Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

Discussion Papers

921

Olaf J. de Groot

Measuring Ethno-Linguistic Affinity between Nations

Berlin, September 2009

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the institute.

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN print edition 1433-0210
ISSN electronic edition 1619-4535

Available for free downloading from the DIW Berlin website.

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN.
Papers can be downloaded free of charge from the following websites:

http://www.diw.de/english/products/publications/discussion_papers/27539.html
http://ideas.repec.org/s/diw/diwwpp.html
http://papers.ssrn.com/sol3/JELJOUR_Results.cfm?form_name=journalbrowse&journal_id=1079991

Measuring Ethno-Linguistic Affinity Between Nations*

Olaf J. de Groot†
DIW Berlin, Department of International Economics

This version: June 2009

Abstract

Research on ethno-linguistic ties has so far mostly focused on domestic measures of ethno-linguistic heterogeneity. Little attention has been given to the possibility that ethno-linguistic relations between countries may affect outcomes, particularly in a spatial econometric context. In this paper, I propose a way of measuring Ethno-Linguistic Affinity between nations. This new index measures the degree of similarity two randomly drawn individuals from two different populations can be expected to display. I show that this measure has a number of attractive theoretical characteristics, which make it particularly useful and continue to actually construct such a measure for all countries in Africa. Finally, using this measure of Ethno-Linguistic Affinity, I show that civil conflict in Africa is likely to spill over between contiguous ethno-linguistically similar countries.

Keywords: Ethno-Linguistic Heterogeneity; Spatial Econometrics; Conflict; Africa

JEL code: C21, F51, O10

*I am grateful to Guido Tabellini, Eliana La Ferrara, Kristian Skrede Gleditsch, Idil Göksel, Jochen Mierau and the participants of the *Graduate Students Seminar* (February 4, 2008) at Bocconi University, the *4th IUE International Student Conference* (April 14, 2008) in Izmir, Turkey, the *2e Conférence Euro-Africaine en Finance et Economie* (June 5-6, 2008) in Tunis, Tunisia, the *12th Annual International Conference on Economics and Security* (June 11-13, 2008) in Ankara, Turkey and the *Jan Tinbergen European Peace Science Conference* (June 30-July 2, 2008) in Amsterdam, the Netherlands for their helpful comments. All views expressed here are my own.

†Address of the author: Olaf J. de Groot, DIW Berlin, Mohrenstrasse 58, 10117 Berlin, Germany.

1 Introduction

Measures of Ethno-Linguistic Fractionalisation (ELF) and other types of ethno-linguistic heterogeneity, have been around for quite some time now. The most well-known contribution in this field is of course by Easterly and Levine (1997), who analyse the relationship between ethnic diversity and a range of economic indicators. They argue that a high level of ethno-linguistic fractionalisation may lead to strong rent-seeking and other growth-retarding policies. Easterly and Levine, however, like others publishing in this field, only concern themselves with the ethno-linguistic relationships within countries. In my opinion, on the other hand, ethno-linguistic (dis)similarities may also go a long way in explaining relationships between countries. In this paper, I propose a simple measure that should be able to improve many spatial macroeconometric analyses by including an ethno-linguistic component in addition to the common spatial parameter. After all, the original premise in spatial econometrics is that influence is exerted over space and that this influence reduces when the physical distance between observations becomes larger[1]. I agree with this statement, but I also believe that one should not merely consider physical distance, but include an ethno-linguistic component as well, when performing any kind of spatial macroeconometric analysis. In the current paper, after introducing this measure and showing how it can be constructed, I apply it in the field of conflict spillovers in Africa. Civil conflict is one of a number of events that is regularly analysed with a spatial component, but has so far received little attention in an ethno-linguistic spillover framework, except for Alesina *et al.* (2006), who look at non-natural borders as a proxy for states that may share one ethnic group.

In the following section of the paper, I give some more background information on ethno-linguistic fractionalization, international ethno-linguistic relations and the spatial econometric literature on conflict spillovers. In the third section, I introduce my own measure, describe how it is set up and show its most important characteristics. In the fourth section, I show my empirical application in the field of spillovers of conflict and the final section concludes.

2 Related Literature

2.1 Ethno-Linguistic Fractionalisation Indices

As mentioned earlier, Easterly and Levine (1997) are the most well-known early adopters of a measure of Ethno-Linguistic Fractionalisation. Their measure of ELF measures the probability of two randomly drawn people from a population being part

[1] This is the spatial version of saying that events from the recent past have a stronger influence than similar events from a more distant past, as one could say in time series analysis.

of two different ethnic groups: $1 - \sum_{j=1}^{J} \alpha_j^2$, where $j \in J$ are all different ethnic groups in a society and α_j is the share of population of group j. The data used in their analysis comes from the Soviet *Atlas Narodov Mira* from 1964, which has long been considered to be the most precise data available at the most disaggregated level. A significant disadvantage of this source, however, is the fact that it focuses strongly on linguistic differences and does not in fact take ethnic differences into consideration. As a result, the famous example of Rwanda is awarded an ELF of 0.14[2], as Hutus and Tutsis speak the same language and are thus considered to be one single group. Of course, history has taught us the flaw in that assumption when up to 1 million were killed in the mid-nineties in ethnic violence between the Hutu majority and the Tutsi minority.

Easterly and Levine (1997) may be among the most famous for using a measure of ELF, but they were neither the first, nor the last to do so. Mauro (1995) is generally credited with being the first to introduce the measure to wider scientific interest in the field of economics, although it had been around for quite some time already. He uses the same ELF as Easterly and Levine do two years later, as an instrument for corruption to analyse its effect on growth in a cross-section of countries.

After Easterly and Levine, others have introduced alternative measures to deal with some of the drawbacks of the simple ELF used by earlier authors. Alesina *et al.* (2003) introduce several new measures for ethnic, linguistic and religious fractionalisation in a large set of countries. The first point they wanted to address is the problem that standard ELF focuses too strongly on linguistic groups, whereas these may not be the ties that distinguish groups the most from each other. They therefore calculate separate indices for religion, language and finally ethnicity. For the ethnicity data, the authors use an interesting approach, changing the definition of ethnicity per country. For example, they make racial distinctions in most of Latin America, whereas linguistic differences are used in Europe. The final important difference between the measures of Alesina *et al.* and previous measures of ELF is the source of the data. For their work, Alesina *et al.* use recent data, instead of the 1960s Atlas Narodov Mira. This has the advantage that the data is more detailed, more precise and possibly more trustworthy. It does, however, bring up the issue of endogeneity, where it can no longer be guaranteed that the ethno-linguistic composition of countries is independent of the outcomes that they are trying to measure (output, particularly). However, it is argued that such endogenous changes are extremely rare and that all measures of fractionalisation show very strong persistence over time.

Another alternative measure for ELF is discussed by Laitin (2000), Fearon (2003) and others, who take a more linguistic approach, based on Greenberg (1956). The use of distance in a tree diagram of languages reflects the expectation that languages that

[2]According to Appendix 3 of Mauro (1995), Rwanda has a level of ethno-linguistic fractionalization of 0.14, which is very low compared to their neighbours Zaire/DRC (0.90), Uganda (0.90) and Tanzania (0.93), but similar to Burundi (0.04).

have branched out from each other more recently are expected to be more culturally similar. While the signal is recognised by Fearon to be noisy, he argues that the proposed fractionalisation measure is informative nonetheless: $1 - \sum_{i=1}^{I}\sum_{j=1}^{J}\pi_i\pi_j r_{ij}$, in which π_i and π_j are population shares and r_{ij} is the so-called resemblance factor proposed by Greenberg. $r_{ij} \in [0,1]$ measures how much two ethnic groups resemble each other, but the specification of this factor is still inconclusive. Greenberg, for example, wants to use the proportion of resemblances between each pair of languages on the most recent version of the glottochronology list. Unfortunately, the science of glottochronology has since lost most of its credibility and is no longer practiced on a large scale. Fearon, instead, proposes $r_{ij} = \left(\frac{l}{m}\right)^{\alpha}$, where l is the number of linguistic levels shared between language i and j, m is the largest number of linguistic levels recorded in the dataset and $\alpha \in [0,1]$. Laitin, finally, counts the number of branchings that are shared between languages according to the *Ethnologue* (Gordon, 2005) dataset.

Bossert *et al.* (2008) take all of the aforementioned indices of fractionalisation and combine them in order to create a Generalised Index of Fractionalisation (GELF). For this they use Census results that provide characteristics of the individual members of the population and they continue to calculate the level of fractionalisation. As shown in their contribution, this approach actually includes most measures of ELF and also possesses a number of desirable characteristics.

A completely different approach to Ethnic Fractionalisation that should be discussed here is proposed by Posner (2004) and only looks at politically relevant groups. He refers to his index as PREG (Politically Relevant Ethnic Groups) and it consists of a standard ELF, but instead of using all subgroups as is done in the traditional ELF literature, he takes only the politically relevant groups for each country, which are normalised to 100%. So, for example in Kenya, instead of using the 21 groups named in the *Atlas Narodov Mira* or the 64 groups mentioned by Gordon (2005), only the population sizes of the politically relevant Luo, Kalenjin and Kikuyu are used. Of course, while this may be an interesting measure, it endogenises the problem of having to decide which groups are the politically relevant ones in a country. Posner seems to have done this very carefully, but it leaves considerable room for criticism. In the final part of his paper, he replicates the results of Easterly and Levine and confirms that ethnic fractionalisation has a significant and strong negative impact on economic growth in his selection of African nations. A final problem with Posner's approach is that it does not cover issues that are caused by the difference between included and excluded groups. That is, certain political events that one is trying to research can be caused precisely by the fact that certain groups that are significant portions of the population are excluded from the political process.

Two further papers that use alternative approaches are Fearon and Laitin (2003) and Michalopoulos (2007). The first combine an array of quasi-standard ELF indices, with some unconventional measures, such as the number of distinct languages spoken

by at least 1% of the population. Of course, this latter measure has typical problems of how to define distinct languages, and what makes a person to be a speaker of a particular language. Michalopoulos is worth mentioning because his results go in the opposite direction of the conventional wisdom. He argues that ethnic heterogeneity is the result of geographic heterogeneity and that, therefore, a measure of geographic heterogeneity can be used as an Instrumental Variable to analyse the influence between ELF and economic outcomes. According to his results, the strong effect found by previous authors was a spurious relationship and there is no real correlation between ELF and development.

Finally, many of the more recent authors[3] argue in favour of collecting more modern data. Often, the central thesis is that the potential endogeneity bias is of less concern than the actual problems with the data from the *Atlas Narodov Mira*. The specific data problems are manifold, but particularly the groupings, the definition of ethnic groups and the actual measurement of different group sizes have all been called into question. More recent data sources, on the other hand, are able to provide data that is more accurate and less biased in favour of any of the participating groups. Another advantage is the possibility of collecting the same data from different sources, comparing them and being able to arrive at a more precise estimate than when using a single source from 1964.

2.2 International Component of Ethno-Linguistic ties

One thing that has received little attention in the literature so far, is the international component of ethno-linguistic ties. The relationship between nations has been subject of many studies, but in nearly every case pure geographical proximity is used as variable of interest[4]. For some particular purposes, such a contiguity-measuring variable may occassionally be augmented with a measure of economic interrelatedness, such as the size of total trade between nation dyads. However, strong arguments can be made in favour of using a measure of ethno-linguistic proximity, as augmentation of simple geographic proximity.

One can think of many instances, in which such ethno-linguistic ties exacerbate existing geographic connections. From conflict literature, there is the example of the conflict in Rwanda, which spilled over into Burundi, due to their shared ethno-linguistic ties, whereas Uganda and Tanzania, both also bordering on Rwanda, were spared. Another example is the relatively large size of trade between Austria and

[3]These include most of the aforementioned authors, as well as Roeder (2001), who compares older results with more modern data and also proposes different definitions of ethnic groups to come up with an array of indicators for ethno-linguistic heterogeneity. Annett (2001) uses data from the *World Christian Encyclopedia* to derive more precise estimates for both ethnic and religious levels of fractionalisation.

[4]Among those studies are Sambanis (2002), Murdoch and Sandler (2004) and Abreu *et al.* (2004), as well as Ward and Gleditsch (2002) and Gleditsch (2007), which are discussed in the following subsection.

Germany. This is not merely explained by the fact that these countries are contiguous, but the historical and ethno-linguistic ties between them explain why, *ceteris paribus*, people from these countries may have a preference for each other over their other neighbours. The existence of such cross-border effects, particularly in Africa, where borders have been randomly drawn by European colonisers, should not come as a surprise. It is possibly more surprising that no comprehensive measure exists to cover this issue.

One paper that does address the ethno-linguistic aspects of conflict spill-overs is by Buhaug and Gleditsch (2008), who research whether conflicts indeed spill over or whether they actually cluster in space due to the clustering of other factors that explain conflict. They conclude that transnational ethnic links are indeed a key element in conflict clustering. Unfortunately, these authors do not provide a thorough description of their measure of *ethnic linkages*, but it is one of only few studies that consider the ethno-linguistic ties in the field of conflict spillovers.

Another paper that includes an international dimension is by Alesina *et al.* (2006), who introduce an innovative way of measuring how artificial international borders are. They do this with a so-called fractal measure, according to the following procedure. For the border of a particular country, a grid is laid out on the border and the number of boxes within the grid that cross the border are counted. Afterwards, the grid size is increased and a new count is made. When this procedure has been repeated several times, the authors have a dataset containing box-sizes and box-counts for a particular border and when one then regresses the natural logarithms of these on each other as follows, coefficient β will give an indicator of artificialness of a border:

$$\ln(box\_count) = \alpha + \beta \cdot \ln(box\_size)$$

The authors continue to combine this fractal measure with an indicator of "partitioned groups", defined as the percentage of the population of a country that belongs to a "partitioned group", where a partitioned group is defined as a group that appears in one or more adjacent countries as well[5]. Using these results, the authors then focus on using their measures as explanatory variables for economic and political success and accomplish satisfactory results. Unfortunately, Alesina *et al.* do not actually apply any of their measures on relations between countries, although they do mention that, with additional work, research into international conflict could come from their line of research.

A final approach that is becoming more prominent recently is the use of genetic distances. First popularised by Cavalli Sforza *et al.* (1994), this way of measuring ethnic distances is used by a number of authors (e.g. Spolaore and Wacziarg, 2006, and Guiso *et al.*, 2007). Spolaore and Wacziarg carefully explain how they apply the genetic distance data in their paper. The measure for genetic distance supposedly

[5]One issue Alesina *et al.* (2006) do not deal with carefully is group definitions. This could have a strongly distortive effect on their measure of "group partitioning" and should be discussed thoroughly.

measures how differently distributed the alleles of different populations are. Given that a population has a particular genetic profile, with particular distributions of the relevant alleles, it is possible to compare populations with each other and comment on their genetic level of similarity. The more different these allele distributions are, the further away these groups are from a common ancestor. The actual index is constructed as follows:

$$F = 1 - \frac{p_a q_a + p_b q_b}{2\overline{pq}}$$

where p, q are frequencies of different alleles in populations a and b and $2\overline{pq} = 1 - \left[\left(\frac{p_a+p_b}{2}\right)^2 + \left(\frac{q_a+q_b}{2}\right)^2\right]$ [6].

While this is an interesting measure, with a strong scientific basis, it faces two problems. Firstly, populations may be genetically similar despite belonging to different ethno-cultural groups, particularly when considering cultural features that have not yet existed for a long time (such as religion). The second issue is the fact that such data is only available at a highly aggregated level. In total, only 42 population groups are available, which are supposed to capture all of Earth's population. It would seem reasonable to say that ethnic competition or cooperation takes place at a more disaggregated level.

2.3 Causes of Civil Conflict

For now, it seems the debate on the causes of civil conflict has converged on the greed versus grievance theory of Collier and Hoeffler (2004), who argue that there can be two main sources for civil conflict. The grievance caused by an atypically unfair distribution of wealth or power or another kind or repression of a significant minority, appears to have relatively little explanatory power, whereas the greed explanation of opportunity for rebellion has much stronger results in regressions regarding the occurrence of civil conflict. However, not everyone agrees and there are still authors who continue to argue that an ethnic component may play an important role in the occurrence of conflict. Whether this is purely grievance-based remains a question, because there may be greed-based explanations related to ethnic division as well. More particularly, Fearon and Laitin (2003) investigate the greed versus grievance issue in their own way and they conclude that while grievance may be a small source, it is mostly economical reasons that cause civil war and not ethnic heterogeneity, in any of the ways they measure it. However, while Fearon and Laitin look at several aspects of ethno-linguistic heterogeneity, they do not approach one source of conflict in ethnic relations. This concerns the conflict between insiders and outsiders, which

[6] Actually, the authors claim that $2\overline{pq} = 1 - \left(\frac{p_a+p_b}{2}\right)^2 + \left(\frac{q_a+q_b}{2}\right)^2$, which must clearly be a typo, even though they repeat the same mistake in equations 18, 19 and 22. The final results, however, are consistent.

is an issue that relies both on heterogeneity, but also on the way heterogeneity is represented in the political system. The untested theory in this case is whether countries of which the governments are less representative of the ethno-linguistic heterogeneity, are more likely to suffer conflict.

Montalvo and Reynal-Querol (2005) put forward another strong critique of previous analyses of the relationship between ethnic (or religious) heterogeneity and conflict. They convincingly argue that it is not ethno-linguistic or religious fractionalisation that matters for the probability of conflict, but polarisation. In their excellent contribution, the authors show that their proposed index measures polarisation properly and is also compatible with a discrete version of the generalised polarisation index proposed by Esteban and Ray (1994):

$$EP = 4\sum_{j=1}^{J}\pi_j^2\left(1-\pi_j\right)$$

In their paper, Montalvo and Reynal-Querol show the excellent properties of this simple index and then continue to show that, contrary to previous results, ethnic polarisation is relevant in predicting civil conflict. Economic variables remain important, but ethnic polarisation (and to some degree: religious polarisation) plays an important role as well.

Of course, there are many other variables that are generally used to analyse conflict probabilities as well. These include the percentage of rough terrain, population density, population size, democratic freedoms and dependence on primary exports (particularly oil). I return to this list in section 4, where I run my own regressions regarding civil conflict.

One last feature of civil conflict that is relevant in the context of this paper, however, is the existence of spill-over theories. Some papers have included different kinds of geographical features in their conflict analyses, including a few that have followed the same line of reasoning that I apply in section 4, in that (civil) conflict is more or less likely to spill over from one country to another. Among the most relevant references in this case are Ward and Gleditsch (2002) and Gleditsch (2007). In their excellent contribution, Ward and Gleditsch (2002) use Markov Chain Monte Carlo estimations to show that conflict spillovers occur and they are able to correctly forecast a significant number of conflicts.

There are several reasons why international ethno-linguistic linkages may affect the initiation of conflict. These include a revisitation of the grievance literature, which in an international context would argue that when majority group A slaughters minority group B in country 1, majority group B in country 2 may want to take revenge on minority A. Additionally, a conflict spillover could also be the direct result of the export of combatants, if refugees in a neighbouring country choose to continue their battles there. Another potential source of spillover is found when neighbouring populations are inspired by a conflict. For example, with 2 neighbouring countries

in which majority group A represses minority group B, a succesful uprising by B in country 1 may lead the people in country 2 to be equally inspired to rise against their oppressors. A final theory on how ethno-linguistic linkages could play a role in conflict is related to precariously stable nations. Imagine a country consisting of two fairly balanced ethnic groups who share power in a reasonable way. Now, due to neighbouring conflict, one of the groups becomes more dominant, which can cause them to use the opportunity to repress the other group, which can lead to conflict.

Gleditsch (2007) argues that it is surprising how underlit the transnational dimensions of civil conflict actually are. He suggests there are several ways through which transnational links may influence civil conflict outcomes. Shared ethnic links, spillovers of autocratic tendencies and economic ties. When regressing both domestic and transnational features on conflict, and using Maximum Pseudo-Likelihood (MPL) techniques to approximate the likelihood function, he concludes that his measure of ethnic linkages is indeed significant.

3 A Measure of Ethno-Linguistic Affinity

In this paper, I propose to construct an index that I shall refer to as a measure of Ethno-Linguistic Affinity (ELA) between nations. Such an index can be used to augment existing distance measures to include both geographic and ethno-linguistic ties, which is important in order to achieve more accurate results regarding the existence of spill-over effects. In my opinion, there is a strong argument to be made in favour of saying that ethno-linguistic ties between nations are a factor that could strongly improve such research. Of course, there are natural phenomena which are driven purely by geographical proximity. Spatial correlations may be found due to similar weather patterns (consider Miguel *et al.*, 2004, for example) or due to regional resource abundance (consider the Middle East, for example). In other cases, neighbouring states may be influenced due to direct spillovers. In this context, think of the increased economic growth in Northern Ireland resulting from increased demand from the Republic of Ireland during the 1990s. Another example could be the recent oil-driven boom in Russia. Culturally close Belarus has benefitted from this economic improvement, while culturally distant Mongolia, despite its long Russian border, has not.

Obviously, it could be that economic ties are a determining feature in the relationship between countries. However, this is not necessarily the case. While in the previous examples economic ties are a relevant channel, for spillovers of anything else than growth (e.g. conflict), other channels may play a role too. But even for purely economic spill-overs, there could be other channels than simply trade. Ethno-linguistic ties could also play a role in how strongly one country responds to events in their neighbouring countries. Finally, a significant problem with using economic ties to analyse spillovers is the endogeneity of the measure. After all, when trying to

analyse when e.g. economic growth in one country spills over into another country, the use of economic channels confuses the analysis, as the existence of the channel may be the source of growth in the first place. It is also important to remember that such an analysis would not answer the question why the economic ties are there in the first place. I am arguing that, while economic relations may play a role in all kinds of spill-over effects, these economic ties are the result of ethno-linguistic ties between nations. Therefore, measuring the ethno-linguistic ties between nations and combining that with geographic spillover analyses makes more sense, because it captures both spillovers that stem directly from the ethno-linguistic ties and the spillovers that happen due to ethno-linguistically induced economic ties.

Of course, this leads to the problem of measuring ethno-linguistic ties. One thing that one could do is related to what Alesina *et al.* (2006) did and consists of looking at dyads of countries and measuring the percentage of the population that belongs to ethnic groups existing in both countries. However, this imports many of the problems that haunt the ELF literature, particularly group definitions. When using this method, it is very important to decide on which level of disaggregation the ethnic groups are measured. Considering northern Africa, the measure will give very different results when e.g. Berbers are considered one group or whether all individual Berber clans are considered separately. Additionally, the strict boundaries between ethnic groups are simply unrealistic, both in theory and in practice, for measuring purposes.

Instead, I propose an alternative way of measuring ELA. The first step in the construction of this measure is the recognition that ethnic identities consist of a number of different so-called *identity characteristics*. One could argue that different historical periods and different regions of Earth may require a different set of identity characteristics and I will therefore not define them yet. However, the kind of characteristics that one could think of are race, national origin, language, religion, clan identification, et cetera. An important feature of these identity characteristics is their cumulative nature: the more characteristics shared between two individuals, the more ethno-linguistically similar they are. Assuming that one is able to come up with a satisfactory set of identity characteristics, it is easy to see that it should be possible to classify all ethnic groups within a population according to them. Particularly, when using a very disaggregated ethnic dataset, it is possible to strictly identify each of the different characteristics that make up a particular ethnic group. This solves another problem from the ELF literature: how an ethnic group is defined. With this method, it can be recognised that two groups are highly similar, while still recognising their individuality. In fact, as I will mention later, using this same technique it is also possible to set up a measure of within-nation ethno-linguistic fractionalisation that does not have the problem of having to choose a level of aggregation of the data, and combines different features that one might deem important. Of course, such a measure is automatically closely related to Fearon's (2003) measure.

After having constructed a dataset of the region of interest that consists of ethnic

groups at the most disaggregated level possible with values for each of the different identity characteristics, one can construct a measure that incorporates both the sizes of the different ethnic groups and how different these groups really are. After all, groups i and j that share all-but-one of their characteristics are more likely to feel affinity towards each other than groups i and k, who share only one of these characteristics. The measure I am proposing to use is the following

$$ELA = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\alpha_i \cdot \beta_j \cdot c_{ij}\right)$$

where α_i is the share of population that ethnic group $i \in I$ has in country A and β_j is the share of population that ethnic group $j \in J$ has in country B. c_{ij}, finally, is the percentage of identity characteristics that are shared between groups i and j. This parameter can be anywhere between 0, if the two groups have nothing to do with each other, and 1, if they are actually the same group but live in different countries. This measure is closely related to the way Greenberg (1956), as reproduced by Fearon (2003), proposed to construct an alternative measure of ELF, except that it involves the relationship between countries instead of measuring just within one country and that I have approached the definition of their *resemblance factor* in a different way. In fact, it is easily possible to apply the same type of resemblance factor they use, instead of my identity characteristics. However, I feel that linguistic distances alone do not appropriately capture the entire arena of ethno-linguistic ties that one can describe with my proposed identity characteristics. On the other hand, it would also be easy, and feasible, to use my identity characteristics as their resemblance factor and come up with an alternative measure of ELF[7].

An advantage of this measure is its clear interpretation, similar to that of the ELF. Remember the interpretation of the ELF is *the probability that two individuals randomly drawn from a population are of the same ethnic group*[8]. This measure of ELA, on the other hand, measures *the percentage of shared identity characteristics of two individuals randomly drawn from two different populations.* In other words, how much affinity can a random person from country A be expected to have with a random person from country B.

In a practical application, however, it is probably rare to expect that Ethno-Linguistic Affinity is the only channel through which spillovers take place. In most

[7]Such a measure would look as follows: $ELF = \sum_{i=1}^{I}\sum_{j=1}^{J}(\pi_i \cdot \pi_j \cdot c_{ij})$, with π_i the percentage of group i in the total population and c_{ij} still the shared percentage of identity characteristics. Such a measure would take the linguistic focus of the measure Greenberg (1956) and Fearon (2003) propose and yield a measure that is not as perceptible to definition changes.

[8]Actually, ELF usually measures the probability that two individuals are from different ethnic groups, but to show the similarity between the measures, the current interpretation is more convenient.

conceivable examples, a combination of the Ethno-Linguistic and Geographic channels should be expected. A combination of these two channels is very easy, when using standard spatial econometric techniques. When setting up a contiguity matrix, one can simply multiply each of the observations for geographic distance with the corresponding measure of ELA, before performing the required row-normalisation. Like in other spatial econometric analyses, the kind of geographic distance measure is still open to debate, but this technique works, independent of whether center-point distances, distances of closest approach, border-lengths or another measure of contiguity are used.

Another thing to remember is that, so far, this measure has a range of potential applications and gives the researcher a lot of room for adjusting it to a suitable situation. The set of *identity characteristics* is, so far, undefined and can be chosen in order to accommodate the particular issue that is being researched. Channels of Ethno-Linguistic Affinity can be expected to differ strongly, depending on time and space. Examples of characteristics that may be relevant in some regions, but not in others include clan affilition (in Africa), caste (in India) or ancestry (in North America). This implies that it is important to decide which are the identity characteristics that fully describe the type of ethno-linguistic group association one is trying to measure.

3.1 Features of the ELA measure

Looking at the way the measure has been constructed, it appears to possess many appealing characteristics. First of all, the range of the measure is linear and clearly defined: $ELA \in [0, 1]$, where 0 means that the populations of two nations have absolutely nothing in common and 1 means that the two countries have completely homogeneous native populations, who share all their identity characteristics. There are two ways in which a country dyad can have a lower level of ELA. First, the populations can become more different. For example, two completely homogeneous nations, of which the two population groups share only half the identity characteristics is going to yield an ELA of 0.5. After all, without any uncertainty, two randomly drawn individuals will always share 50% of their characteristics[9]. The second way is when, instead of two equal homogeneous populations, the two countries both consist of the same two, equally-sized, ethnic groups that do not share any identity characteristics with each other. This would also lead to an ELA of 0.5, because there is a 50% probability of drawing two completely equal individuals and 50% probability of drawing two completely different individuals. The expected value is therefore 0.5.

Another desirable characteristic of the measure is its divisibility. After all, the

[9]In fact, the condition that both countries have a completely homogeneous population is unnecessary, as long as $c_{ij} = 0.5 \ \forall i, j$, the level of Ethno-Linguistic Affinity between the two countries will always be 0.5.

current measure is simply a sum of the separate distributions of the different identity characteristics:

$$ELA = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\alpha_i \cdot \beta_j \cdot c_{ij}\right) = \frac{1}{C}\sum_{c=1}^{C}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\alpha_i \cdot \beta_j \cdot (1 \mid c_i = c_j)\right)$$

where $c \in C$ are the different identity characteristics and c_i is the value of identity characteristic c for population group i.

A final attractive feature of this measure is the fact that the value of ELA does not change when a particular ethnic group is subdivided incorrectly. Measuring ethnic groups at a subdivision that is more detailed than strictly necessary will not change the value of ELA. After all, when several small groups share the same identity characteristics, these are summed up again when calculating the actual measure[10]:

$$\begin{aligned} ELA &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\alpha_i \cdot \beta_j \cdot c_{ij}\right) = \sum_{i=1}^{I}\left[\left(\sum_{j=1}^{J-1}\left(\alpha_i \cdot \beta_j \cdot c_{ij}\right)\right) + \left(\alpha_i \cdot \beta_J \cdot c_{iJ}\right)\right] = \\ &\sum_{i=1}^{I}\left[\left(\sum_{j=1}^{J-1}\left(\alpha_i \cdot \beta_j \cdot c_{ij}\right)\right) + \left(\alpha_i \cdot \beta_{J_1} \cdot c_{iJ_1}\right) + \left(\alpha_i \cdot \beta_{J_2} \cdot c_{iJ_2}\right)\right] \\ \text{if } \beta_{J_1} + \beta_{J_2} &= \beta_J \text{ and } c_{iJ} = c_{iJ_1} = c_{iJ_2} \end{aligned}$$

In fact, it is reasonable to say that for each identity characteristic the distribution over different groups is actually irrelevant. More particularly, the sub-measure for a single identity characteristic, where $k \in K$ are the different values a particular identity characteristic can take can be summarised as following:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\alpha_i \cdot \beta_j \cdot (1 \mid c_i = c_j)\right) = \sum_{k=1}^{K}\left(\min\left(\alpha_k, \beta_k\right)\right)$$

where α_k and β_k are the population shares that k has in nations A and B respectively.

3.2 Practical Construction of Measure

In this subsection, I set up a measure of Ethno-Linguistic Affinity between nations in Africa, which is used to analyse the spill-over effects of conflict in the following section. To construct my measure, I utilise an unusual source that does not seem to have been used often before: The *Joshua Project* (2007). This was a project originally started in 1995 and is currently an official ministry of the *U.S. Center for World Mission*,

[10] In fact, like Bossert *et al.* (2008) show in the case of a measure of ethno-linguistic fractionalisation, the optimal result is achieved when using actual individuals instead of groups. However, it is imply unfeasible to have such detailed data available for any sizeable group of countries.

an evangelical organisation aiming to spread the word of their religion to the so-called "unreached peoples of the Earth". While this is an unorthodox source, one can make strong arguments in favour of using this particular one. The data provided by the *Joshua Project* is extremely detailed and seems to combine many of the sources used in other papers[11,12], with an extensive local network that is able to provide more detail from a local perspective. Often, religious data may be questionable in its veracity, but the stated goal of the *Joshua Project* shows why this data is worth using: "The mission of *Joshua Project* is to help bring definition to the unfinished task of the Great Commission by identifying and highlighting the people groups of the world that have the least exposure to the Gospel and the least Christian presence in their midst"[13]. The religious fervency with which this organisation collects data works in our advantage. After all, no religion would want to underestimate their own follower base, but this project especially is trying to analyse which particular groups need their "help" the most, and therefore, it is also imperative not to overestimate their own following either. In fact, the *Joshua Project* is clearly best-served with true and correct data. Of course, one should not trust blindly, and where possible, I have consulted alternative sources to see whether the data provided by the *Joshua Project* was compatible and by and large, this did seem to be the case. Nowadays, the most popular source is the *World Christian Encyclopedia* (Barrett *et al.*, 2001) and in order to check the compatibility of that source and the Joshua Project, I have tried to match the entries from each of these sources. This process is not very easy, because different names are used for the same population groups and the level of detail differs per source as well. However, despite these problems and despite the difference in the time frame of the different sources, the correlation coefficient between the different entries is approximately 0.96. This re-enforces my premise that the *Joshua Project* is a valid data source.

A large advantage of the *Joshua Project* data is its amazing level of detail. Ethnic groups are split into micro-groups, as a result of which one gets a proper overview of all the information available. Most other sources (and other papers) use only groups that contain at least 1% of the population, but with my measure of ELA, this would not be convenient. After all, when calculating a traditional ELF-index, ethnic shares are multiplied with themselves and a group that is smaller than 1% of the population exerts influence of less than $0.01 \times 0.01 = 0.0001$ on the total ELF. However, with my index of ELA, in the most extreme case, where the 1%-group of one nation is 100% compatible with 100% of the population of a neighbouring state, the total influence would be significant at $0.01 \times 1 = 0.01$[14]. For Africa, the *Joshua Project* reports results

[11] Including Ethnologue, the World Christian Encyclopedia, the CIA World Factbook, Peoples-Groups.org and Harvest Information System.

[12] Of course, this source is also usable when analysing other world regions. Additionally, if one were to set up a domestic version of this measure, as suggested in section 3, any data required are available in there as well.

[13] From www.joshuaproject.net

[14] Of course, this example is extreme but for example the Ukrainians make up some 72% of the

for 3704 country-groups[15] and it is this level of detail that outweighs the problems this source may inherently contain. Of course, this does not answer any of the standard questions regarding endogeneity. When researching the influence of ethno-linguistic composition on some macro effect (particularly civil conflict), one should use the ex ante ethno-linguistic composition, as the ethno-linguistic composition may have been endogenously determined by the occurrence of conflict or anything else you are trying to measure. However, it has been argued in the past that due to the strong level of persistence among ethnic composition, one does not need to worry much about this problem. Roeder (2001) shows that, particularly in Africa, the ethnic composition persistence is indeed very high and I use this as a basic assumption to be able to continue with this dataset.

When an appropriate dataset is found, the foremost issue that comes to mind when using the previously proposed way of setting up a measure of ELA, is the recognition of the relevant identity characteristics. Given that the aim of this exercise is to explain conflict spillovers and the geographic area of interest is Africa, this already points in the direction of the type of characteristics that one should be looking for. They are largely determined by ethnic characteristics, which are unfortunately typically hard to classify. A first measure, however, is the self-identified (internationalised) ethnic group. This is the most basic level of ethnic affiliation and is directly connected to the second identity characteristic, original language spoken by an ethnic group. While these two characteristics are likely to be strongly correlated, they capture in fact two different aspects, because different groups may speak the same language and in extraordinary cases, the same group may be speaking different languages in different nations.

These first two characteristics are at a very micro level, but it is important to try and capture the interconnections of these groups at a slightly higher level as well. For this, I use macro-measures for both of the first two identity characteristics. In the case of linguistic ties, I follow Greenberg's (1956) theory that languages that have split more recently belong to ethnic groups that are more closely related and have therefore used the *Ethnologue* (Gordon, 2005) and *Rosetta Project* (2007) databases to construct separate linguistic groups at the level of a sub-family. The subfamily of Atlantic Congo (family: Niger-Congo), however, since it is so prevalent in Africa, turned out to contain a majority of the country-groups and of the population involved, so I have split this subfamily into smaller sections[16], following Gordon and the *Rosetta Project*.

---

population in Ukraine. In nearby countries like Georgia and Kyrgyzstan, Ukrainians indeed make up around 1% of the population, so the influence of this relationship on the total level of affinity is still quite significant.

[15] The whole world contains a total of 15,965 country-groups.

[16] To be precise, the Atlantic Congo subfamily has been split in Atlantic, Ijoid and Volta-Congo, where the last was split into Dogon, Kru, Kwa, Northern Languages and Benue Congo. Finally, Benue Congo was split in West Benue Congo, Cross-River, Platoid, Bandoid (non-Bantu) and Narrow Bantu.

Table 1: Summary statistics for different identity characteristics

	categories	average		largest category	
	Nr.	groups (nr.)	ppl (mln)	groups (nr.)	ppl (mln)
Language	2079	1.8	0.5	83	45.7
Linguistic group	65	57.0	14.4	906	259.5
Ethnic group	2435	1.5	0.4	53	44.4
People Cluster	98	37.8	9.5	319	65.9
Religion	22	168.4	42.4	988	378.4
Total	3704	1	0.3	1	43.4

Note: These summary statistics include the size of the average and largest categories within an identity characteristic. For the largest category, the number of groups and the number of people are not necessarily in the same category (for example, while the largest number of groups can be found in the Benue people cluster, it is the Egyptian people cluster that contains the largest number of people).

The ethnic equivalent of this last characteristic is a division made according to "People Cluster". This term is posited by Johnstone (2007) and defined as "[a] smaller grouping of peoples within an affinity bloc, often with a common name or identity, but separated from one another by political boundaries, language or migration patterns", where an affinity bloc is defined as "[a] large grouping of peoples related by language, history, and culture, and usually indigenous to a geographical location". Johnstone's objective in the construction of the measure of People Clusters is a simplification of the task of evangelisation. While not his original aim, he does provide a framework for the logical clustering of all these ethnic groups that makes the list of available groups significantly smaller. The total number of People Clusters in Africa is 98 (after merging some non-native African groups that were too small to exist on their own) and this measure truly seems to capture an appropriate subdivision of all different ethnicities in Africa.

The final characteristic I use is one that moves away from evolutionary development over time and applies another, relatively recent, phenomenon: religion. In the area of conflict, religion is known to be a significant divider between different sides and a unifier among those who follow the same religion. Therefore, religion is used as one of the five identity characteristics. The *Joshua Project* provides data on what the main religion of an ethnic group is for all groups, but unfortunately the subdivision is not always provided. I have used generally accessible sources, such as the *Encyclopedia Brittanica* to fill in some of the blanks of ethnicities that did not yet have a subgrouping. As a result, about 72% of all ethno-linguistic groups, covering more than 90% of the population, have been given a religious subgrouping. The missing groups do have the general religious affinity (i.e. Christianity or Ethnic Religions), and this data has been used to replace the missing values. In fact, a strategy has been

followed in which an observation of the affinity between subgroups of religions has been replaced by the affinity between actual religions whenever one of the two groups did not have an observation for the religious subgrouping. This clouds the actual estimation, but due to the fact that the missing groups are only small in number and are the relatively smaller groups, I think the estimation is still very reasonable. Table 1 contains the most important summary statistics for all five identity characteristics. *Average* refers to the number of groups and the number of people in the average category of an identity characteristic. *Largest category* refers to the largest category for each.

Having so constructed a profile of ethno-linguistic identity characteristics, it is thus possible to construct the measure of ELA as proposed in the previous subsection. Doing this for the African data that are available to me now, generates a matrix of $53 \times 53 = 2809$ dyadic relations and a corresponding measure of Ethno-Linguistic Affinity between them. Table 2 reports the summary statistics for the dyadic relations. As can be seen, the spread is quite large. Whereas Burundi's maximum is reached at 0.591, implying that a randomly drawn individual from Burundi shares 59.1% of its characteristics with a randomly drawn individual from Rwanda, the Central African Republic's highest value[17] is only 0.115. On average, two randomly drawn individuals from two different countries in Africa share 8.1% of their identity characteristics and two individuals from two neighbouring states share 19.8% of their characteristics. Remembering that I use 5 different identity characteristics, it can be said that two individuals from two neighbouring countries share on average approximately one identity characteristic. ELA is related to distance, but not as strongly as one might expect. The correlation coefficient between the logarithmic distance in kilometers between centre points and ELA is -0.42.

Unfortunately, some caution is in order for the current measure. Two major problems should be discussed, although I think they can be dismissed in the end. First, there is the *ex post* definition of the individual groups. As discussed before, previous authors have dismissed this problem as minor but I am afraid that the level of detail of the data used makes them more susceptible to problems. Additionally, the fact that the data are mostly very recent should also create worries, because the cumulative amount of dislocated people due to civil war has increased significantly over time. Roeder (2001) compares ELF's from different time periods, but his most recent one employs data from 1985. This excludes the increased number of violent civil conflicts from the nineties, most importantly the Rwanda and Burundi ethnic conflicts. Unfortunately, there is little to be done about this and I simply have to follow previous authors and their argument that ethnic heterogeneity persistence really is very high.

The second problem is the actual data source. While I argued earlier that there are strong arguments in favour of using this particular source, potential contamination

[17] Surprisingly, the Central African Republic's highest value of ELA is achieved in its relationship with non-contiguous Burkina Faso.

Table 2: Summary statistics for ELA dyads

	Max	Average	Min
Highest ELA	0.591	0.315	0.115
	BUR-RWA	-	CAR-BFA
Avg ELA	0.134	0.081	0.031
	COM	-	MAD
Avg ELA (contiguous)	0.439	0.198	0.067
	TUN	-	CAR
Lowest ELA	0.025	0.003	2.3×10^{-8}
	TZA-ETH	-	ERI-LES
Lowest ELA (contiguous)	0.391	0.140	0.005
	TUN-LIB	-	DRC-SUD

Note: Max, average and Min values are reported for the highest ELA, the ELA country averages, ELA country averages including contiguous states only, lowest ELA and the lowest ELA including contiguous states only. The table also reports between/in which nations these extremes are found. For the neighbour-only values, only directly contiguous neighbours are included and island nations are left out of the analysis completely.

due to its religious purpose cannot be excluded. Again, as explained earlier, I have made all possible effort to make sure that this contamination is as limited as possible, but it would be interesting to replicate the results with data from an alternative source. Unfortunately this is not feasible in the short run, as no comprehensive alternative data source exists that contains as much information in such detail as this particular one.

4 Practical Application: Conflict Spillovers

In this section, I set up a practical application of my index of ELA between nations. Several authors, including the ones I have quoted earlier, have published papers which try to explain the occurrence of civil conflict. A smaller number of authors have used spill-over effects as one of the mechanisms involved in this and I think it is very important to do so. I am therefore proposing to use a combination of ELA and a measure of geographic distance, in addition to a number of other variables that are known to predict civil conflict.

Spatial econometrics makes use of a number of specially developed techniques, as explained so well in Anselin (1988). More recently, Beck *et al* (2006) is an excellent contribution to the estimation issues when using spatial econometrics in a

political economics context. Their explanation of the particularities regarding the interpretation of coefficients and the estimation techniques is illuminating.

4.1 Model

Collier and Hoeffler (2004), Montalvo and Reynal-Querol (2005) and Gleditsch (2007), all use logit models to analyse the impact of different factors on the incidence of conflict. I use the same technique, with the exception of using conflict initiation instead of conflict incidence as a dependent variable. As a source of the conflict data, I use the PRIO database (Gleditsch *et al.*, 2002). Among the different standard explanatory variables used by these authors and that I use as well, are the natural logarithms of GDP and population (Heston *et al.*, 2006), the years since the most recent conflict in a country (Gleditsch *et al.*, 2002)[18], a measure of political freedom (Center for Global Policy, 2008)[19], a measure of ethnic polarization (Montalvo and Reynal-Querol, 2005) and a measure for mountainous terrain (Gerrard, 2000)[20]. The entire sample consists of 53 countries over 45 years (1960-2004). However, since some countries were not yet independent during some time of this sample, the maximum number of observations is 2134.

The most important part of the analysis, however, is obviously the way I make use of my measure of ELA. It is important to recognise that even if conflict is likely to follow ethno-linguistic patterns when spillovers take place, there has to be a geographical proximity factor involved as well. I therefore combine ELA with a measure of geographic distance. In fact, the measure of geographic proximity I use is border length between nations (CIA, 2007)[21]. The possibility of conflict spilling over from one country to a noncontiguous one seems dismissable and including that would only increase the possibility of spurious correlations[22]. So, the contiguity matrix W con-

[18] Following Gleditsch (2007), instead of just the number of years since the last conflict, an exponential function is included: $e^{\left[-\frac{y}{\alpha}\right]}$, in which y is the number of conflict-free years and a takes the experimentally determined value of 4. For the number of years since the last conflict, only years are included after 1950 and after independence.

[19] Following Gleditsch (2007), footnote 14, I do not directly use the Polity2 index. As Gleditsch warns, the makers of the index replace all missing values with a value 0. This is a dubious choice, since missing values are generally caused by an extreme flux in the political variables. Instead, Gleditsch awards the lowest possible value of -10 to these observations with 'irregular policy values'.

[20] For this variable, I follow Collier and Hoeffler (2004), who use this same source because it gives a good estimate for mountainous regions that give an opportunity for rebels to hide. The measure combines elevation, relative relief and area in order to identify mountainous areas.

[21] As often in spatial econometrics, islands present a problem. I deal with the issue on a case-by-case basis and employ several formulas. The assumed border length influence of i on j, when i is either a coastal nation or an island and j is an island is $\delta_{ij} = 100 \cdot \frac{coastline_i}{distance_{ij}}$; the border length influence of an island i on coastal nation j is $\delta_{ij} = coastline_i \cdot \frac{coastline_j}{\sum borderlength_j} \cdot \frac{distance_{ij} \cdot coastline_j}{\sum\limits_k (distance_{jk} \cdot coastline_k)}$, where k stands for all the islands that are within reach of j.

[22] As Beck *et al.* (2006, p. 28) note, "The assumption that these connectivities are known a priori

sists of a square matrix with all nations along the horizontal and vertical axes and with the matrix elements e_{ij} describing the relation between i and j, normalised over rows, and is defined as follows:

$$e_{ij} = \frac{ELA_{ij} \circ \delta_{ij}}{\sum_{i=1}^{N} (ELA_{ij} \circ \delta_{ij})}$$

where δ_{ij} is the geographical distance measure in use. A potential criticism is that this assumes equality between all borders and that spillovers should be more likely in the cases where borders clearly divide particular ethnic groups, as opposed to distant borders across mostly impassable terrain (e.g. the southern borders of Algeria). However, such prohibitive geographical features that limit potential spillovers (sea, desert, mountains) also lead to a stronger separation of the populations on opposite sides of the division. As a result, the level of Ethno-Linguistic Affinity can be expected to be low in such cases and this is unlikely to be countered by the border lengths. In the case of Algeria, the presence of the Sahara to separate the northern Algerian population centres from their southern neighbours in Mali and Niger, leads to a low level of Ethno-Linguistic Affinity between Algerians and their souther neighbours and consequently a lower spill-over probability.

Putting all these factors together in a logit model results in a complete regression that looks as follows:

$$\Pr(y_{i,t} = 1) = \frac{e^{\eta_{i,t}}}{1 + e^{\eta_{i,t}}}$$

where

$$\begin{aligned} \eta_{i,t} &= \beta_0 + \beta_1 \ln(gdp_{i,t-1}) + \beta_2 \ln(pop_{i,t-1}) + \beta_3 peace_{i,t} + \\ &\quad \beta_4 mount_i + \beta_5 dem_{i,t} + \beta_6 W(conf_t) + \varepsilon_{it} \end{aligned}$$

in which $y_{i,t}$ is a variable that takes value 1 if a new conflict was initiated in country i during year t, $gdp_{i,t-1}$ is the one-period lagged level of GDP, $pop_{i,t-1}$ is the one-period lagged population size, $peace_{i,t}$ is a measure for the years of peace at the start of the year, $dem_{i,t}$ is the adjusted Polity2 score for an observation-year and $conf_{i,t}$ is a dummy that takes value 1 if conflict is taking place in a country during period t. Estimating this seemingly easy regression is not completely trivial, however, as $y_{i,t} \in conf_t$ and as a result, the value change in the dependent variable influences that particular independent variable. This is one of the things that Beck *et al.* (2006) mean when they say that caution has to be exercised when interpreting the coefficients in a spatial econometric model. There is, however, a solution. According to Ward and

is both a strong assumption and a cricitical one for the methods of spatial econometrics to work". For this reason, it is important to reject other mechanisms (such as distance between center points or distances of closest approach) on theoretical grounds.

Gleditsch (2002) and Gleditsch (2007), one can use either Markov Chain Monte Carlo simulation or Maximum Pseudo-Likelihood (MPL) methods. The results, however, are very similar and as the MPL method is easier to apply I use this in the estimations.

4.2 Results

Column 1 of table 3 contains a baseline model for conflict initiation, in which there are no spillovers. As can be seen, lagged GDP, lagged population and the years of peace since the last conflict are all significant and have the expected sign. *mountain* and *polity*2 have the expected sign, but are not significant. The table then continues to show the model described in the previous subsection, of which the results are shown in column 2. Again, lagged GDP has a negative impact on the probability of new conflict initiation and lagged population a positive one. Due to the way the number of peace years has been defined, the positive and significant coefficient of *peace* implies that a country that has been in peace for a longer time has a lower probability of conflict outbreak. *mountain*, which measures the inaccessibility of terrain, is positive, as expected, but not significant. Finally, the *polity*2 score has an insignificant negative impact. The most interesting and relevant variable, however, is of course $W \cdot conf$. It can be seen immediately that the spillover of conflict along the combination of a geographic and my proposed ethno-linguistic channel is positive and significant, which means that a country whose ethno-linguistically close neighbours are suffering from conflict is more likely to suffer conflict initiation as well. In order to check whether the claim is warranted that the ethno-linguistic channel plays an important role in this, column 3 of table 3 shows the same result, but with W defined using only border-lengths. As can be seen, the significant relation between neighbouring conflict and home-country conflict disappears. Finally, in column 4, the result is shown when only ELA is used in the contiguity matrix. In this case, the result also disappears, which is not surprising, because it includes linkages that are too far-sought to influence conflict spillovers (such as the strong ethno-linguistic ties between the north-east and north-west of Africa).

In table 4, a number of variations are shown. The first three columns contain results for the same regression, but with a measure of ethnic polarisation included (from Montalvo and Reynal-Querol, 2005). The strength of the results is slightly reduced, but overall the conclusion remains the same. I believe, however, that ethnic polarisation should not be included in these regressions, due to the interference it has with the measurement of Ethno-Linguistic Affinity, as these variables are both measuring along the same dimension.

The final two columns (4 and 5) drop the insignificant variables of *polity*2 and *mountain* respectively. This is not done so much on theoretical grounds, but based on the fact that these variables are the bottlenecks for the number of observations. Dropping either of these variables increases the number of observations significantly, and the results are unaffected. The regressions that exclude either the ELA element

Table 3: Regression results of logit MPL estimations

	1	2	3	4
spill-overs:	baseline	ELA*border	border	ELA
C	-2.747**	-3.182**	-2.934**	-3.085**
	1.294	1.294	1.287	1.273
$\ln(gdp_{t-1})$	-0.335**	-0.319**	-0.327**	-0.340**
	0.145	0.147	0.146	0.144
$\ln(pop_{t-1})$	0.224***	0.239***	0.225***	0.212**
	0.084	0.081	0.083	0.084
peace	0.668***	0.667***	0.669***	0.695***
	0.251	0.251	0.251	0.253
mountain	0.460	0.268	0.353	0.435
	0.410	0.440	0.433	0.409
*polity*2	-0.014	-0.013	-0.014	-0.019
	0.020	0.020	0.020	0.020
$W \cdot conf$		0.752**	0.496	2.224
		0.347	0.356	1.432
N	1837	1837	1837	1837
$LR - \chi^2$	32.68	36.30	34.36	34.73
df	5	6	6	6

*Note: Results of the most important regressions, using robust Maximum Pseudo-Likelihood estimations in a logit model. Variables defined in the text. *, ** and *** imply significance at 10%, 5% and 1% respectively.*

Table 4: Further regression results of logit MPL estimations, using alternative specifications

	1	2	3	4	5
	ELA*dist	dist	ELA	ELA*dist	ELA*dist
C	-3.440***	-3.173**	-3.353***	-3.173**	-2.923**
	1.308	1.299	1.287	1.264	1.247
$\ln(gdp_{t-1})$	-0.382**	-0.399**	-0.408**	-0.368***	-0.347**
	0.158	0.158	0.158	0.141	0.143
$\ln(pop_{t-1})$	0.258***	0.245***	0.236***	0.286***	0.230***
	0.085	0.086	0.089	0.073	0.077
$peace$	0.522*	0.520*	0.548**	0.518**	0.700***
	0.275	0.274	0.278	0.237	0.239
$mountain$	0.292	0.366	0.412	0.259	
	0.447	0.442	0.424	0.439	
$polity2$	-0.009	-0.009	-0.013		-0.016
	0.020	0.020	0.020		0.020
$polar$	1.154**	1.161**	1.198**		
	0.568	0.574	0.578		
$W \cdot conf$	0.643*	0.340	1.968	0.742**	0.973***
	0.358	0.364	1.482	0.345	0.320
N	1753	1753	1753	1950	1894
$LR - \chi^2$	40.68	39.28	39.82	35.33	39.18
df	7	7	7	5	5

*Note: Results of the most important regressions, using robust Maximum Pseudo-Likelihood estimations in a logit model. Variables defined in the text. *, ** and *** imply significance at 10%, 5% and 1% respectively.*

or the geographic element (not shown) also yield the same results as before. A final check (not shown) leaves out both insignificant variables and the results still remain the same (with N=2041).

Collier and Hoeffler (2004) also argue that a measure for exports of primary commodities should be used as an explanatory variable. According to them, countries that have large exports of primary commodities are more likely to have rebellion due to the opportunity of rebel financing it presents. The authors also show that in their dataset they get significant results implying that the proposed mechanism is indeed at work. However, Fearon (2005) argues that the relationship between primary commodity exports and conflict is not nearly as clear. His argument is that these exports actually provide an easy source of finance for the government, which may lead to less stable institutions, but also potentially to a government that is better able to fight off a rebel uprising. In order to see whether there is any effect, I have also added a

measure for primary commodity exports to my model[23], but the influence is insignificant. The spill-over effect remains strong and other explanatory variables are also unaffected.

The interpretation of the results is one thing to look at carefully, however. Due to the logit structure of the analysis, one must be careful in interpreting the coefficients. In fact, it is most convenient to report the estimates for the influence the different variables have, keeping the others constant at their mean. Taking the original model, as shown in column 2 of table 3 and keeping all variables at their mean values, the ceteris paribus addition of one standard deviation of conflict among neighbours increases the probability of conflict by 1.0 percentage point. To compare, increasing lagged GDP, lagged population or *peace* by one standard deviation, leads to changes of -1.2, 1.5 and -1.1 percentage points respectively. So changes in those factors impact the probability of conflict initiation to a similar degree as changes in neighbouring conflict. The implied changes appear to be quite small, but it should be taken into account that the probability of conflict initiation is small to begin with at 4.8%. Therefore, a change by 1.0 percentage points, implies an increased probability of 21.3%, which cannot be considered small.

Finally, it is good to have another look at the measure of ELA. While it would be desirable to do a factor analysis of some kind to analyse whether the set of identity characterics used now is appropriate to use, this is unfortunately not possible. Any analysis of such kind is based on the assumption that the final measure is a linear sum of the different components. This is not the case here, due to the fact that the actual measure used in the regression is the result of the row-normalisation of a multiplication of ELA and border length. This makes different components non-linear. However, as an alternative measure, it is possible to re-create alternative ELAs. In table 5, for each of the columns, a different identity characteristic is dropped. At that moment, a new ELA is calculated on basis of only four identity characteristics, which is then combined with the distance measure and finally used in these regressions. The results show that little changes when dropping any of the variables. Only the Linguistic Subgroup has a strong effect on the point estimate and causes a reduction in the significance level (to $z = 1.95$), which shows its importance. In the other direction, dropping the People Cluster variable increases the point estimate for that parameter by a small amount. This can safely be ignored.

5 Conclusion

In this paper, an index of Ethno-Linguistic Affinity between nations is set up. Using relatively simple tools, it appears to be easy to set up such an index that is

[23] I employ the dataset that Fearon (2005) uses, which is a dataset created on basis of Collier and Hoeffler's (2005) data. The latter use 5-year periods for their analyses, whereas the first uses yearly data, just like me. Unfortunately, the dataset finishes is 1999, which causes a fairly large drop in the number of available observations (N=1630).

Table 5: Logit regressions, where each column drops one of the identity characteristics

	1	2	3	4	5
missing char	GRP	LAN	LNG	CLUS	REL
C	-3.188**	-3.183**	-3.097**	-3.210**	-3.138**
	1.294	1.294	1.292	1.296	1.293
$\ln(gdp_{t-1})$	-0.318**	-0.319**	-0.319**	-0.316**	-0.324**
	0.147	0.147	0.146	0.146	0.147
$\ln(pop_{t-1})$	0.240***	0.239***	0.233***	0.239***	0.241***
	0.081	0.081	0.082	0.081	0.081
peace	0.666***	0.666***	0.675***	0.663***	0.675***
	0.251	0.251	0.251	0.251	0.251
mountain	0.268	0.268	0.288	0.261	0.287
	0.440	0.440	0.437	0.442	0.439
*polity*2	-0.013	-0.013	-0.014	-0.013	-0.013
	0.020	0.020	0.020	0.020	0.020
$W \cdot conf$	0.757**	0.754**	0.652*	0.779**	0.684**
	0.346	0.346	0.334	0.349	0.343
N	1837	1837	1837	1837	1837
$LR - \chi^2$	36.38	36.34	35.36	36.76	35.49
df	6	6	6	6	6

*Note: Results of logit regressions with alternative measures of ELA. Each column drops one of the identity characteristics for its construction of ELA. The missing characteristics are GRP=Ethnic Group; LAN=Language; LNG=Linguistic Group; CLUS=People Cluster; REL=Religious subgroup and other variables defined in the text. *, ** and *** imply significance at 10%, 5% and 1% respectively.*

able to avoid many of the caveats that haunt ethno-linguistic indices in general. When dissecting ethnicities into separate identity characteristics and considering the (dis)similarity on different levels it turns out to be possible to set up a measure that successfully exploits the varying sizes of differences between ethnicities along the lines of these characteristics.

Many spatial-econometric analyses use geographic measures of distance between nations as core variables, but in this paper I argue that in many of these cases it is not merely physical proximity that has the strongest influence, but the ethno-linguistic (dis)similarity in dyads of nations. Examples from the spatial-econometric literature include the spill-over effects of trade, conflict and economic growth, all of which might benefit from the inclusion of an ethno-linguistic component. The most interesting and promising field, however, is the spillover of institutions. It can easily be argued that such spillovers are among the most likely to spill over along ethno-cultural lines, particularly if one is able to include particularly appropriate identity characteristics. So far, democracy has been the only kind of institution for which there is a fairly substantial body of literature, whereas other kinds of institutions, including trade and social institutions could also be considered for spill-over effects. In all of these, though, the inclusion of an ethno-linguistic component is an idea worth considering.

However, a large challenge for the practical application of the proposed measure concerns data collection. Clearly, this measure benefits from the most detailed level of data collection, but it may be difficult to collect such detailed data and guarantee its accuracy and its completeness. The latter is pivotal to being able to actually use this measure of ELA, so it should not be underestimated. This is also immediately related to the largest drawback of the application worked out in the current paper. The data source is unorthodox, which may lead some to dismiss it. However, I argue that the creators of this database had strong incentives to make it as accurate as possible and that it can therefore be used.

Finally, an example of an application of this measure of ELA is included. Ordinary geographic distances dismiss the existence of conflict spillovers in Africa, but when including my measure of ELA, the results change drastically. Conflict can clearly be seen to be more likely to initiate in countries that have an ethno-linguistically similar neighbouring country suffering from conflict.

References

[1] Abreu, Maria, Henri L.F. de Groot and Raymond J.G.M. Florax (2004), "Space and Growth", *Tinbergen Institute Discussion Paper*, TI 2004-129/3

[2] Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat and Romain Wacziarg (2003), "Fractionalization", *Journal of Economic Growth*, Vol. 8, pp. 155-194

[3] Alesina, Alberto, William Easterly and Janina Matuszeski (2006), *Artificial States*, NBER Working Paper 12328

[4] Annett, Anthony (2001), "Social Fractionalization, Political Instability, and the Size of Government", *IMF Staff Papers*, Vol. 48 (3), pp. 561-592

[5] Anselin, Luc (1988), *Spatial Econometrics*, Kluwer Academic Publishers, Dordrecht, the Netherlands

[6] Barrett, David B., George Thomas Kurian and Todd M. Johnson (eds.) (2001), *World Christian Encyclopedia: A Comparative Survey of Churches and Religions in the Modern World*, Oxford University Press, 2nd edition

[7] Beck, Nathaniel, Kristian Skrede Gleditsch and Kyle Beardsley (2006), "Space Is More Than Geography: Using Spatial Econometrics in the Study of Political Economy", *International Studies Quarterly*, Vol. 50, pp. 27-44

[8] Bossert, Walter, Conchita D'Ambrosio and Eliana La Ferrara (2008), *A Generalized Index of Fractionalization*, Bocconi University Working Paper

[9] Buhaug, Halvard and Kristian Skrede Gleditsch (2008), "Contagion or Confusion? Why Conflicts Cluster in Space", *International Studies Quarterly*, Vol. 52(2), pp. 215-233

[10] Cavalli-Sforza, L. Luca, Paolo Menozzi and Alberto Piazza (1994), *The History and Geography of Human Genes*, Princeton University Press

[11] Center for Global Policy (2008), "Polity IV Project", <http://www.cidcm.umd.edu/polity/>, retrieved January 2008

[12] CIA (2007), "The World Factbook", <https://www.cia.gov/cia/publications/factbook/>, retrieved November 2007

[13] Collier, Paul and Anke Hoeffler (2004), "Greed and Grievance in Civil War", *Oxford Economic Papers*, Vol. 56 (4), pp. 563-595

[14] Easterly, William and Ross Levine (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics*, Vol. 112 (4), pp. 1203-1250

[15] Esteban, Joan and Debraj Ray (1994), "On the Measurement of Polarization", *Econometrica*, Vol 62 (4), pp. 819-851

[16] Fearon, James D. (2003), "Ethnic and Cultural Diversity by Country", *Journal of Economics Growth*, Vol. 8, pp. 196-222

[17] Fearon, James D. (2005), "Primary Commodity Exports and Civil War", *Journal of Conflict Resolution*, Vol. 49 (4), pp. 483-507

[18] Fearon, James D. and Laitin, David D. (2003), "Ethnicity, Insurgency and Civil War", *American Political Science Review*, Vol. 97 (1), pp. 75-90

[19] Gerrard, A.J.W. (2000), *What is a Mountain? Background Paper to definition of mountains and mountain regions*, World Bank mimeo

[20] Gleditsch, Kristian Skrede (2007), "Transnational Dimensions of Civil War", *Journal of Peace Research*, Vol. 44 (3), pp. 293-309

[21] Gleditsch, Nils Petter, Peter Wallensteen, Mickael Eriksson, Margareta Sollenberg and Håvard Strand (2002), "Armed Conflict 1946-2001: A New Dataset", *Journal of Peace Research*, vol. 39(5), pp. 615-637

[22] Gordon, Raymond G., Jr. (ed.) (2005), *Ethnologue: Languages of the World, Fifteenth edition*, Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/

[23] Greenberg, Joseph H. (1956), "The Measurement of Linguistic Diversity", *Language*, Vol. 32 (1), pp. 109-115

[24] Guiso, Luigi, Paolo Sapienza and Luigi Zingales (2007), *Cultural Biases in Economic Exchange?*, EUI Working Papers, ECO 2007/42

[25] Heston, Alan, Robert Summers and Bettina Aten (2006), *Penn World Table Version 6.2*, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania

[26] Johnstone, Patrick (2007), "Affinity Blocks and People Clusters: An Approach Toward Strategic Insight and Mission Partnership", *Mission Frontiers*, March-April, pp. 8-15

[27] Joshua Project (2007), "Joshua Project: Bringing Definition to the Unfinished Task", <www.joshuaproject.net>, retrieved November 2007

[28] Laitin, David D. (2000), "What Is a Language Community?", *American Journal of Political Science*, Vol. 44 (1), pp. 142-155

[29] Mauro, Paolo (1995), "Corruption and Growth", *The Quarterly Journal of Economics*, Vol. 110 (3), pp. 681-712

[30] Michalopoulos, Stelios (2007), *Ethnolinguistic Diversity: Origins and Implications*, working paper, Brown University

[31] Miguel, Edward, Shanker Satyanath and Ernest Sergenti (2004), "Economic Shocks and Civil Conflict: An Instrumental Variables Approach", *Journal of Political Economy*, Vol. 112 (4), pp. 725-753

[32] Montalvo, José G. and Marta Reynal-Querol (2005), "Ethnic Polarization, Potential Conflict, and Civil Wars", *American Economic Review*, Vol. 95 (3), pp. 796-816

[33] Murdoch, James C. and Todd Sandler (2004), "Civil Wars and Economic Growth: Spatial Dispersion", *American Journal of Political Science*, vol. 48(1), pp. 138-151

[34] Posner, Daniel N. (2004), "Measuring Ethnic Fracionalization in Africa", *American Journal of Political Science*, Vol. 48 (4), pp. 849-863

[35] Roeder, Philip G. (2001), "Ethnolinguistic Fractionalization (ELF) Indices, 1961 and 1985." February 16, <http//:weber.ucsd.edu/~proeder/elf.htm>, retrieved January 3, 2008

[36] Rosetta Project (2007), "The Rosetta Project Digital Language Archive", <http://www.rosettaproject.org>, retrieved December 2007

[37] Sambanis, Nicholas (2002), "A Review of Recent Advances and Future Directions in the Quantitative Literature on Civil War", *Defence and Peace Economics*, vol. 13(3), pp. 215-243

[38] Spolaore, Enrico and Romain Wacziarg (2006), *The Diffusion of Development*, NBER Working Paper Series 12153

[39] Ward, Michael D. and Kristian Skrede Gleditsch (2002), "Location, Location, Location: An MCMC Approach to Modeling the Spatial Context of War and Peace", *Political Analysis*, Vol. 10 (3), pp. 244-260