Predicting language-learning difficulty

Michael Cysouw

1. Introduction[1]

The difficulty people have in learning a foreign language strongly depends on how different this language is from their native tongue (Kellerman 1979). Although this statement seems uncontroversial in the general form as it is formulated here, the devil lies in the detail, namely in the problem how to define differences between languages. In this paper, I investigate various factors that quantify differences between languages, and explore to which extend these factors predict language learning difficulty. This investigation results in concrete predictive formulas that derive the learning difficulty for native English speakers depending on a small selection of linguistic factors of the language to be learned.

Section 2 presents the data for language learning difficulty that will be used in this paper. This data originates at the Foreign Services Institute (FSI) of the US Department of State and it includes only approximate average learning times of foreign languages for English speakers. The data is rather rough, but it is highly interesting because it gives comparable estimates for language learning difficulty for a large number of strongly different languages from all over the world. Section 3 investigates the relation of these estimates for language learning difficulty to very general predictors like geographical distance and genealogical affiliation. In both cases, the further away a language is from English, both geographically and genealogically, the more difficult a language is expected to be. All empirical effects point in the expected direction, though the factor Germanic vs. non-Germanic turns out to be the strongest predictor for language-learning difficulty.

Section 4 takes up the differences in writing systems as used for the various languages in the current sample. Using the Universal Declaration of Human Rights, the orthographic similarity between English and other languages is established. For languages with a Latin script, there is a strong correlation between language learning difficulty and the similarity in frequency distribution of orthographic symbols. Section 5 investigates structural grammatical properties of languages using data from the World Atlas of Language Structures. I establish which structural differences from English most strongly correlate with language learning difficulty for English speakers.

Section 6 combines all these factors and searches for suitable models to predict language-learning difficulty. Two different kinds of models are proposed: one based on agglomerative similarity values between English and other languages and one based on more practical binary predictors describing actual characteristics of the languages to be learned. In the agglomerative models, language learning difficulty language learning difficulty can be predicted by a strong factor related to structural typological similarity and a weaker subsidiary factor related to the writing system. In the binary models, the main factors were related to having a Latin script or not, being an Indo-European language or not, and various structural characteristics, namely prepositions vs. postpositions, accusative vs. ergative alignment and the presence vs. absence of obligatory plural marking of nouns.

2. Measurements of language learning difficulty

For this paper, I will use two different measurements of the difficulty English speakers appear to have when learning specific foreign languages. Both these rather rough measurements originate at the Foreign Services Institute (FSI) of the US Department of State. They arose in the context of planning the amount of resources necessary for language teaching when preparing US citizens for foreign detachment.

The first measurement of language learning difficulty is an assessment of the number of class hours it takes to achieve general proficiency in speaking and reading in a foreign language (where "general proficiency" is defined by the language skill level descriptions from the Interagency Language Roundtable). It is basically a three-level scale (I for "easy", II for "middle", III for "hard"), which Jackson and Kaplan (2001: 77) explain as follows:

> "The categories indicate gross differences in how hard it is for native speakers of American English to learn different languages. […] These categories […] are based solely on FSI's experience of the time it takes our learners to learn these languages. […] The more commonalities a language shares with English – whether due to a genetic relationship or otherwise – the easier and faster it is for a native English speaker to learn that language. […] The more dissimilar a new language is - in structure, sounds, orthography, implicit world view, and so on - the longer learning takes." (Jackson and Kaplan 2001: 77)

Actual assessments for different languages are not currently available through any documentation from the FSI. However, there used to be a website from the FSI with information about the languages of the world, which can still be accessed through the Internet Archive. On this website a classification of various languages is given according to the three-level scale, as reproduced here in table 1.2 In addition, it is noted that various languages are "somewhat more difficult for native English speakers to learn than other languages in the same category". These languages are marked with a star in the table. Further, German is specifically indicated to fall in between category I and I*, so I have added a separate category for German. I will use the resulting seven-level scale as a measurement of difficulty and refer to this measurement "FSI-level" in the remainder of this paper. If not specifically indicated, I will interpret the seven levels as a linear numerical scale from one to seven, as shown in table 1.

The second measurement of language learning difficulty used in this paper is reported in Hart-Gonzales and Lindemann (1993), as cited in Chiswick and Miller (2005: 5-6). As above, I have not been able to get hold of the original source by Hart-Gonzales and Lindemann, so I am simply using the numbers as presented in Chiswick and Miller. They explain this measurement as follows:

> "The paper by Hart-Gonzalez and Lindemann (1993) reports language scores for 43 languages for English-speaking Americans of average ability after [24 weeks] of foreign language training. […] The range is from a low score (harder to learn) of 1.00 for Japanese to a high score (easier to learn) of 3.00 for Afrikaans, Norwegian and Swedish. The score for French is 2.50 and for Mandarin 1.50. These scores suggest a ranking of linguistic distance from English among these languages: Japanese being the most distant, followed by Mandarin, then French and then Afrikaans, and Norwegian and Swedish as the least distant." (Chiswick and Miller 2005: 5)

I will refer to this measurement as the "24-week ability" score in the remainder of this paper. The individual scores from Hart-Gonzales and Lindemann are reproduced here in table 2.

Table 1. FSI levels of difficulty for various languages (higher levels represent greater difficulty).

FSI Level	Languages
1 (I)	Afrikaans, Danish, Dutch, French, Italian, Norwegian, Portuguese, Romanian, Spanish, Swedish
2 (I)	German
3 (I*)	Indonesian, Malay, Swahili
4 (II)	Albanian, Amharic, Armenian, Azerbaijani, Bengali, Bosnian, Bulgarian, Burmese, Czech, Greek, Hebrew, Hindi, Icelandic, Khmer, Lao, Latvian, Lithuanian, Macedonian, Nepali, Pashto, Persian, Polish, Russian, Serbo-Croatian, Sinhalese, Slovak, Slovenian, Swahili, Tagalog, Turkish, Ukrainian, Urdu, Uzbek, Xhosa, Zulu
5 (II*)	Estonian, Finnish, Georgian, Hungarian, Mongolian, Thai, Vietnamese
6 (III)	Arabic, Cantonese, Korean, Mandarin
7 (III*)	Japanese

Table 2. Average ability scores for various languages after 24 weeks of foreign language training (low values represent less communicational ability).

24-week ability	Languages
1.00	Japanese, Korean
1.25	Cantonese
1.50	Arabic, Lao, Mandarin, Vietnamese
1.75	Bengali, Burmese, Greek, Hindi, Nepali, Sinhalese
2.00	Amharic, Bulgarian, Czech, Finnish, Hebrew, Hungarian, Indonesian, Khmer, Mongolian, Persian, Polish, Serbo-Croatian, Sinhalese, Tagalog, Thai, Turkish
2.25	Danish, German, Spanish, Russian
2.50	French, Italian, Portuguese
2.75	Dutch, Malay, Swahili
3.00	Afrikaans, Norwegian, Romanian, Swedish

These two measurements of language learning difficulty are strongly correlated (Pearson $r = -0.85$, $p = 1.8\text{e-}12$). The correlation is negative because more difficult languages have a high FSI-level score, but learners will have a low ability after 24 weeks of language training. Although the two measurements are highly correlated, there are still notable differences (e.g. concerning the position of Danish and Spanish). Also, there are more languages with an FSI-level than with a 24-week ability score, which makes the somewhat more coarse-grained FSI-level scale more telling for quantitative comparisons. For these reasons, I will use both measurements in the rest of this paper.

It should be noted that these measurements of language learning difficulty are extremely rough. Not only do they just distinguish a few levels of "difficulty", they also do not include any information about the background of the learners and the process of the learning itself, both factors known to have significant influence on the language learning difficulty (cf. Schepens, van der Slik and van Hout, this volume). However, given the origin of the current data, it can be assumed that the kind of people entering the learning and the kind of lessons presented to them are rather homogeneous, so that ignoring these factors – while unfortunate – is probably not influencing the current results significantly.

3. Geography and genealogy

The difficulty in learning a language is supposedly related to the degree of difference between the language(s) a learner already knows and the language the learner wants to learn. There are two factors that are known to be strongly correlated to the degree of difference between languages. First, the closer two languages are geographically, the smaller the differences are expected to be. And, second, the closer the genealogical relationship between two languages, the smaller the differences will be.

To assess the geographical distance between English and the other languages considered in this paper, I will locate English in the City of London. This is of course a completely illusory point of origin considering the current world-wide distribution of English speaking communities. At best, it represents the most prestigious location of English speakers up until a century ago. Likewise, I will use point locators for all the other languages listed in Section 2, which in many cases are also spoken over widely dispersed territories. The measurements of geographical proximity are thus to be taken as very rough approximations (verging on the nonsensical) of the actual social distance between real speakers. In practice, I will use the coordinates as listed in the *World Atlas of Language Structures* (WALS, Haspelmath et al. 2005) as the point locations for the computation of geographical distance between languages.

As the distance between two point locators representing languages, I will use the distance "as the bird flies", i.e. the great-circle distance, further assuming the world to be a perfect sphere and ignoring elevation difference.

Such a simplistic assumption will of course further lessen any real-world impact of the current conceptualization of geographical distance between languages. However, the correlation between this notion of geographical distance and language learning difficulty is still clearly significant, though not very strong (FSI-levels: $r = 0.38$, $p = 0.003$; 24-week ability scores: $r = –0.43$, $p = 0.004$). So, indeed, languages that are further away geographically from English are in general more difficult to learn for English speakers. As expected, Afrikaans, Swahili, Malay and Indonesian are the most extreme outliers to the one side, being much easier to learn than expected from their large geographical distance from English. In contrast, Arabic is more difficult to learn compared to the relative geographical proximity to English (see figure 1).

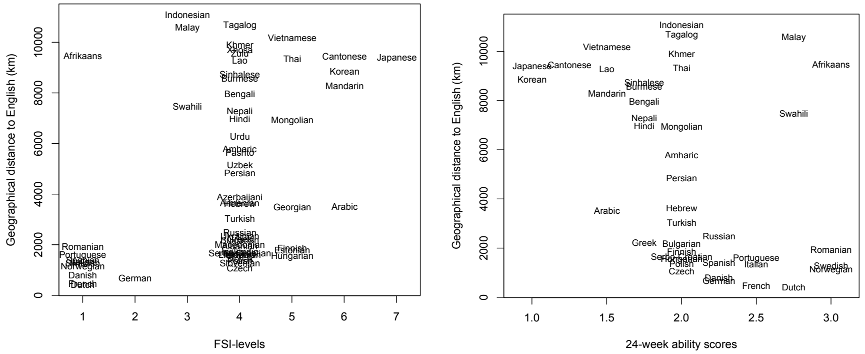

Figure 1. Correlations between geographical proximity to English and language learning difficulty, FSI-levels to the left and 24-week ability scores to the right.

Genealogically closely related languages – i.e. languages from the same language family – are also expected to be relatively similar, and thus be easier to learn. Closely related languages are often structurally similar and also share part of the lexicon, which might ease learning. A substantial amount of shared lexicon of course also increases the chance of the occurrence of false friends, inhibiting ease of learning. However, this effect is probably not relevant for the relatively low proficiency levels with which we are dealing in this paper. Further complicating matters is that it is not immediately obvious how to quantify genealogical proximity of languages. Although it is clear that English is genealogically closer to German than to Greek, Hindi, or Cantonese (in decreasing order), giving numbers to such qualifications strongly depends on the details of the historical reconstruction. As a practical solution, I will use the two-level genealogical classification from Dryer (2005b). Dryer distinguishes a level of closely related languages ("genus") and a level of more distantly related languages ("family").

As expected, languages from within the same genus as English, viz. Germanic, are easier to learn for English speakers than languages from different genera (FSI-levels: Germanic mean 1.57 vs. non-Germanic mean 4.00, $t = – 5.27$, $p = 7.5e\text{-}4$; 24-week ability scores: Germanic mean 2.71 vs. non-Germanic mean 1.95, $t = 4.52$, $p = 0.002$). Similarly, languages from the same family as English, viz. Indo-European (IE), are easier to learn for English speakers than languages from different families (FSI-levels: IE mean 3.06 vs. non-IE mean 4.58, $t = – 4.86$, $p = 9.2e\text{-}6$; 24-week ability scores: IE mean 2.27 vs. non-IE mean 1.83, $t = 3.16$, $p = 0.003$).

So, languages that are geographically far away, and such that are non-Germanic or better still non-Indo-European, are difficult to learn for English speakers. But, these three factors are to some extend measuring the same facts and are clearly all related to each other (cf. Cysouw 2012). Non-Indo-European languages are by definition also non-Germanic languages, and both these groups of languages will generally be geographically far away. To assess the relative impact of these factors for language learning difficulty, I combined the three factors in a linear regression model, as shown in table 3.

These numbers can be interpreted as follows. For the FSI-levels, the default level to learn a foreign language is at 1.52 (viz. the intercept estimate), while the presence of any of the other factors increases the difficulty of the language: geographical distance leads to an increase of 0.25 per 10.000 km, being non-Germanic increases learning difficulty with 1.85, and being non-Indo-European increases learning difficulty with 1.03. For example, for Hindi (located at about 7.000 km from English) this model predicts an FSI-level of 3.55 (=1.52+0.25·0.70+1·1.85+0·1.03), while the actual FSI-level is at 4. However, the geographical factor is not significant, and removing this factor indeed results in a simpler model with equal residual deviance. So, while both genealogical levels are significant factors, the influence of geographical distance is already accounted for to a large extend by the genealogical factors.

For the 24-week ability the results in table 3 are similar, though in this case the factor non-Indo-European is also not significant. Further note that the intercept estimate (2.77) here represents the maximum ability after 24 weeks, while all factors reduce the predicted ability. Again taking Hindi as an example, the model predicts an ability of 2.00 (=2.77–0.28·0.70–1·0.57–0·0.19), while the actual ability as listed in the data used in this paper is 1.75. In this table, only the Germanic vs. non-Germanic parameter is significant, and this parameter was also the strongest in the calculations for the FSI-levels. This suggests that the strongest effect for language learning difficulty stem from the rather local effect of whether a language is Germanic or not.

Table 3. Regression model of geographical and genealogical factors.

	Estimate	Std. Error	t-value	Pr(> |t|)
FSI-levels				
Intercept	1.52	0.44	3.45	0.001 **
Geography	0.25	5.18	0.49	0.62
Non-Germanic	1.85	0.48	3.88	0.0002 ***
Non-Indo-European	1.03	0.37	2.79	0.007 **
24-week ability				
Intercept	2.77	0.18	15.72	< 2e-16 ***
Geography	– 0.28	2.17	– 1.29	0.21
Non-Germanic	– 0.57	0.20	– 2.85	0.007 **
Non-Indo-European	– 0.19	0.17	– 1.02	0.31

4. Writing system

Another obvious factor influencing the effort needed to learn a foreign language is the writing system that is used. Languages with a similar writing system to English are expected to be easier to learn than languages with a completely different writing system. To quantitatively assess the similarities of writing systems between languages I used the translations of the Universal Declaration of Human Rights as prepared in Unicode encoding by Eric Muller.[3] For several languages there is more than one translation available. For German and Romanian, I chose the version with the most recent orthography. For Chinese I chose the simplified orthography and for Greek the version with the monotonic script. For Malay, Bosnian and Azerbaijani I selected the translation using the Latin script, while for Serbian I chose the Cyrillic script, because these scripts seem to have the most widespread usage for these languages. Finally, for Sinhalese and Cantonese no translations were available.

It is well known that the orthographic structure of texts is a good approximation for language similarity (Damashek 1995). The most widespread application of this finding is the usage of so-called “n-gram” statistics for the identification of languages or even individual authors. The same statistics can also be used to approximate genealogical relationships (Huffman 2003; Coppin 2008). The basic idea of n-gram statistics is that the number of occurrences of each sequence of *n* character is counted in the text. I will here basically use 1-gram statistics, i.e. the simple frequency of each character. However, the situation is a bit more complicated, because sequences of Unicode characters that include combining characters are treated as one character. For Latin scripts, the most widespread combining characters are various diacritics, like tildes and accents. All possible combinations of letters with diacritics are treated as separate characters of the orthographic structure in this paper. This makes the Devanagari scripts of Indian languages especially complicated, because the syllabic combinations of consonants and vowels are treated as one character.[4] Further, not taken into account here is the widespread occurrence of multigraphs in orthographies all over the world, i.e. combinations of multiple letters to signify one element of the orthography, like <sh> or <ng>. Languages with frequent multigraphs will be estimated here to have a simpler orthography than they in reality have.

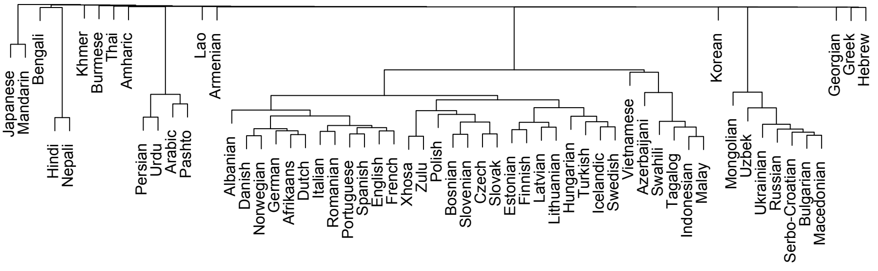

*Figure 2.*Dendrogram of 1-gram similarities of writing systems.

The similarity between two orthographies is computed by taking the Pearson correlation coefficient between the frequencies of occurrence of each character per language. The correlation matrix of all pairs of languages can be used to make a hierarchical clustering of orthographies (see figure 2). In this hierarchical clustering, the following groups are clearly discernible:

- A large cluster with all Latin scripts, including Vietnamese as an outlier;
- A cluster with the Cyrillic scripts of Mongolian, Uzbek, Ukrainian, Russian, Serbian, Bulgarian and Macedonian;

- A cluster with the Arabic scripts of Persian, Pashto, Urdu, and Arabic;
- The Devanagari script of Hindi and Nepali cluster together with a minor link to Bengali (which has its own Unicode range of characters, though uses the same separation sign as Devanagari, viz. the "danda", Unicode U+0964).
- Japanese and Mandarin cluster together based on the frequent usage of Chinese Kanji in Japanese;
- The scripts of Khmer, Burmese, Thai, Amharic, Lao, Armenian, Korean, Georgian, Greek and Hebrew do not cluster with any other script in the current set of languages.

The similarity between the English orthography and the orthographies of other languages (as measured by the Pearson correlation coefficient) is strongly negatively correlated with the difficulty of learning the language (FSI-levels: $r = – 0.56$, $p = 4.6e-6$; 24-week ability scores: $r = 0.66$, $p = 2.9e-6$). So, the more different a script is from English, the more difficult it is to learn the language for an English speaker. This correlation only makes a statement about languages that have a Latin script. For all other languages the similarity to the English script is basically zero, so they are all treated as "just different". Yet, intuitively there seems to be a great difference between learning the Cyrillic characters of Russian and the Kanji of Japanese. Simply because there are much more Kanji, the Japanese script should be more difficult. For all languages that do not have a Latin script, I investigated the difficulty of learning the languages in relation to the number of different characters used in the script. Although there is a trend discernible, this trend is not significant (FSI-levels: $r = 0.32$, $p = 0.11$; 24-week ability scores: $r = – 0.39$, $p = 0.097$). The crucial outliers in this correlation are Korean and Arabic, which are both far more difficult to learn than the (limited) size of their orthographic inventory would predict. Removing these outliers from the correlation makes the correlation between size of the orthographic inventory and the difficulty in learning the language highly significant (FSI-levels: $r = 0.51$, $p = 0.01$; 24-week ability scores: $r = – 0.73$, $p = 0.0009$).

5. Language structure

A further factor influencing the difficulty of language learning is the structural similarity between languages. The more similar the grammatical structure of two languages is, the easier it is – supposedly – for speakers of the one language to learn the other language. The notion of "grammatical structure" is interpreted here rather all-encompassing, including phonological, morphological, syntactical, lexical, semantic and discourse structures. To quantitatively assess the similarity of grammatical structure, I will use the data from the *World Atlas of Language Structures* (WALS, Haspelmath et al. 2005). This atlas provides data on the worldwide distribution of 142 structural typological parameters, including parameters concerning all above-mentioned domains of grammar. The data on sign languages and on writing systems in WALS will not be used in this paper, so there are 139 remaining structural parameters to be included in the comparison here.

There are numerous different ways to derive an overall measure of structural similarity between languages from the WALS data (cf. Albu 2006; Cysouw 2012). For this paper I will use the most basic measure of similarity, namely a relative Hamming distance. This similarity is defined as the number of similar parameters between two languages divided by the number of comparisons made. For example, English and Hindi differ in 55 structural parameters from WALS, but are similar in 69 parameters. For the remaining 15 parameters (=139–55–69) there is no data available for both languages, so no comparison can be established. This results in a structural similarity of 69/(55+69) = 0.56 between English and Hindi. On this scale, a value of one would indicate complete structural identity, while a value of zero would signify that the two languages do not share any characteristic in WALS. Because of limited data availability in WALS, Afrikaans, Malay, Slovak and Bosnian are excluded from the computations in this section.

The overall structural similarity between English and all other languages is strongly negatively correlated with the difficulty of learning those languages (FSI-levels: $r = -0.65$, $p = 4.8e\text{-}8$; 24-week ability scores: $r = 0.69$, $p = 7.4e\text{-}7$). So, the more different a language is structurally from English, the more difficult it is to learn this language. Even more interesting, is the question, which of the 139 structural parameters correlate strongly with language learning difficulty, because such parameters are indicative of structural characteristics that are difficult to learn for English speakers.

For all parameters individually, I computed the absolute value of the Goodman-Kruskal gamma for the distribution of same vs. different compared to English across the seven FSI-levels. Likewise, I computed the probability values of t-tests testing the difference of the 24-week ability scores between the set of languages with similar vs. different structure compared to English. The resulting rankings of parameter-difficulty are strongly correlated (Spearman's $\rho = -0.73$, $p = 2.2e\text{-}16$), arguing that both difficulty measures roughly agree on which parameters from WALS are difficult for English language learners. The combination of the two assessments of difficulty is plotted in figure 3, higher values indicating more difficult features. For reasons of better visibility, the negative logarithm of the probability values of the t-test is shown in this figure.

There are various interesting structural parameters that end up high in both rankings. I will specifically discuss here those parameters that have a t-test probability of less than 0.01 (for the 24-week ability scores) and, at the same time, an absolute value of gamma that is higher than 0.60 (for the FSI-levels). These boundaries do not have any special meaning. They are only used here as a practical limitation to restrict the discussion of individual features. The following WALS parameters are strongly correlated with language learning difficulty.

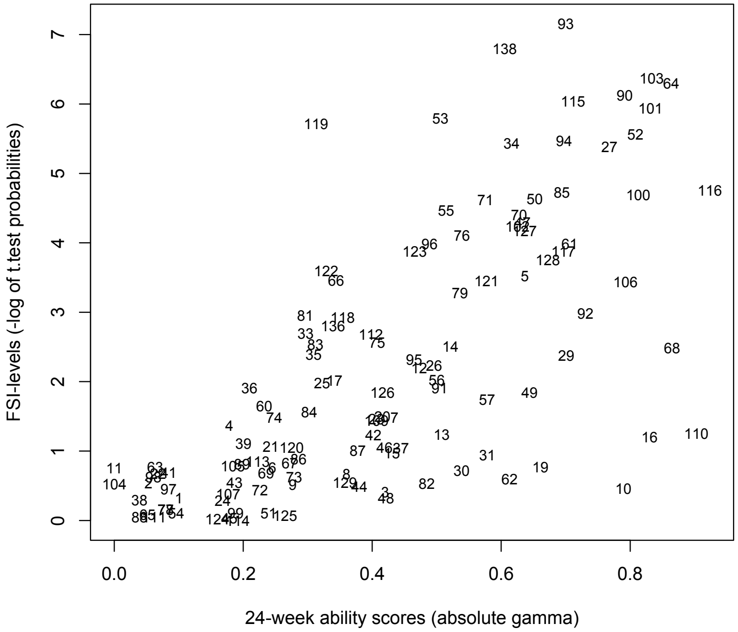

Figure 3. Difficulty assessment of individual WALS features (numbers refer to the chapters in WALS).

The parameters 93: "Position of interrogative phrases in content questions" (Dryer 2005g) and 116: "Polar questions" (Dryer 2005f) both relate to the structure of questions. Apparently, it is difficult for English speakers to learn a language in which the structure of questions is different from English. With regard to parameter 93, English consistently places the content interrogate (*who*, *what*, etc.) in the first position of the sentence, like most European languages. The most widespread other option used among the world's languages is the so-called "in-situ" interrogative, which appears in the same position in the sentence as the corresponding answer. Concerning parameter 116, English - like all Germanic languages - uses a special word order for polar questions (the so-called "inversion" construction, triggering *do*-support in English). This is a highly unusual construction from a world-wide perspective. Most languages use a special interrogative particle to formulate polar questions.

The parameters 100: "Alignment of verbal person marking" (Siewierska 2005a), 101: "Expression of pronominal subjects" (Dryer 2005a) and 103: "Third-person zero of verbal person marking" (Siewierska 2005b) all relate to the person cross-referencing as marked on the verb (so-called "person inflection", also often called "agreement"). Apparently it is difficult for English speakers to learn a language that uses a different kind of person inflection compared to English. Regarding parameter 100, English uses accusative alignment, i.e. the intransitive subject and the transitive subject trigger the same inflection. This is the most widespread strategy from a world-wide perspective. Other approaches, like ergative or active alignment, make a language more difficult to learn for English speakers. Concerning parameter 101, English needs obligatory pronouns in subject position. Most languages do not force such marking ("pro-drop"), again apparently making learning difficult for English speakers. Finally, with regard to parameter 103, English overtly marks a third person singular subject by verb inflection (with the suffix *-s*, though not in all tenses), whereas all other persons are unmarked. This is a highly idiosyncratic structure from a world-wide perspective. Languages with another distribution of zero person inflection are relatively more difficult for English learners. However, this difficulty actually conflates two phenomena. First, languages without any person inflection (i.e. all person marking is zero) are generally more difficult for English learners, but, likewise, are languages with person inflection for all persons, though zero-marked in the third person.

The parameters 52: "Comitatives and instrumentals" (Stolz, Stroh and Urdze 2005) and 64: "Nominal and verbal conjunction" (Haspelmath 2005b) both relate to the semantic distribution of linguistic structures. In both parameters, English - like all European languages - does not differentiate formally between the coding of two different semantic structures. Languages that do differentiate are apparently more difficult for English Learners. Concerning parameter 52, English uses the same construction for comitatives (*John went to the cinema* ***with Mary***) and instrumentals (*John fixed the lamp* ***with a screwdriver***), which is actually a minority pattern from a worldwide perspective. Similarly, English uses the same conjunction between noun phrases (***The lion and the monkey*** *eat bananas)* and verb phrases (*The lion* ***eats and sleeps***). Such an identity of conjunction structure is widespread, though roughly half of the world's languages would use different marking in these two situations.

The parameters 85: "Order of adposition and noun phrase" (Dryer 2005c), 90: "Order of relative clause and noun" (Dryer 2005e) and 94: "Order of adverbial subordinator and clause" (Dryer 2005d) all relate to the ordering of elements in the sentence. Ever since Greenberg's (1963) seminal paper on word order universals, there has been a strong interest in the interrelation between such parameters (cf. Dryer 1992 as a major reference). It is not completely clear why exactly these three parameters (and not any of the other word-order parameters) end up high on the scale of difficult to learn grammatical characteristics. It appears to depend on the rather limited set of languages in the current sample. Yet, it is clear that languages with different word-order characteristics from English are difficult to learn for English speakers.

There are various other parameters that appear to make languages difficult to learn for English speakers, for example the use of reduplication as a structural mechanism in the grammar of a language (parameter 27, (Rubino 2005)) and the fact that nouns with a plural meaning are not always obligatorily marked as such (parameter 34, (Haspelmath 2005c)). The remaining parameters high on the difficulty scale are less obvious to explain. Parameter 115: "Negative indefinite pronouns and predicate negation" (Haspelmath 2005a) described the difference whether negative indefinites like *nowhere, nobody* or *nothing* can co-occur with a further negation in the sentence. However, English is described as having "mixed behavior", so almost all the world's languages are different from English in this respect. Finally, parameter 138: "The word for *tea*" (Dahl 2005) classifies languages as to whether they use a word for *tea* derived from Sinitic *cha,* or from Min Nan Chinese *te*. Although it is slightly amusing that such a parameter appears to be correlated with learning difficulty, it simply seems to be an accidental side effect that will be ignored subsequently in this paper.

6. Predicting language learning difficulty

Given the numerous factors that strongly correlate with language learning difficulty for English speakers, it seems likely that we can reverse the approach and predict the difficulty of a language from these factors. Such a prediction might be useful to get an indication of expected difficulty for languages that are not included in the FSI data used here. Furthermore, statistical predictive models offer a more detailed indication of the relative importance of the various factors discussed in this paper. However, remember that the following predictive models are based on the very restricted difficulty-assessments as prepared by the FSI. For example, that data does not include any control for the individual background of the learners, but treats all English speakers as equal. Differently formulated, the current factors only deal with the target of the learning, while better models should also include factors relating to the background and personality of the learners. Also, the levels of difficulty distinguished are rather rough, and the number of languages available for the establishment of the models is rather limited. The models that will be proposed in this section should thus be interpreted with these limitations in mind.

The basic approach to find suitable predictive models is to include various factors into a linear regression model and try to find a model by reducing the number of factors, while optimizing the relative goodness of fit as measured by the Akaike information criterion (AIC).[4] By including many factors it will almost always be possible to produce well-fitting predictive models. However, the more factors included, the less clear the interpretation of such models becomes. It is thus more interesting to search for models with a limited number of factors that still predict the observed measurements to a reasonable degree.

Before turning to the concrete models, there is one further problem with the current data. The problem is that the values to be predicted (i.e. the values of language learning difficulty) are strongly biased towards mid values. In the FSI-levels, the largest group of languages is of level 4 (viz. 34 of 60 languages, i.e. more than half of the sample, cf. table 1), while for the 24-week ability scores, the largest group of languages has a score of 2.00 (viz. 15 of 42 languages, cf. table 2). To counterbalance this skewed distribution, I weighted all observations in the regression model by the inverse of the number of languages in the level. For example, the languages with FSI-level 4 were weighted as counting only 1/34.

Table 4. Predictive model of language-learning difficulty with continuous factors.

	Estimate	Std. Error	t-value	Pr(> |t|)
FSI-levels				
Intercept	8.98	0.68	13.21	< 2e-6 ***
Typology	– 7.25	1.50	– 4.82	1.34e-5 ***
Writing System	– 2.19	0.46	– 4.71	1.93e-5 ***
24-week ability				
Intercept	0.72	0.29	2.53	0.016 *
Typology	1.77	0.59	3.00	0.0049 **
Writing System	0.66	0.19	3.51	0.0013 **

The first kind of model to predict language-learning difficulty consists mainly of continuous factors. I included the following factors in the search for optimal models (the actual values used can be found in the Appendix A):

– Typological similarity, defined as a value between 1 (completely similar to English) and 0 (completely dissimilar from English), cf. Section 5;
– Geographical distance from London, defined as the great circle distance in kilometers;
– Orthographic similarity, defined as a value between 1 (completely similar to English) and 0 (completely dissimilar from English), cf. Section 4;
– Size of the orthographic system, defined as the number of Unicode graphemes used in the writing system;
– Genealogical similarity to English, defined as two binary parameters: first, whether the language belongs to the Germanic genus or not, and, second, whether the language belongs to the Indo-European family or not.

The optimal models only include the typological similarity and the orthographic similarity, as shown in table 4. For the FSI-levels, this model starts from an intercept of almost 9, which can be interpreted as saying that language learning is very difficult. Then, depending on the typological and orthographic similarity to English, the FSI-level is reduced. The typological similarity counts for a relative reduction of 7.25, while the orthographic similarity only results in a relative reduction of 2.19. Consider for example Norwegian, with an FSI-level of 1 (i.e. easy to learn for English speakers). Based on the typological similarity of Norwegian to English of 0.78 and a writing system similarity to English of 0.93, the linear regression in table 4 predicts an FSI-level of 1.29 (=8.98–7.25·0.78–2.19·0.93). For the 24-week ability scores, the model starts from an intercept of 0.72, which likewise represents maximum language learning difficulty. Typological similarity adds a fraction of 1.77, while orthographic similarity adds a fraction of 0.66 to this score. Again taking Norwegian as an example (24-week ability score of 3.00) the model predicts a score of 2.71 (=0.72+1.77·0.78+0.66·0.93). Only by including these two factors, a reasonable good prediction can be made of the learning difficulty of a language for English speakers.

Although the models in table 4 have a good predictive power, they are not very practical in actual usage. To predict language-learning difficulty with these models it is necessary to assess the complete typological similarity to English based on the WALS data. Furthermore, an extensive analysis of the writing system is necessary. To obtain simpler predictive models, I searched for optimal models using only binary factors, i.e. simple yes/no questions about the languages in question. A well-fitting model with only a few such simple questions could be of enormous practical value for predicting the difficulty English speakers might have when learning a foreign language. I included the following factors in the search for optimal models:

- Whether the language has a Latin script, or not;
- Whether the language is of the same genus as English (i.e. Germanic), or not;
- Whether the language is of the same family as English (i.e. Indo-European), or not;
- Whether the language has the same grammatical structure as English for any of the WALS parameters as discussed above with reference to figure 2, i.e. parameters 93, 116, 100, 101, 52, 64, 85, 90, 94, 27, and 34.6

To be able to search through all combinations of WALS parameters, a complete data table for the 11 parameters is necessary. Unfortunately, the data in WALS is highly incomplete, so I had to reduce the number of languages even more for this search. In the end, I decided on a set of 28 languages for which the parameters are almost completely available, and added the missing data points by choosing the parameter values most commonly attested in closely related languages and/or in the linguistic area in which the language is spoken (see Appendix B and C). The resulting predictive models (after optimizing for AIC, as above) are shown in table 5. With only four binary factors (as with the FSI-levels model) it is maximally possible to predict $2^4 = 16$ different levels of learning difficulty. With only three factors (as with the 24-week ability scores model) the number of possibly difficulty levels is even less, namely only $2^3 = 8$ levels. These models can thus not be very precise in their predictions. Statistically, it seems to be possible to reduce the number of factors even further, but I have decided to add more typological parameters as statistically necessary for an optimal model to get somewhat more different levels of prediction (which leads to non-significance of some of the parameters).

I will take Greek as an example for how these models predict language-learning difficulty. First, Greek has an FSI-level of 4, while the model in table 5 predicts a level of 3.41 (=6.46–0·1.72–1·1.68–1·0.84–1·0.53), based on the facts that Greek does not have a Latin script, but is Indo-European, has (predominantly) prepositions, and has accusative alignment. Second, Greek has a 24-week ability score of 1.75, while the model in table 5 predicts a score of 1.99 (=1.21+0·0.47+1·0.28 +1·0.50). The predications are equally accurate for all other languages investigated.

Table 5. Predictive model of language-learning difficulty with only binary factors.

	Estimate	Std. Error	t-value	Pr(> |t|)
FSI-levels				
Intercept	6.46	0.30	21.68	< 2e-6 ***
Latin script	– 1.72	0.43	– 4.00	5.70e-4 ***
Indo-European	– 1.68	0.40	– 4.21	3.32e-4 ***
Prepositions (85)	– 0.84	0.41	– 2.04	0.052
Accusative (100)	– 0.53	0.47	– 1.13	0.27
24-week ability				
Intercept	1.21	0.12	9.83	2.63e-9 ***
Latin script	0.47	0.15	3.18	0.0045 **
Prepositions (85)	0.28	0.16	1.71	0.10
Nominal plural (34)	0.50	0.15	3.27	0.0037 **

7. Conclusion

Language learning becomes more difficult the more different the language to be learned is from the learner's native tongue. There are many different ways in which differences between languages can be quantified, and this paper has investigated a few possibilities. It turns out that, indeed, larger differences between languages are correlated with larger difficulty, though not all differences are equally important. For English native speakers it appears to be particularly difficult to learn a language that does not have a Latin script, is non-Indo-European, has postpositions, is ergatively aligned and does not have obligatory nominal plural. The fact that such differences make a language difficult to learn is not very surprising. The more interesting result of this paper is, first, exactly which factors are the strongest predictors amongst the many possible factors quantifying similarity between languages, and, second, the detailed quantitative predictions of language learning difficulty based on such few characteristics of the language to be learned.

Appendix A: Complete data for continuous factors

WALS code	Language name	FSI-level	24-week score	Germanic	Indo-Europ.	Geograph. distance	Typology similarity	Script similarity	Script invent.
afr	Afrikaans	I	3.00	+	+	9469	NA	0.912	53
dsh	Danish	I	2.25	+	+	789	0.848	0.929	60
dut	Dutch	I	2.75	+	+	412	0.750	0.915	59
fre	French	I	2.50	–	+	467	0.662	0.954	60
ita	Italian	I	2.50	–	+	1343	0.696	0.947	60
nor	Norwegian	I	3.00	+	+	1112	0.782	0.935	58
por	Portuguese	I	2.50	–	+	1571	0.638	0.929	62
rom	Romanian	I	3.00	–	+	1929	0.657	0.920	61
spa	Spanish	I	2.25	–	+	1368	0.615	0.947	60
swe	Swedish	I	3.00	+	+	1283	0.862	0.934	63
ger	German	I*	2.25	+	+	684	0.698	0.916	70
ind	Indonesian	I**	2.00	–	–	11086	0.481	0.766	52
mly	Malay	I**	2.75	–	–	10553	NA	0.750	58
swa	Swahili	I**	2.75	–	–	7476	0.463	0.687	62
alb	Albanian	II	NA	–	+	1947	0.590	0.858	62
amh	Amharic	II	2.00	–	–	5787	0.446	0.000	164
arm	Armenian	II	NA	–	+	3651	0.494	0.000	83
aze	Azerbaijani	II	NA	–	–	3858	0.415	0.668	70
ben	Bengali	II	1.75	–	+	7924	0.439	0.000	383
bos	Bosnian	II	NA	–	+	1674	NA	0.850	62
bul	Bulgarian	II	2.00	–	+	2145	0.620	0.000	68
brm	Burmese	II	1.75	–	–	8573	0.373	0.000	357
cze	Czech	II	2.00	–	+	1070	0.630	0.849	78
grk	Greek	II	1.75	–	+	2225	0.581	0.000	67
heb	Hebrew	II	2.00	–	–	3620	0.562	0.000	31
hin	Hindi	II	1.75	–	+	6968	0.556	0.000	390
ice	Icelandic	II	NA	+	+	1737	0.689	0.855	61
khm	Khmer	II	2.00	–	–	9905	0.402	0.000	336
lao	Lao	II	1.50	–	–	9288	0.432	0.000	268
lat	Latvian	II	NA	–	+	1635	0.580	0.816	69
lit	Lithuanian	II	NA	–	+	1612	0.565	0.821	72
mcd	Macedonian	II	NA	–	+	2000	0.692	0.000	68
nep	Nepali	II	1.75	–	+	7260	0.426	0.000	352
psh	Pashto	II	NA	–	+	5654	0.429	0.000	34

WALS code	Language name	FSI-level	24-week score	Ger-manic	Indo-Europ.	Geograph. distance	Typology similarity	Script similarity	Script invent.
prs	Persian	II	2.00	–	+	4842	0.430	0.000	48
pol	Polish	II	2.00	–	+	1364	0.605	0.841	70
rus	Russian	II	2.25	–	+	2488	0.652	0.000	66
scr	SerboCroatian	II	2.00	–	+	1662	0.608	0.000	64
snh	Sinhalese	II	1.75	–	+	8739	0.523	NA	NA
svk	Slovak	II	NA	–	+	1447	NA	0.870	81
slo	Slovenian	II	NA	–	+	1277	0.677	0.885	51
tag	Tagalog	II	2.00	–	–	10653	0.387	0.607	56
tur	Turkish	II	2.00	–	–	3044	0.437	0.861	56
ukr	Ukrainian	II	NA	–	+	2336	0.659	0.000	73
urd	Urdu	II	NA	–	+	6285	0.500	0.000	30
uzb	Uzbek	II	NA	–	–	5148	0.442	0.000	75
xho	Xhosa	II	NA	–	–	9697	0.360	0.808	61
zul	Zulu	II	NA	–	–	9569	0.397	0.812	60
tha	Thai	II*	2.00	–	–	9336	0.455	0.000	249
hun	Hungarian	II*	2.00	–	–	1540	0.511	0.851	56
est	Estonian	II*	NA	–	–	1796	0.583	0.862	52
fin	Finnish	II*	2.00	–	–	1858	0.593	0.863	56
geo	Georgian	II*	NA	–	–	3454	0.400	0.000	46
vie	Vietnamese	II*	1.50	–	–	10181	0.423	0.649	117
kha	Mongolian	II*	2.00	–	–	6903	0.432	0.000	54
aeg	Arabic	III	1.50	–	–	3520	0.504	0.000	60
cnt	Cantonese	III	1.25	–	–	9450	0.447	NA	NA
kor	Korean	III	1.00	–	–	8855	0.453	0.000	64
mnd	Mandarin	III	1.50	–	–	8286	0.462	0.001	532
jpn	Japanese	III*	1.00	–	–	9379	0.385	0.000	505

Appendix B: Data added to WALS

Language	WALS	Feature	Value	Notes
Burmese	brm	34	1	common in South-East Asia
Burmese	brm	52	2	common in South-East Asia
Burmese	brm	64	1	common in South-East Asia
Dutch	dut	27	3	same as all of Europe
Dutch	dut	52	1	same as all of Europe
Dutch	dut	64	1	same as all of Europe
Dutch	dut	93	1	same as all of Europe
Georgian	geo	116	6	
Georgian	geo	90	1	
Hindi	hin	101	2	same as most Indic
Italian	ita	93	1	same as all of Europe
Khalka	kha	115	1	typical Eurasian
Khalka	kha	64	2	same as Mangghuer

Language	WALS	Feature	Value	Notes
Khmer	khm	101	5	common in South-East Asia
Khmer	khm	64	1	common in South-East Asia
Korean	kor	34	4	
Latvian	lat	27	3	same as most of Europe
Mandarin	mnd	94	5	same as Cantonese
Persian	prs	34	6	same as all Iranian and European
Spanish	spa	52	1	same as all of Europe
Swahili	swa	64	2	same as most Bantu
Tagalog	tag	101	1	
Tagalog	tag	85	2	same as all Austronesian
Thai	tha	34	1	common in South-East Asia
Vietnamese	vie	52	2	common in South-East Asia
Zulu	zul	52	2	typical Bantu

Appendix C: Complete data for WALS parameters

WALS	100	101	103	93	116	52	64	115	34	27	90	94	85
dut	2	1	2	1	4	1	1	2	6	3	1	1	2
fre	2	1	2	1	1	1	1	3	6	3	1	1	2
ita	2	2	2	1	6	1	1	3	6	3	1	1	2
spa	2	2	2	1	4	1	1	3	6	3	1	1	2
ger	2	1	2	1	4	1	1	2	6	3	1	1	2
ind	2	1	2	3	1	3	1	1	4	2	1	1	2
swa	2	2	2	2	1	2	2	1	6	1	1	1	2
brm	1	5	1	2	2	2	1	1	1	2	2	4	1
grk	2	2	2	1	1	1	1	1	6	3	1	1	2
heb	2	6	4	1	1	1	1	1	6	1	1	1	2
hin	2	2	2	2	1	2	1	1	6	1	4	1	1
khm	1	5	1	2	1	2	1	1	4	1	1	1	2
lat	2	1	4	1	1	1	1	1	6	3	1	1	2
prs	2	2	4	2	1	1	1	1	6	1	1	1	2
rus	2	1	2	1	1	3	1	1	6	3	1	1	2
tag	1	1	1	1	1	2	1	1	4	1	1	1	2
tur	2	2	4	2	1	1	1	1	6	1	2	5	1
zul	2	2	2	2	1	2	2	1	6	2	1	1	2
tha	1	5	1	2	1	2	1	1	1	1	1	1	2
hun	2	2	2	2	1	3	1	1	6	1	7	1	1
fin	2	6	4	1	1	2	1	1	6	3	1	1	1
geo	2	4	2	2	6	3	1	3	6	1	1	1	1
vie	1	5	1	2	1	2	1	1	4	1	1	1	2
kha	1	5	1	2	1	2	2	1	4	1	2	2	1
aeg	2	2	4	2	1	2	1	1	6	1	1	1	2
kor	1	5	1	2	2	2	2	1	4	1	2	2	1
mnd	1	5	1	2	1	2	2	1	2	1	2	5	2
jpn	1	5	1	2	1	2	2	1	2	2	2	2	1

Notes

[1] I thank the editors and the anonymous reviewers for their assistance with the preparation of the current paper. This research was supported by ERC Starting Grant 240816 “Quanthistling”.

[2] The following FSI website contains the assessments of language difficulty: http://www.nvtc.gov/lotw/months/november/learningExpectations.html. This website is available through the internet archive at http://www.archive.org/. A website at WikiBooks claims to have information about the FSI-levels for an even larger number of languages, but I have not been able to trace the origin of these additional assessments, so I have not used them for this paper: http://en.wikibooks.org/wiki/Language_Learning_Difficulty_for_English_Speakers (all pages accessed on 22 March 2011).

[3] Available online at http://unicode.org/udhr/.

[4] For some reason, the Unicode standard treats the Hangul script of Korean differently, as the syllabic combinations are not treated as combining. This results in Korean being treated here rather differently from Devanagari, while the difference in the script structure is not that profound.

[5] In practice, I used the implementation *step* as available in the statistical environment R (R Development Core Team 2010) for this optimization.

[6] The parameters 103 and 115 are not included, because their similarity/difference to English is difficult to interpret. Also parameter 138 is not included because of lack of relevance (cf. the discussion in Section 5).

References

Albu, Mihai
2006 Quantitative analyses of typological data. Ph.D. Thesis, University of Leipzig.

Chiswick, B. R. and P. W. Miller
2005 Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development* 26(1). 1-11.

Coppin, Ben
2008 Automatic language similarity comparison using n-gram analysis. Master Thesis, Queens’ College: Cambridge.

Cysouw, Michael

2012 Disentangling geography from genealogy. In: Benedikt Szmrecsanyi (ed.), *Space in language and linguistics: geographical, interactional, and cognitive perspectives,* Walter de Gruyter: Berlin.

Dahl, Östen

2005 Tea. In: Martin Haspelmath et al. (eds.), 554-557.

Damashek, M.

1995 Gauging similarity with n-grams: Language-independent categorization of text. *Science* 267(5199). 843.

Dryer, Matthew S.

1992 The Greenbergian word order correlations. *Language* 68(1). 80-138.

2005a Expression of pronominal subjects. In: Martin Haspelmath et al. (eds.), 410-413.

2005b Genealogical language list. In: Martin Haspelmath et al. (eds.), 582-642.

2005c Order of adposition and noun phrase. In: Martin Haspelmath et al. (eds.), 346-349.

2005d Order of adverbial subordinator and clause. In: Martin Haspelmath et al. (eds.), 382-385.

2005e Order of relative clause and noun. In: Martin Haspelmath et al. (eds.), 366-369.

2005f Polar questions. In: Martin Haspelmath et al. (eds.), 470-473.

2005g Position of interrogative phrases in content questions. In: Martin Haspelmath et al. (eds.), 378-381.

Hart-Gonzalez, Lucinda and Stephanie Lindemann

1993 Expected achievement in speaking proficiency. School of Language Studies, Foreign Services Institute, Department of State: Washington DC.

Greenberg, Joseph H.

1963 Some universals of grammar with particular reference to the order of meaningful elements. In: Joseph H. Greenberg (ed.), *Universals of Language,* 73-113. MIT Press: Cambridge, Mass.

Haspelmath, Martin

2005a Negative indefinite pronouns and predicate negation. In: Martin Haspelmath et al. (eds.), 466-469.

2005b Nominal and verbal conjunction. In: Martin Haspelmath et al. (eds.), 262-265.

2005c Occurrence of nominal plurality. In: Martin Haspelmath et al. (eds.), 142-145.

Haspelmath, Martin, Matthew S. Dryer, Bernard Comrie and David Gil (eds.)
2005 *The World Atlas of Language Structures,* Oxford University Press: Oxford.

Huffman, Stephen M.
2003 The genetic classification of languages by n-gram analysis. PhD Thesis, Georgetown University: Washington, D.C.

Jackson, F. H. and M. A. Kaplan
2001 Lessons learned from fifty years of theory and practice in government language teaching. In: James A. Alatis and Ai-Hui Tan (eds.), *Language in Our Time,* 71-87. Georgetown UP: Washington, D.C.

Kellerman, Eric
1979 Transfer and Non-Transfer: Where We Are Now. *Studies in Second Language Acquisition* 2(1). 37-57.

R Development Core Team
2010 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria.

Rubino, Carl
2005 Reduplication. In: Martin Haspelmath et al. (eds.), 114-117.

Schepens, Job, Frans van der Slik, Roeland van Hout
this vol. The Effect of Linguistic Distance across Indo-European Mother Tongues on Learning Dutch as a Second Language.

Siewierska, Anna
2005a Alignment of verbal person marking. In: Martin Haspelmath et al. (eds.), 406-409.
2005b Third-person zero of verbal person marking. In: Martin Haspelmath et al. (eds.), 418-421.

Stolz, Thomas, Cornelia Stroh and Aina Urdze.
2005 Comitatives and Intrumentals. In: Martin Haspelmath et al. (eds.), 214-217.