A Database for Measuring Linguistic Information Content

[1]Richard Sproat, [2]Bruno Cartoni, [3]HyunJeong Choe, [4]David Huynh,
[1]Linne Ha, [1]Ravindran Rajakumar, [4]Evelyn Wenzel-Grondie

Google, Inc
[1]New York, [2]Zurich, [3]Seoul, [4]Mountain View
{rws,brunocartoni,hyunjeongc,dfhuynh,linne,ravirajakumar,evelynw}@google.com

Abstract

Which languages convey the most information in a given amount of space? This is a question often asked of linguists, especially by engineers who often have some information theoretic measure of "information" in mind, but rarely define exactly how they would measure that information. The question is, in fact remarkably hard to answer, and many linguists consider it unanswerable. But it is a question that seems as if it ought to have an answer.

If one had a database of close translations between a set of typologically diverse languages, with detailed marking of morphosyntactic and morphosemantic features, one could hope to quantify the differences between how these different languages convey information. Since no appropriate database exists we decided to construct one. The purpose of this paper is to present our work on the database, along with some preliminary results. We plan to release the dataset once complete.

Keywords: Information content, morphosyntax, morphosemantics

1. Introduction

In April 2013 one of the authors was asked by a science journalist to comment on the question of which languages can pack the most information into a Tweet (Taylor, 2013). To a first approximation, Twitter defines a character as a Unicode code point, and thus 'A' or the Chinese character 肉 count the same, even though the former is from a set that makes a few tens of distinctions, whereas the latter is from a set that makes a few thousand distinctions. Thus for this very limited notion of "space", the question turns out to be relatively easy to answer: Chinese, with its morphosyllabic writing system, where each character corresponds to a morpheme, is a strong contender since one can pack a lot more message into 140 morphemes than into 140 characters of a language like English.

But that question led naturally to another broader question, one often asked of linguists: Which languages are the most "efficient" at conveying information in a given amount of space, where "space" is not as arbitrarily defined as on Twitter, and where "information" is at least implicitly more broadly intended than just, e.g., how many words one can convey? Often the questioner may have an information-theoretic notion of "information" in mind (Shannon, 1948), though the questioner rarely defines how they would actually compute the information. This question is, in fact, remarkably difficult to answer and as Taylor notes, many linguists consider it unanswerable.

Yet it is a question that seems as if it ought to have an answer – if one could define what one means by information. The problem is that to convey what is theoretically the same message, different languages frequently mark different kinds of information. For example in English, one might say *Grandfather died.* An appropriate translation of this sentence into Korean might be something like

할아버지께서 돌아가셨어요
harabeoji-kkeseo doraga-sy-eoss-eo-yo
(grandfather+ref-resp-nom
pass-away+ref-resp+past+linking-vowel+addressee-resp)

This conveys the same message, but it also conveys other information too. In particular, by using the verbal ending *-yo*, the speaker is indicating politeness towards the hearer. Even more importantly, several pieces indicate respect towards the grandfather: the honorific nominative marker *-kkeseo*, the choice of doraga as the expression of "die" (literally "return") and the honorific marker *-sy-* on the verb. In Korean one cannot avoid marking respect — for the hearer and for the referent. To drop such markers and say, for example

할아버지가 죽었어
harabeoji-ga chugeoss-eo
(grandfather+nom die+past+linking-vowel)

with the normal nominative marker *-ga*, the common verb for "die" *chug*, with no honorific marker, and failing to use the politeness marker *-yo*, would convey a lack of respect for the grandfather and for the hearer. It is these kinds of differences that are at the core of what makes it tricky to quantify the amount of "information" conveyed by different languages. As Jakobson (Jakobson, 1959) famously noted, "languages differ essentially in what they must convey and not in what they may convey."

As the previous example suggests, many such differences between languages involve morphosyntactic or morphosemantic features, which very often are expressed by morphological marking, though in some cases (as with the honorific verb for *die* in Korean) may be expressed by choices of lexical items.

If one had a database of close translations between a set of typologically diverse languages, with detailed marking of morphosyntactic and morphosemantic features, one could hope to quantify the differences between how these different languages convey information. Since no appropriate database exists we decided to construct one. The purpose of this paper is to present our design for the database, along with some preliminary results. We plan to release the dataset once complete.

2. Dataset

Our data are taken from a few domains of interest to Google including driving directions and answers generated from structured data for Google Now™.[1] Obviously such examples are but a subset of the ways in which language is used to communicate: The reason for picking data from this circumscribed set of domains is that for part of the data at least, the text corresponds to, and in a real application would be automatically generated from, data in a defined format (see below for an example). Therefore the basic intended meaning of a message is to a large extent given, thus obviating the need to do semantic annotation. By producing parallel target sentences in various languages, and making sure that the translations are as close as possible, while still being stylistically and socially acceptable, we can be minimize differences in information content that might arise for irrelevant reasons, such as liberal choices of wording taken by the translators. We are therefore focusing as much as possible on what the languages must convey, rather than one what they may convey.

Our initial dataset consists of 85 sentences from a mix of domains for the following languages: English, French, Italian, German, Russian, Arabic, Korean and Mandarin Chinese. These languages were chosen from among languages for which we have very good resources, to be somewhat typologically balanced, representing languages of the "isolating" or quasi-isolating type (English, Mandarin), "inflectional" (French, Italian, German, Russian, Arabic) and "agglutinative"' (Korean). We are also interested in languages with rich case systems (German, Russian, Korean), gender systems (French, Italian, German, Russian, Arabic),and a variety of language families — four in this case.[2]

For the current dataset, translators were given the English original in a spreadsheet, and were given the following instructions:

> This is a request for natural sounding and socially appropriate translations which should be inserted directly into the provided spreadsheet in the column for your language. Important: There is no character restriction for these translations. However, we want translations that are succinct as possible, natural sounding, and socially appropriate.

As noted above, some of the examples in actual applications would be generated from a universal data structure that represents the basic message being conveyed. To give a simple example, take the sentence *in 100 meters, turn left*, and its Russian equivalent

через 100 метров поверните налево
(through 100 meter+gen-plu turn+imper-addr-resp left)

In each case, these are assumed to be hand translated (though of course automatically generated in a real application) from a data structure that looks roughly as follows:

```
{ distance: { number: 100 units : "meter"}
  action: "turn"
  direction: "left" }
```

3. Annotation

The text messages are annotated by native speakers of each language, using an internal annotation tool designed for annotating spans of text with features. The annotations use a hierarchy of universal morphosyntactic and morphosemantic features, based heavily on GOLD 2010 (GOLD, 2010), and those of the Surrey Morphology Group (Kibort, 2008), with subsets of the hierarchy carefully defined for each language in consultation with linguistic experts of that language. Annotators were asked to indicate for each sentence all the elements that express a given feature.

The annotators thus need to identify language-specific information that is secondary to the basic meaning of the message. Thus for the example above the English annotators would mark:

- PluralNumber: *-s*
- ImperativeForce: *turn*

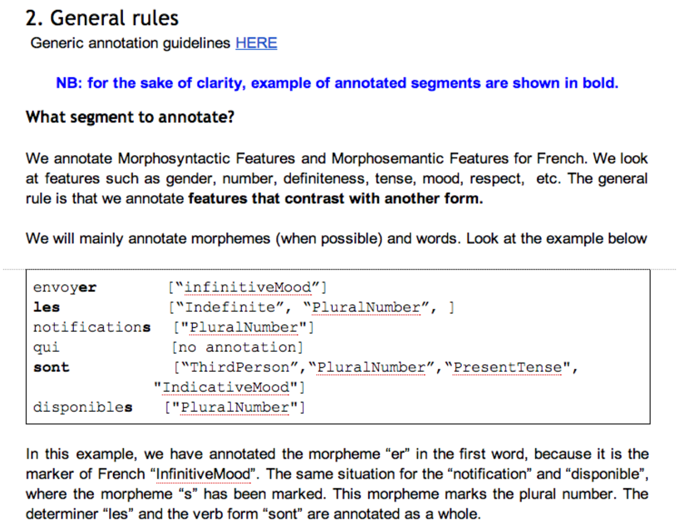
The Russian annotators would need to mark:

- PluralNumber+GenitiveCase: -ов
- SecondPerson+ AddresseeRespect+ImperativeForce: -ите

For each language, a set of features needed for that language was decided in consultation with a language expert. Then detailed instructions were developed with examples of what to annotate. In general, annotators were instructed to mark a feature only if there is an overt expression of that feature either in the form of an explicit morphological mark, or else in the form of contrast with a marked form. Thus for example *dog* is SingularNumber even though there is no morpheme

[1] Note that no Google user data is included in our data collection.

[2] At the time of writing, data for Indonesian are in preparation.



2. General rules

Generic annotation guidelines HERE

NB: for the sake of clarity, example of annotated segments are shown in bold.

What segment to annotate?

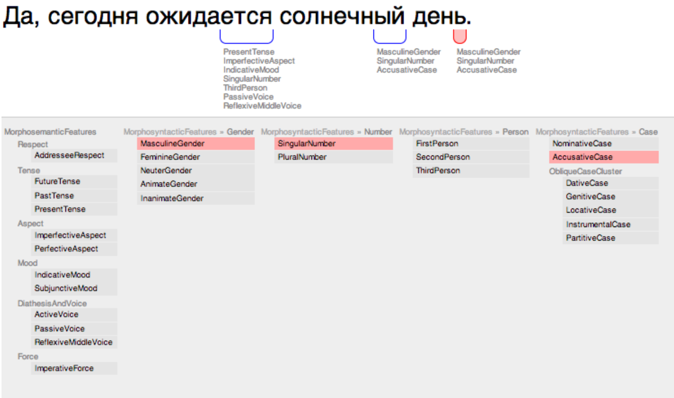
We annotate Morphosyntactic Features and Morphosemantic Features for French. We look at features such as gender, number, definiteness, tense, mood, respect, etc. The general rule is that we annotate **features that contrast with another form.**

We will mainly annotate morphemes (when possible) and words. Look at the example below

```
envoyer          ["infinitiveMood"]
les              ["Indefinite", "PluralNumber", ]
notifications     ["PluralNumber"]
qui               [no annotation]
sont              ["ThirdPerson","PluralNumber","PresentTense",
                "IndicativeMood"]
disponibles     ["PluralNumber"]
```

In this example, we have annotated the morpheme "er" in the first word, because it is the marker of French "InfinitiveMood". The same situation for the "notification" and "disponible", where the morpheme "s" has been marked. This morpheme marks the plural number. The determiner "les" and the verb form "sont" are annotated as a whole.

Figure 1: Sample of instructions for French annotators.



Да, сегодня ожидается солнечный день.

Figure 2: Example of an annotation window in the annotation tool for a Russian sentence.

marking the singular, since it contrasts with *dogs*. Similarly the French singular definite article *le* is MasculineGender since it contrasts with FeminineGender *la*. However, in the plural, *les* never shows a contrast in gender, so gender would not be marked. A fragment of the instructions for French giving a concrete example is given in Figure 1. Figure 2 shows an example of an annotation of a Russian sentence using the annotation tool.

The total mean number of features, as well as the number of bytes annotated for each language is as follows:

	# feats	# bytes		# feats	# bytes
Ar	2034	4441	De	2275	5972
En	851	4842	Fr	1582	6297
It	1757	5748	Ko	672	2234
Ru	1657	5046	Zh	187	1879

We hope eventually to expand the set to around 1,000 sentences per language. However, as we will see below, we can already learn a lot from the dataset thus far developed.

4. Measuring Information

Before we turn in the next section to the analysis and the actual measures we develop, we wish to lay out a few general issues on how different kinds of features relate to information content.

The simplest approach to measuring information would simply be to count the morphosyntactic and morphosemantic features that are marked for each sentence. However, this would surely overreward some bits of information that are less informative than others. For example, consider grammatical gender. On the one hand gender could be useful in language comprehension: gender agreement of a predicate adjective with a preceding noun, for example, might help to reinforce the fact that the adjective is predicated of the noun. On the other hand, grammatical gender is purely formal feature and does not really convey much information in the common-language sense of this term. This is in contradistinction to, say, case marking, which can convey quite a bit of information about the role of a noun phrase in the sentence. Put another way, formal gender (as opposed to natural gender) generally corresponds to nothing in the world. On the other hand, case marking, even if indirectly, marks the role of particular nominals in an action or state of affairs, and thus can be said to reflect something in the world.

Another possible dimension to compare features is in terms of their contribution to the truth value versus their pragmatic status. If a speaker mixes up nominative and accusative case, the result could be a sentence that is no longer true: if the situation is that John ate a fish, marking *John* with accusative and *fish* with nominative would result in a sentence that is probably false. On the other hand, using an inappropriate respect marker would not change the truth condition, though it would probably render the result socially inappropriate. Depending now how we weigh the importance of denotation versus pragmatics, we could quantify the "importance" of these different kinds of features.

We also do not want to count the same piece of information twice: in our Korean example above, there were multiple parts that conveyed respect for the referent (grandfather), but taken together these mark a single instance of ReferentRespect. This latter point can in fact easily be taken care of either during the annotation process by requiring the annotators to describe the multiple exponents of a particular feature as being the same feature instance, or by post hoc processing of the annotations.

As noted above, we also need a good definition of "space": as we discuss below we consider several measures of this, including byte count, character count, and measures correlated with phoneme count.

5. Analysis

Not all of the sentences in our set currently are generated from a template of the kind discussed above for the driving directions example. Thus in order to get

DENOTATIONAL	
AccusativeCase	ActiveVoice
ComitativeCase	Comparative
ConditionalMood	DativeCase
Definite	FirstPerson
Focus	FutureTense
GenitiveCase	ImperativeForce
ImperfectTense	ImperfectiveAspect
Indefinite	IndicativeMood
InstrumentalCase	InterrogativeForce
LocativeCase	NominativeCase
Ordinal	PartitiveCase
PassiveVoice	PastParticiple
PastTense	PerfectiveAspect
PluralNumber	PresentParticiple
PresentTense	ProgressiveAspect
ReflexiveMiddleVoice	SecondPerson
SingularNumber	SpeculativeForce
SubjunctiveMood	Superlative
ThirdPerson	Topic
VolitiveForce	
SOCIOLINGUISTIC	
AddresseeRespect	ReferentRespect
FORMAL	
AnimateGender	FeminineGender
InanimateGender	MasculineGender
NeuterGender	
InfinitiveMood	EMPTY

Table 1: Features used for the eight languages currently considered and their broad classification.

an estimate of the amount of information conveyed by the basic message, we cannot currently simply count slots in the formal template, so we need to consider an alternative.

One reasonable estimate is based on the idea that that the amount of basic "meaning" conveyed by a message is loosely proportional to the number of words in the message: on balance, if there are more words in a sentence, the amount of information being conveyed is greater. The next question then is which language to assume as the basis for this estimate. The source language for all of these translations was English, so one could take English as the basis. On the other hand, English at least contains words such as articles that are obligatorily present but do not always contribute much to the basic message being conveyed. Because it is the most "compact", we have chosen the number of words in the Chinese sentences as the basis of the computation of amount of basic meaning being conveyed. In any case, the results we report below are not much affected by the choice of language as the basis. Since Chinese is written with no spaces between words, we employed the Google-internal CJK segmenter, which is based on a word bigram model. Thus, to recap, the basic semantic contribution of the "message" in a sentence is simply the number of words in the Chinese version.

Over and above the basic meaning being conveyed are various morphosyntactic and morphosemantic features. The features that are needed for the eight languages currently in our set are given in Table 1. These we have divided into four categories:

- DENOTATIONAL: These are features that relate directly or indirectly to entities, relations or events. Thus case features relate indirectly to relations between named entities and other entities or events. Similarly, mood, tense and aspect features relate for the most part to things like the time of an event, its completion, whether the addressee is being instructed to perform a certain task, and so forth. Similarly, number features relate to the number of participants in an event.

- SOCIOLINGUISTIC: Currently in this category are AddresseeRespect and ReferentRespect. These do not have any denotational content, but they do convey information about the relation between the speaker and the addressee or the person being discussed.

- FORMAL: Currently in this category are grammatical (not natural) gender features, which by and large serve no purpose other than as a purely formal part of the concord system of the language.

- EMPTY: Currently in this category is InfinitiveMood, which while it was tagged (since in some of the languages in our set it has an identifiable morphological mark) it seems to convey no information by itself.

Plausibly, different features convey different amounts of information. For example, GenitiveCase conveys a lot of information in that it often specifies that there is a relation of possession between two entities in a sentence.[3] On the other hand, as noted above, using the feminine form of an adjective does not really convey any information, though it serves to indicate that the adjective is related (as modifier or predicate) with a particular noun phrase. To a first approximation, then, one could weight the amount of information contributed DENOTATIONAL, SOCIOLINGUISTIC, FORMAL, and EMPTY features differently. As a first cut at this we consider the following options, with the assumption that EMPTY features always carry no weight:

DENOT.	SOCIOLGSTC	FORMAL	EMPTY
d=1.0	s=0.5	f=0.0	e=0.0
d=1.0	s=1.0	f=0.0	e=0.0
d=1.0	s=1.0	f=1.0	e=0.0

That is, each instance of a marked DENOTATIONAL feature adds 1.0 to the total score. For FORMAL features, we have two options: either 1.0 or 0.0. For SOCIOLINGUISTIC features, the choices are 1.0 or 0.5.

[3] Of course what *possession* itself means is itself a complex question, and many languages distinguish between different notions of possession.

The amount of *lexical* information in a sentence or set of sentences is thus the number of words in the Chinese version of the sentence(s), plus the weighted sum of the features marked in the sentence(s), with the weights chosen as above.

The next question is what we are comparing the number of features against. When people ask about which languages are most "efficient" at conveying information they implicitly have in mind a ratio of information per unit. What unit is the right unit for this? One might consider **word** as a basic unit, but that would of course unfairly reward languages that tend to have longer words, which will often pack more information into those words: on that measure, the most "efficient" languages would be polysynthetic languages where, often, a sentence may consist of a single word.

To take another extreme, one might consider the written representation of the text in UTF8, counted in **bytes**. Since normal Chinese morphosyllabic characters and Korean Hangeul syllables are encoded in 3 bytes, this fairly captures the fact that one can pack more information into a single character in Chinese or Korean than one can in Western European languages that use a (mostly) one-byte encoding. However, this unfairly penalizes Russian and Arabic, which use alphabetic scripts that require two bytes to encode in UTF8.

Probably the most reasonable simple option for written text is what we will call **normalized characters**. These are just the number of Unicode characters in the input text, *multiplied by the log of the size of the inventory* (see below), except:

- **Korean** (hangeul) syllables are recoded with their individual segmental glyphs (*jamo*).
- **Chinese** characters are replaced with a pinyin-plus-tone representation of their pronunciation.
- Since **Arabic** orthography is *defective* in lacking vowels and other segmental information, we estimate the length of the text in characters if all this information were written.[4]

Normalized character length correlates very well with phoneme length. Indeed for some languages with highly transparent grapheme-phoneme correspondences, or in systems like pinyin that are quasi-phonemic transcriptions, the correlation is almost perfect. Since obviously a larger inventory allows one to pack more information into a unit we normalize these text lengths by the log of the size of the inventory, where we assume the following sizes:

Arabic	31	(28 consonants plus 3 vowels)
Chinese	29	pinyin plus tone
English	26	letters
French	40	letters incl. accents
German	30	letters incl. accents
Italian	30	letters incl. accents
Korean	51	jamo + combinations
Russian	30	letters

Thus we have three settings for weighting morphosyntactic and morphosemantic features, and three different ways of computing the lengths of the texts. Figure 3 shows the 9 plots for the ratio $\frac{\text{information}}{\text{length}}$ for eight languages. In the cases we have more than one annotation for a language (English, French, Korean and Chinese), each annotator's data are shown separately; in any case it can easily be seen that there is little difference between the annotators for a given language.

The weightings of the features above however ignore the fact that in many cases the same setting of a feature — say MasculineGender — on several words really reflects a *single* morphosyntactic event. For example for the French phrase *la maison verte* 'the green house', both the article *la* and the adjective *verte* mark FeminineGender, but since both of these are really instances of agreement with *maison*, they do not really reflect multiple instances of that feature. In Figure 4, we approximate $\frac{\text{information}}{\text{length}}$ for this situation by eliminating duplicates of morphosyntactic features within a sentence.

Figure 5 shows $\frac{\text{information}}{\text{length}}$ when we set all the morphosyntactic and morphosemantic feature weights to 0, thus treating all languages as conveying the *same* amount of information, and simply measuring how efficient the communication of that information is according to the different notions of text length.

Finally Figure 6 shows $\frac{\text{information}}{\text{length}}$ when we weight "lexical" features as before, but weight the morphosyntactic and morphosemantic features by the negative log probability of that feature, where the probabilities are estimated separately for each annotator for each language. Obviously given the small sample size this estimate is very crude, but the more commonly used features (e.g. gender in Arabic) would tend to get a lower weight by this measure, as desired. This measure is more related to information theoretic measures of information, and so is useful to compare whether such approaches are likely to give radically different answers to the simpler measures already discussed.

If we compare the various settings in Figure 3, within each length measure (word, byte, normalized character) the orderings of the languages are essentially the same.

feats per word	ar>ru>de>ko>it>fr>en>zh
feats per byte	ko>zh>ar>de>ru>it>fr>en
feats per norm. char	de>ru>it>ar>en>fr>ko>zh

For the features-per-byte measure, Korean and Chinese pack in the most information, giving those languages a big advantage on Twitter as we noted in the introduction. With elimination of feature

[4] For the purpose these analyses, we randomly picked 10 of the Arabic sentences, fully transcribed them in a standard Romanization system with all the vowel and other diacritic information fully marked, computed the ratio between the length of that transcription and the original text, in characters, and then used that ratio as a multiplier for all the sentences.

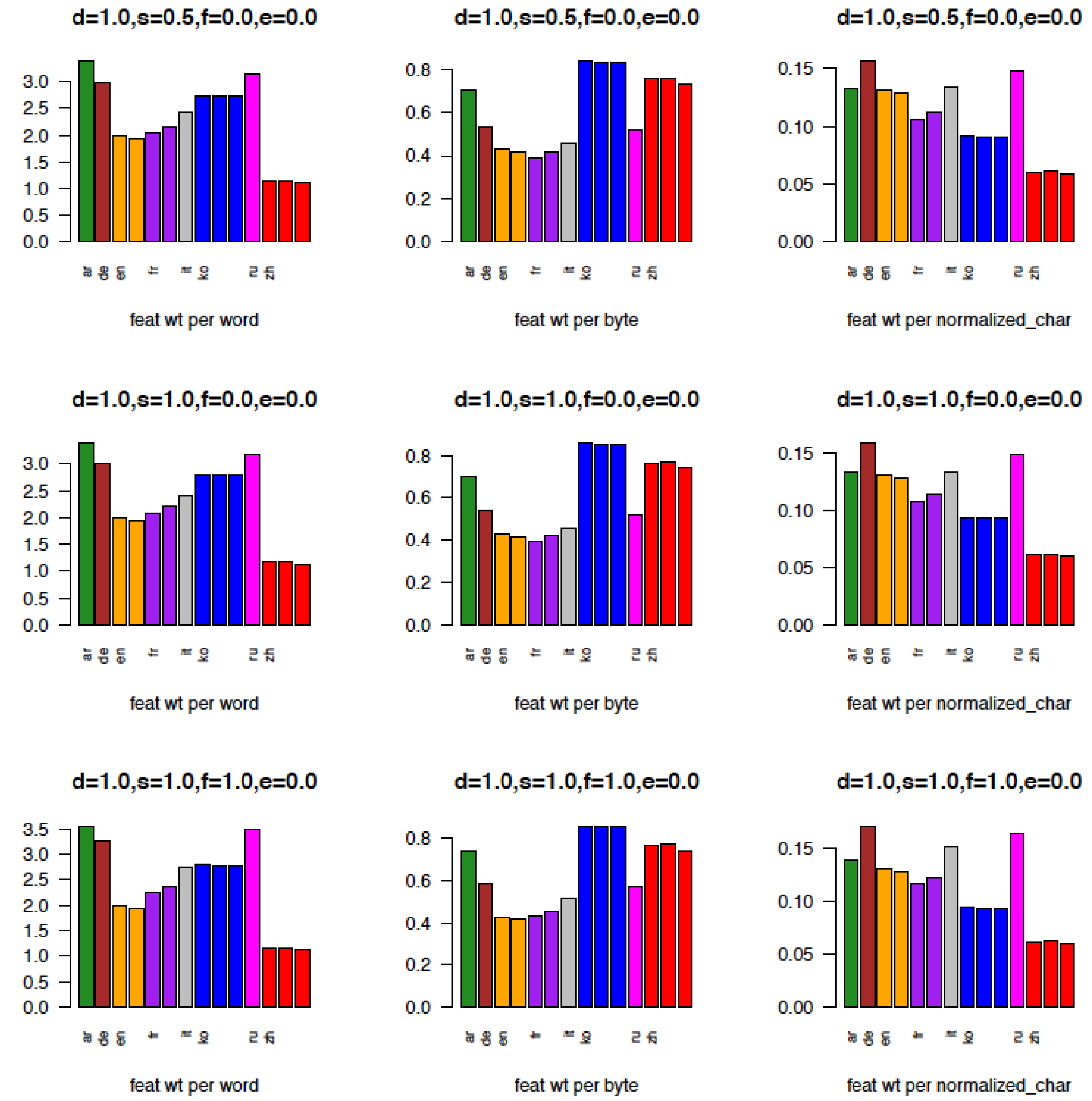


Figure 3: $\frac{\text{information}}{\text{length}}$ (vertical axis) for eight languages. See the text for the explanation of the headers and footers for each plot. Note that the feature weights are the same along each *row*. Annotators for each language (where there was more than one), are grouped together. Colors for the languages for all plots are: Arabic – green, German – brown, English – orange, French – purple, Italian – grey, Korean – blue, Russian – magenta, Chinese – red.

duplicates (Figure 4) the following patterns emerge:

feats per word	ko>ru>ar>de>en>it>fr>zh
feats per byte	ko>zh>ar>ru>en>de>it>fr
feats per norm. char	ru>en>de>it>ko>ar>fr>zh

Here again Korean and Chinese top the list by the features-per-byte measure. Weighting morphosyntactic and morphosemantic features as 0 (Figure 5) comes out as follows, with Korean being about the same as German, and French as Chinese, in the features-per-normalized-character condition:

feats per word	ko>ru>ar>en>de>it>fr>zh
feats per byte	zh>ko>ar>en>ru>it>de>fr
feats per norm. char	en>ru>it>ko ~ de>fr ~ zh>ar

Finally, the statistical negative log probability weighting (Figure 6) yields:

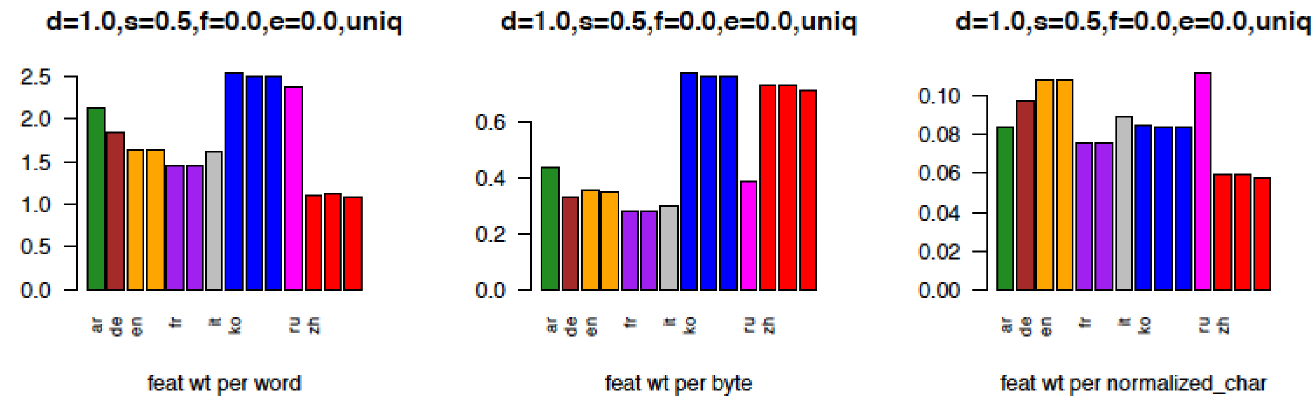


Figure 4: $\frac{\text{information}}{\text{length}}$ for eight languages, when we eliminate duplicates for any feature within a sentence.

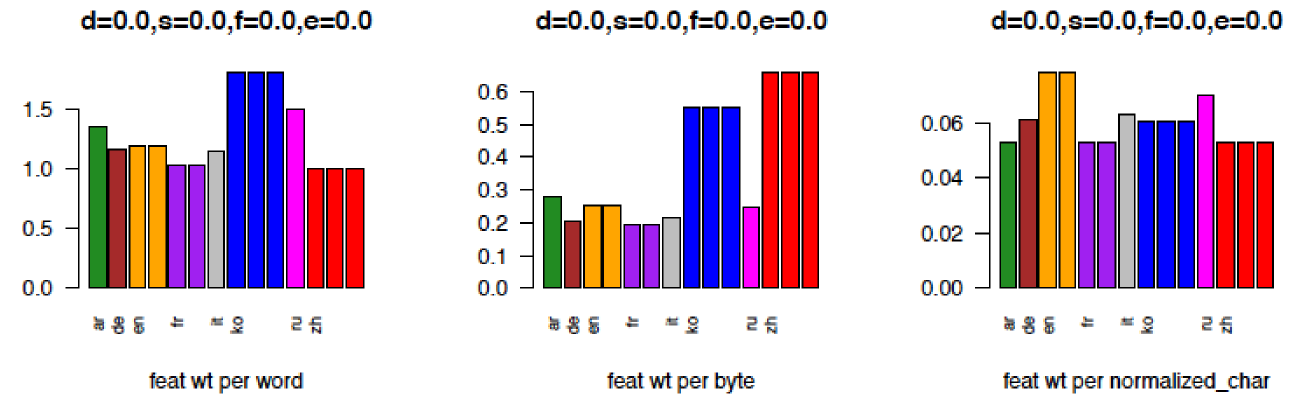


Figure 5: $\frac{\text{information}}{\text{length}}$ for eight languages, when we *only* consider the lexical features (setting all weights for morphosyntactic and morphosemantic features to 0).

feats per word	ru>ar>de>it>ko ~ fr>en>zh
feats per byte	ar>ko>de>ru>it>zh>fr>en
feats per norm. char	ru>en>de>it>ko>ar>fr>zh

If now we consider *all* of the ways in one might measure information, both in terms of feature weightings and the length against which that is being compared, and for each condition, weights the language by the position in the partial ordering, and then sums over all conditions, one arrives at the following ranking:
ru > ar > de > ko > it > en > zh > fr
To a large extent this ranks the highly inflected languages at the top of the hierarchy, with the single exception of French, whose position at the bottom may be due in part to the French translations being more verbose than for the other languages, but in part to other differences — e.g. French being non-pro-drop and thus expressing pronouns where, say, Italian does not, and more use of possessives in French.

This, then, constitutes an answer to the question of which languages are the most efficient at communicating information. While it is not the only possible answer, it is at least defensible.

6. Further work

The dataset thus far developed is more modest in scope than what we eventually intend. Nonetheless, as we have seen, we already have results that are informative about the question of how efficient different languages are at encoding information. Also, now that we have a baseline set of annotations against which to compare, we will explore using automatic morphological taggers (which we have for many of the languages we have examined) to replace human annotators for this task.
There are still further issues to consider that we will address in future work. One important issue is that features that are encoded morphologically in many languages may be encoded structurally in others. English does not mark AccusativeCase for any words outside the pronominal system. But it does mark the functions that AccusativeCase marks in many languages structurally. This is currently not taken into account, and so languages like English and Chinese are being penalized for their lack of morphological expression. Note that this is different from the situation with sociolinguistic markers such as Korean respect markers, since in general most English sentences do not mark such information even implicitly.

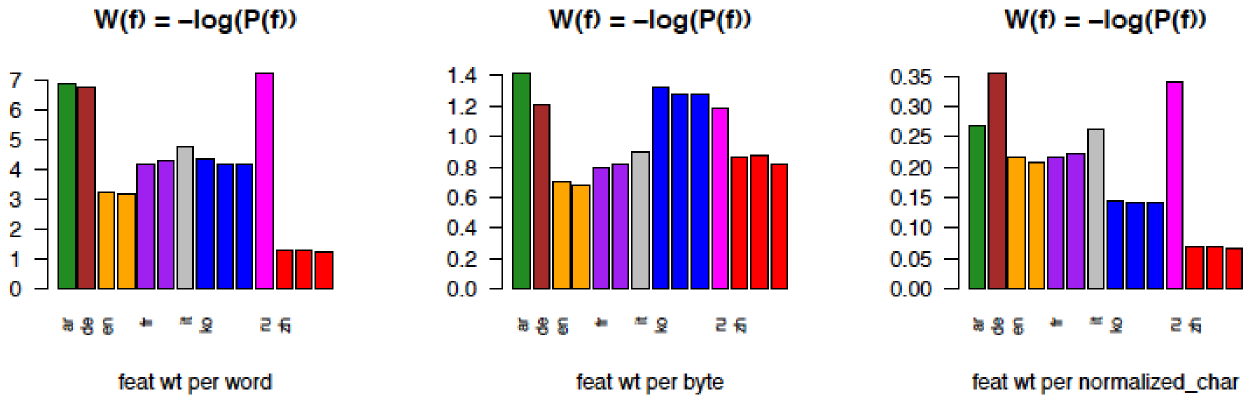


Figure 6: $\frac{\text{information}}{\text{length}}$ for eight languages, weighting the lexical features as before, but weighting each feature as the negative log probability of that feature setting, estimated separately for each annotator.

Other measures of the "importance" of a feature may be considered. For example (Acedański and Przepiórkowski, 2010) discuss various measures that better assess the importance of tagging errors for different applications, and such measures may be relevant here too as a way to assess the salience of a feature.

Many of the assumptions that have gone into the database and its analysis may strike the reader as silly. For example, why assume that a given morphosyntactic feature has the same amount of information as some other, or as a lexeme? Why do all case features have the same weight in some of our analyses? And so forth. One could of course as easily ask, why not? The point is that in order to make any sense at all of the question of how much information is conveyed by different languages, in the same amount of space, one must make concrete assumptions about what one means by "information" and by "space". Only then can one even begin to answer the question. Since to our knowledge nobody has ever tried to formally quantify these notions with a view to answering the main question, one may view the current work as an initial step towards an answer, which can hopefully be improved upon with further research. In the worst case, this work is a *reductio ad absurdum* of the question.

Beyond the linguistic questions that have motivated this work, we believe that this work is also relevant for practical applications such as language modeling and natural language generation (NLG). Considering just NLG, one of the problems in multilingual generation from the same core data structure is how to make the result sound both correct and appropriate. The annotated data we will provide will be of direct use to constructing models for NLG since it will indicate exactly what morphosyntactic and morphosemantic information must be added beyond what is in the information in the core message, and where that information must be added.

We plan to make the current dataset along with the features annotated for each language freely available in the near future.[5]

7. Acknowledgements

We wish to thank the language experts who helped us with designing language-particular feature sets and annotating the data: Costanza Asnaghi, Elixabete Murguia Gomez, Zainab Hossainzadeh, Josie Li, Thomas Meyer, Fayeq Oweis, Tanya Scott. Thanks also to Daniel van Esch for helping arrange for some of the annotation work.

8. References

Szymon Acedański and Adam Przepiórkowski. 2010. Towards the adequate evaluation of morphosyntactic taggers. In *COLING*.

GOLD. 2010. `http://linguistics-ontology.org/gold/2010`.

Roman Jakobson. 1959. On linguistic aspects of translation. In R. A. Brower, editor, *On Translation*, pages 232–239. Harvard University Press, Cambridge, MA.

Anna Kibort. 2008. A typology of grammatical features. 11. `http://www.features.surrey.ac.uk/inventory.html`.

Claude Shannon. 1948. A mathematical theory of communication. *Bell Labs Technical Journal*, 27:379–423, 623–656, July, October.

Ashley Taylor. 2013. Twitter in foreign language. `http://www.theconnectivist.com/2013/04/twitter-in-foreign-languages/`.

[5] Please contact the first author for further details.