent and then by Southeast Asia. In many classifications, the British are considered more HC than Americans, as in Rösch and Segler (1987) based on Hall (1976).

Although English is a "hot" language that is not regarded as favouring pro-drop, it is interesting to test if we can detect cross-lectal differences that can be correlated with the LC-HC continua. The data from the variationist literature partly contradicts data from cross-cultural comparisons. Most notably, the African cultures are considered HC, but the corresponding English varieties are reported to use more explicit subjects.

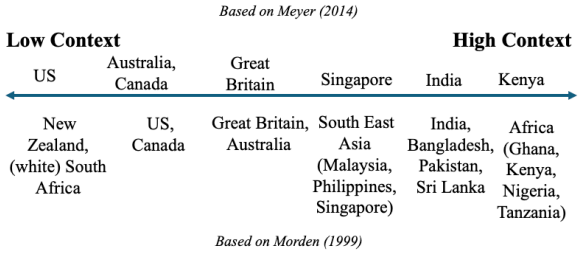

Figure 3: Ranking of cultures according to Meyer (2014) (top) and Morden (1999) (bottom).

To investigate subject omission rates, I sampled randomly 2,000 sentences for every variety from the Global Corpus of Web-based English (blogs) and annotated them with UD using the default Stanza neural model for English, performing the same analyses as the ones described in Section 2. In addition to that, all examples without subject were manually checked to avoid errors due to poor editing or sentence segmentation. Fixed expressions such as *thank you* and *see you* and boilerplate expressions (*The entry was posted on…)* were excluded.

After the semi-automatic cleaning procedure, the dataset contained 45,610 instances of finite predicates. Only 329 of them, or 0.7%, had no overt subject. Figure 4 displays the percentages in each variety. All of them are very low, which means that the web-based varieties of English are "hot".

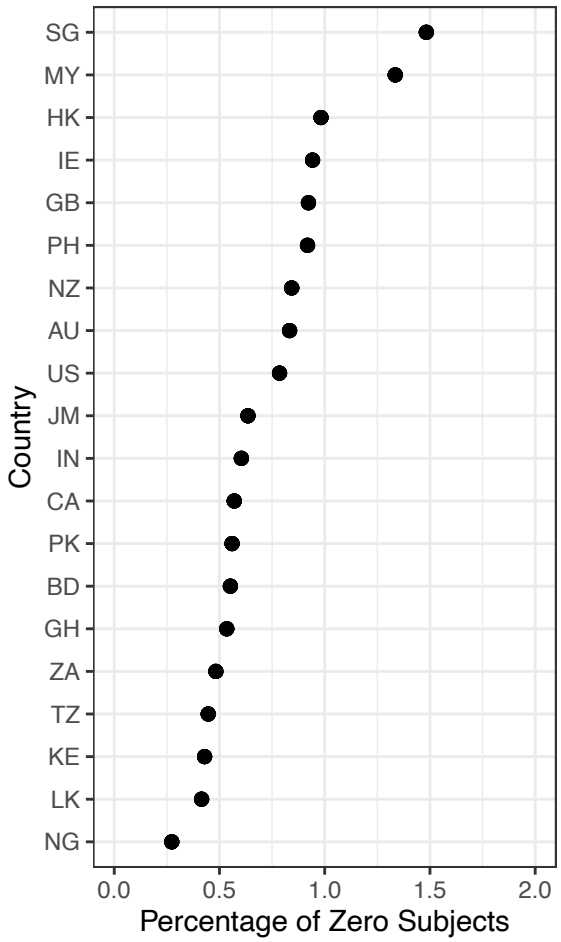

Figure 4: Percentage of zero subjects of finite non-imperative predicates in GloWbE blogs.

At the same time, we observe some variability, which to some extent matches the predictions based on eWAVE: Singapore and Malaysia have the highest proportion of zero subjects, while Nigeria has the lowest, as well as the other African countries in the sample. Although there are some unexpected findings, e.g., relatively low omission rates for South Asian varieties, a statistical analysis of standardized residuals (Agresti, 2002) reveals that only Singapore, Malaysia and Nigeria have observed frequencies that significantly (|res| > 2) deviate from expected values based on marginal frequencies.

A few examples of predicates without overt subjects are provided below, with the predicate tokens in bold. Very often, the first person subject (*I*) is omitted:

(4) *me was a bright young student, full of hope and confidence. hoping to step up into the world and prove my worth.* ***studied*** *my arse off, wanting to prove the world that me is good, me can do it.* ***told*** *meself that i have time to relax when me is done studying.* ***wasted*** *all my youth, the best of my years into mugging those damn tomes. they might as well be tombstones for all they are worth. yeah* ***got*** *the grade,* ***got*** *a place where i hoped to get.* (GloWbE Malaysia)

(5) *Just* ***got*** *back from SMTown Live In Singapore concert.* (GloWbE Singapore)

(6) *Kinda* ***love*** *the 100% sarcastic tone of RPS, tho.* (GloWbE US)

(7) ***Agree*** *that Australia's university system has powerful linkages with Asian regions.* (GloWbE Great Britain)

The subject is also omitted in questions addressed to the 2nd person:

(8) ***Remember*** *mom brownies?* (GloWbE Canada)

Highly accessible abstract 3rd person referents (some event or situations) can be omitted, too:

(9) ***Happens*** *when you have 6 pups together in a pen with pee and poop all over the place.* (GloWbE Australia)

(10) ***Means*** *the industry is growing* (GloWbE Canada)

Among other predicates used without overt subject are *looks (like)*, *sounds (great)*, *reminds (you of)*…, which express the writer's opinions and impressions. Examples of dummy pronouns without any referent (cf. Feature 44 in eWAVE) occur, too, e.g., *seems*/*turns out (that) P*, as in the example below, but they are relatively infrequent.

(11) *Ahhh ...* ***turns*** *out these people were just as confused as us.* (NZ)

It must be emphasized that the frequencies of finite predicates without subject are extremely low, which suggests that grammatical constraints in this register largely override communicative, cultural or any other factors.

Note that it can be reasonable to exclude lexical or clausal subjects from the analysis, as most previous research focused on the omission of pronouns (pro-drop). When proportions of zero subjects are calculated with nouns and clauses excluded, the cross-country differences remain virtually identical (Pearson's $r = 0.99$).

4. Conclusions and Outlook

This study is a first attempt, to the best of my knowledge, to explore the rates of subject omission in a large number of languages and language varieties. The potential of Universal Dependencies is used here to find quantitative evidence automatically in large corpora.

The results lend support to previous claims about "hot", "medium" and "cool" languages and the LC-HC continuum and expand the typology to many new language and varieties. Although the view of English as a "hot" language with predominantly overt subjects is supported by the data, we also find interesting quantitative variation, with

Singapore and Malaysian English having higher rates of zero subjects than Nigerian English and other African varieties, in line with eWAVE.

Explaining these results is challenging, because it is difficult to disentangle linguistic and cultural-communicative factors. In particular, word order and certain types of verb agreement can play a role, as well as language contact. However, the study provides quantitative descriptive data that can be useful for linguistic and sociolinguistic typology.

The present study has several limitations. First, the web-based data are noisy, and the reliability of automatic UD annotations would benefit from more systematic manual validation. The problem of defining *subject* as a comparative category also deserves more attention (de Marneffe et al., 2014). Future work should extend the analysis to additional text types, particularly spoken interaction, in order to get a fuller range of contextual reliance in communicative practice.

5. Bibliographical References

Agresti, A. (2002). *Categorical Data Analysis.* Wiley, Hoboken, NJ.

Bisang, W. (2009). On the evolution of complexity—Sometimes less is more in East and mainland Southeast Asia. In G. Sampson, D. Gil and P. Trudgill (Eds.), *Language complexity as an evolving variable*. Oxford University Press, Oxford, pp. 34–49.

Cardon, P. (2008). A critique of Hall's contexting model: A metaanalysis of literature on intercultural business and technical communication. *Journal of Business and Technical Communication,* 22(4): 399–428.

Hall, E. T. (1976). *Beyond Culture.* Anchor Press/Doubleday, Garden City, NY.

Horn, L. (1984). A new taxonomy for pragmatic inference: Q-based and R-based implicature. In Schiffrin, D. (Ed.). *Meaning, Form, and Use in Context: Linguistic Applications*. Georgetown University Press, Washington, pp. 11–42.

Huang, C.-T. J. (1984). On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry*, 15(4):531–574.

Kittler, M.G., Rygl, D. and Mackinnon, A. (2011). Beyond culture or beyond control? Reviewing the use of Hall's high/low-context concept. *International Journal of Cross Cultural Management*, 11(1): 63–82. https://doi.org/10.1177/1470595811398797

Levshina, N. (2022). *Communicative Efficiency: Language Structure and Use*. Cambridge University Press, Cambridge.

Levshina, N. (2025). The paradox of SOV: A case for token-based typology. *Glottometrics*, *59*: 1-23. https://doi.org/10.53482/2025_59_425

McGloin, N., Endo Hudson, M., Nazikian, F. and Kakegawa, T. (2014). *Modern Japanese Grammar: A Practical Guide*. Routledge, London.

McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. McGraw-Hill, New York.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C.D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 4585–4592, European Language Resources Association (ELRA), Reykjavik, Iceland.

Meyer, E. (2014). *The Culture Map: Decoding how people think, lead, and get things done across cultures*. PublicMatters, New York.

Morden, T. (1999). Models of National Culture – A Management Review. *Cross Cultural Management,* 6(1):19–44.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations*. 2020. https://arxiv.org/abs/2003.07082

Ross, J.R. (1982). Pronoun-deleting processes in German. Paper presented at the annual meeting of the Linguistic Society of America, San Diego, California.

Rösch M. and Segler, K.G. (1987). Communication with the Japanese. *Management International Review, 27*: 56–67.

Tsao, F. (1977). *A Functional Study of Topic in Chinese: The First Step toward Discourse Analysis* [Doctoral dissertation]. USC.

6. Language Resource References

Davies, M. 2013. Corpus of Global Web-Based English. Available online at https://www.english-corpora.org/glowbe/.

Dryer, M.S. (2013). Expression of Pronominal Subjects. In Dryer, M.S. & Haspelmath, M. (Eds.), WALS Online (v2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591. Available online at http://wals.info/chapter/101

Goldhahn, D., Eckart, Th. and Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12*), pp. 759–765. ELRA. Available online at https://wortschatz.uni-leipzig.de/en/download/.

Kortmann, B., Lunkenheimer, K. and Ehret, K. (Eds.) 2020.The Electronic World Atlas of Varieties of English. Zenodo. DOI: 10.5281/zenodo.3712132. Available online at http://ewave-atlas.org.