Better Transcription of UK Supreme Court Hearings

Hadeel Saadany[1,*], Catherine Breslin[2], Constantin Orăsan[3] and Sophie Walker[4]

[1]Centre for Translation Studies, University of Surrey, United Kingdom
[2] Kingfisher Labs Ltd, United Kingdom
[3]Centre for Translation Studies, University of Surrey, United Kingdom
[4]Just Access, United Kingdom

Abstract
Transcription of legal proceedings is very important for enabling access to justice. However, manual speech transcription is an expensive and slow process. In this paper we describe part of a combined research and industrial project for building an automated transcription tool designed specifically for the justice sector in the UK. We explain the challenges involved in transcribing court room hearings and the Natural Language Processing (NLP) techniques we employ to tackle these challenges. We will show that fine-tuning a generic off-the-shelf pre-trained Automatic Speech Recognition (ASR) system with an in-domain language model as well as infusing common phrases extracted with a collocation detection model can improve not only the Word Error Rate (WER) of the transcribed hearings but avoid critical errors that are specific of the legal jargon and terminology commonly used in British courts.

Keywords
Legal Transcription, UK Supreme Court, Automatic Speech Recognition

1. Introduction

There has been a recent interest in employing NLP techniques to aid the textual processing of the legal domain [1, 2, 3, 4]. In contrast, processing spoken court hearings has not received the same attention as understanding the legal text documents. In the UK legal system, the court hearings sessions have a unique tradition of verbal argument. Moreover, these hearings crucially aid in new case preparation, provide guidance for court appeals, help in legal training and even guide future policy. However, the audio material for a case typically spans over several hours, which makes it both time and effort consuming for legal professionals to extract important information relevant to their needs. Currently, the existing need for legal transcriptions (covering 449K cases p.a in the UK across all court tribunals [5] is largely met by human transcribers.

Although there are several current speech-to-text (STT) technology providers which could be used to transcribe this data automatically, most of these systems are trained on general domain data which may result in domain-specific transcription errors if applied to a specialised domain. One way to address this problem is for end-users to train their own ASR engines using their in-domain data. However, in most of the cases the amount of data available is too low to enable them to train a system which can compete with well-known cloud-based ASR systems which are trained on much larger datasets. At the same time, in commercial scenarios, using generic cloud-based ASR systems to transcribe a specialised domain may result in a sub-optimal quality transcriptions for clients who require this service.

Table 1
Examples of Errors Produced by Amazon Transcribe for Legal Hearings. Errors and Corrections are typed in bold.

Model	Transcript
Reference	So **my lady** um it is difficult to..
AWS ASR	So **melody** um it is difficult to...
Reference	**All rise** ...
AWS ASR	**All right** ...
Reference	it makes further **financial order**
AWS ASR	it makes further **five natural**

This holds particularly true for British court room audio procedures. When applying a generic cloud-based ASR system (in our case Amazon Transcribe) on British court rooms, the Word Error Rate (WER) remains relatively high due to hearings' length, multiplicity of speakers, complex speech patterns, and more crucially, due to unique pronunciations and domain-specific vocabulary. Examples in Table 1 show some common problems we faced when transcribing UK court hearings by on-the-shelf ASR systems such as Amazon Web Services (AWS) Transcribe[1]. The references are taken from human-generated ground-truth transcripts of real UK Supreme Court Hearings[2] created by the legal editors

Workshop on Artificial Intelligence for Access to Justice (AI4AJ 2023), June 19, 2023, Braga, Portugal
*Corresponding author.
hadeel.saadany@surrey.ac.uk (H. Saadany)
0000-0002-2620-1842 (H. Saadany)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://aws.amazon.com/transcribe/
[2]https://www.supremecourt.uk/decided-cases/index.html

CEUR-WS.org/Vol-3435/short4.pdf

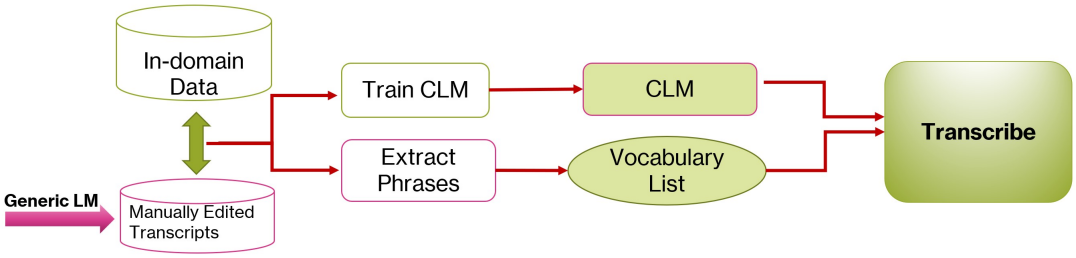

Figure 1: Pipeline for Improving ASR Output for Legal Specific Errors

in our project's team. The first error is due to a special pronunciation of the phrase '*my lady*' in British court rooms as it is pronounced like '*mee-lady*' when barristers address a female judge. Similarly, in the second example, the error relates to the linguistic etiquette of UK court hearings which the ASR system consistently fails to recognise. The error in the third example, on the other hand, is related to legal terminology critical of the specific transcribed case. Errors similar to the third example are numerous in our dataset and also affect named entities such as numbers and names that are vital in understanding the legal argument in the transcribed cases. These errors can lead to serious information loss and cause confusion.

In this paper, we describe a joint research and commercial effort to perform domain adaptation of a generic ASR system to mitigate the errors in the automated UK court transcription services. We propose to minimise legal-specific errors by fine-tuning off-the-shelf ASR systems with a custom language model (CLM) trained on legal documents as well as 139 hours of human-edited transcriptions of UK Supreme Court hearings. We also employ NLP techniques to automatically build a custom vocabulary of common multi-word expressions and word n-gram collocations that are critical in court hearings. We infuse our custom vocabulary to the CLM at transcription time. In this research, we evaluate the benefits of our proposed domain adaptation methods by comparing the WER of the CLM output with two off-the-shelf ASR systems: AWS Transcribe (commercial) and the OpenAI Whisper model (open-source) [6]. We also compare the general improvement in the ASR system's ability to correctly transcribe legal entities with and without adopting our proposed methods. In addition we discuss the transcription time with different ASR settings since transcription time is critical for the commercial pipeline implemented by the industrial partner of the project.

2. Related Work

Automatic speech recognition (ASR) models convert audio input to text and they have optimal performance when used to transcribe data which is similar to the one they were trained on. However, performance degrades when there is a mismatch between the data used for training and the one that is being transcribed. Additionally, some types of audio material are intrinsically harder for speech recognition systems to transcribe. In practice, this means that speech recognition system performance degrades when, for example, there is background noise [7], non-native accents [8, 9], young or elderly speakers [8], or a shift in domain [10].

Performance degradation is typically mitigated by adapting or fine-tuning ASR models towards the domain of the targeted data by using a domain-specific dataset [11, 12, 13]. Some methods for domain adaptation adopt NLP techniques such as using machine translation models to learn a mapping from out-of-domain ASR errors to in-domain terms [14]. An alternative approach is to build a large ASR model with a substantially varied training set, so that the model is more robust to data shifts. An example of this latter approach is the recently released OpenAI Whisper model which is trained on 680k hours of diverse domain data to generalise well on a range of unseen datasets without the need for explicit adaptation [6].

Moreover, ASR models are evaluated using Word Error Rate (WER), which treats each incorrect word equally. However, ASR models do not perform equally on different categories of words. Performance is worse for categories like names of people and organisations as compared to categories like numbers or dates [15]. ASR research targeted improving specific errors such as different named entities using NLP techniques [16, 17].

In this paper, we propose simple techniques to improve the effect of the domain mismatch between a generic ASR model and the specialised domain of British court room hearings. Our proposed method, improves both the system's WER rate as well as its ability to capture

case-specific terms and entities. In the next section, we present the setup of our experiments and the evaluation results.

3. Experiment Setup

Figure 1 illustrates our proposed pipeline to improve the ASR system performance by legal domain-adaptation techniques. First, we build a custom language model (CLM) by fine-tuning the base AWS ASR system, using two types of training data: 1) textual data from the legal domain, 2) a corpus of human-generated legal transcriptions. Second, we use NLP techniques to extract domain-specific phrases and legal entities from the in-domain data to create a vocabulary list. We use both the CLM and the vocabulary list for transcribing legal proceedings. The following sections explain details of our experiment where we implemented this pipeline on the AWS Transcribe base model. We compare the performance of our CLM model with different settings to AWS Transcribe base ASR system and OpenAI Whisper open-source ASR system when transcribing $\approx$ 12 hours of UK Supreme Court Hearings.

3.1. Fine-tuning the ASR system

AWS Transcribe improves the quality of speech recognisers by employing an architecture known as the recurrent neural network-transducer (RNN-T) [18]. It is an end-to-end model for automatic speech recognition (ASR) which has gained popularity in recent years as a way to fold separate components of a conventional ASR system (i.e., acoustic, pronunciation and language models) into a single neural network [19]. The AWS Transcribe platform allows the fine-tuning of their ASR architecture via building custom language models to improve transcription accuracy for domain-specific speech. Creating a robust custom language model requires a significant amount of text data, which must contain spoken domain-specific vocabulary.

For training our CLM, we use two datasets from the legal domain. The first is Supreme Court written judgements of 43 cases consisting of 3.26M tokens scraped from the official site of the UK Supreme Court[3]. The second dataset consists of $\approx$ 81 hours of gold-standard transcripts of 10 Supreme Court hearings. The gold-standard transcripts are created by post-editing the AWS Transcribe output of the court hearings by a team of legal professionals using a specially designed interface. We use both datasets to train a CLM that fine-tunes the base AWS ASR architecture to the UK legal domain.

3.2. Phrase Extraction Model

For the vocabulary list, we use a dataset of $\approx$ 139 hours of gold-standard transcriptions of Supreme Court hearings along with the supreme court judgements used for training the CLM. To extract the vocabulary from this dataset, we implement two methods. First, we use this dataset to train a phrase detection model that collocates bigrams based on Pointwise Mutual Information (PMI) scoring of the words in context [20]. PMI is a measure of association between words; it compares the probability of two words occurring together to what this probability would be if the two words were independent. We train the collocation model using the Gensim Python library with a minimum score threshold for a bigram to be taken into account set to 1 and with PMI as the probability scoring method [21]. The collocation model is trained on the textual data of the Supreme Court transcriptions and the supreme court judgements. The model is then used to extract a list of most common bigrams in this dataset. Figure 2 shows an example of the type of common phrases extracted by our collocation model along with their frequencies. As can be seen from the figure, the extracted phrases include frequent legal terms (highlighted in blue) as well as named entities such as names of institutions and persons (highlighted in yellow) which are specific of the Supreme Court cases included in the training corpus.

```
l7267.38  khan_toddler
l7267.38  illustrative_nonexhaustive
l7267.38  alba_cora
l7267.38  actus_reus  (legal )
l6576.68  sri_lanka
l5348.78  swimming_pool
l5348.78  gillette_industries (institution)
l5348.78  colonel_karuna
l5348.78  blank_cheque
l4095.82  et_cetera
l4095.82  bona_fide (legal)
l3813.9  pottu_amman (names)
l3813.9  mens_rea (legal)
l3813.9  dumbarton_oaks
l3156.1  al_qaeda
l2950.53  wayne_tank
l1937.94  prima_facie
```

Figure 2: Example of Common Collocations Extracted by the Phrase Extraction Model

The second method we employ to create a list of custom vocabulary is to identify named entities in our dataset. For this purpose, we use Blackstone[4], an NLP library for processing long-form and unstructured legal text capable of identifying legal entities. The list of legal entities includes: Case Name, Court Name, Provision (i.e. a clause in a legal instrument), Instrument (i.e. a legal

[3]https://www.supremecourt.uk/decided-cases/

[4]https://research.iclr.co.uk/blackstone

Table 2
Average WER and Transcription Time

Model	WER Case1	WER Case2	WER Average	Transcription Time
AWS base	8.7	16.2	12.3	85 mins
CLM1	8.5	16.5	12.4	77 mins
CLM2	**7.9**	15.5	**11.6**	77 mins
CLM2+Vocab	**7.9**	15.6	**11.6**	132 mins
CLM2+Vocab2	8.0	15.6	11.7	112 mins
Whisper	9.6	**15.3**	12.4	191 mins

term of art) and Judge. We concatenated this Blackstone entity list with the spaCy v3.4 library list of non-legal entities such as: Cardinals, Persons and Dates. The results of applying our domain-adaptation methods for the transcription of 2 Supreme Court case hearings consisting of 12 hours is explained in the next section.

4. Results

Table 2 shows the WER scores and WER average score for the 2 transcribed cases with different CLM system settings, as well as, for the two baseline systems: the AWS Transcribe (AWS base) and Whisper. The different CLM settings are as follows:

1. **CLM1** is trained on only the texts of the Supreme Court judgements.
2. **CLM2** is trained on both the judgements and the gold-standard transcripts.
3. **CLM2+Vocab** uses CLM2 for transcription plus the global vocabulary list extracted by our phrase detection model.
4. **CLM2+Vocab2** uses CLM2 for transcription plus the legal entities vocabulary list extracted by Blackstone and spaCy v3.4 library.

As can be seen in Table 2, the ASR performance is consistently better with the CLM models than with the generic ASR systems for the two transcribed cases. CLM2 model, trained on textual data (i.e. the written judgements) and gold-standard court hearing transcriptions, outperforms AWS base and Whisper with a 9% and 8% WER improvement, respectively. Moreover, we observe around 9% improvement in average WER score over the two generic models when concatenating the list of legal phrases that is extracted by our phrase detection model with the CLM2 system. While ASR error correction indicates an improved transcription quality with our proposed domain adaptation methods, we also evaluated the ASR systems performance with specific errors such as legal entities and terms.

Table 3 shows the average ratio of correctly transcribed legal entities in the two studied court room hearings. We compare the performance of CLM2 infused with the legal terms list (CLM2+Vocab) to the two generic ASR systems. The ratios in Table 3 indicate that CLM2+Vocab is generally more capable of transcribing legal-specific terms than the other two models. It is also better at transcribing critical legal entities such as Provisions.[5] Such legal terminology needs to be accurately transcribed. Our CLM2 model with legal vocabulary demonstrates better reliability in transcribing these terms.

A similar trend is evident with the legal entity Judge which refers to the forms of address used in British court rooms (e.g. 'Lord Phillips', 'Lady Hale'). This entity is typically repeated in court hearings whenever a barrister or solicitor addresses the court. We see that both the generic ASR systems perform badly on this category with ratios of 0.66 and 0.69, respectively. On the other hand, we observe a significant improvement in correctly transcribing this type of entities by the CLM2+Vocab with a ration of 0.84 correct transcriptions. Appendix A shows an example of the output of the AWS base ASR model without our domain-adaptation methods compared to the output of the CLM correcting the mistakes. The transcription errors (highlighted yellow) in the base output includes legal jargon, legal terms and named entities. The errors are corrected by our CLM model (corrections are highlighted in blue).

In addition to evaluating the output of the ASR engines, we also recorded the time required to produce the transcription. The models based on AWS were run in the cloud using the Amazon infrastructure. Whisper was run on a Linux desktop with an NVIDIA GeForce RTX 2070 GPU with 8G VRAM. For all the experiments, the medium English-only model was used. As expected the fastest running time is obtained using the AWS base model. Running the best performing model increases the time by 155%, whilst Whisper more than doubles it. Trade-off between running time and the level of domain-specific accuracy is a variable parameter that can be determined based on the transcription purpose and the end-user needs defined by our project's commercial partner.

[5] A Provision, a statement within an agreement or a law, typically consists of alphanumeric utterances in British court hearings (e.g. 'section 25(2)(a)-(h)' or 'rule 3.17').

Table 3
Ratio of Correctly Captured Legal Entities by the ASR Systems

Entity	AWS BASE	Whisper	CLM2+vocab
Judge	0.66	0.77	0.84
CASE NAME	0.69	0.85	0.71
Court	0.98	1	0.93
Provision	0.88	0.95	0.97
Cardinal	1	0.97	1

5. Conclusion

In this paper, we present a study which shows the effect of domain adaption methods on improving the off-the-shelf ASR system performance in transcribing a specialised domain such as British court hearings. We optimised the performance of the ASR system by training an ASR custom language model on gold-standard legal transcripts and textual data from the legal domain. We also trained a phrase detection model to incorporate extracted list of data-specific bigram collocations at transcription time. We evaluated the ASR quality improvements both in terms of average WER and ratio of correctly transcribed legal-specific terms. We observe significant gains in the ASR transcription quality by our domain adaptation techniques. For commercial use of ASR technologies, improving error rate in general and transcription quality of critical legal terms in particular would minimise manual post-editing effort and hence save both time and money. We plan to evaluate the impact of different configurations proposed in this paper on the editors' postediting effort.

In the future, we will expand to record data from a variety of accents to address another axis of degradation in British audio procedures different than the Supreme Court hearings which are mostly a homogeneous group of speakers. We will also explore the ability to use NLP topic modelling techniques to connect legal entities that were crucial in a court's case decision.

References

[1] E. Elwany, D. Moore, G. Oberoi, Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding, arXiv preprint arXiv:1911.00473 (2019).

[2] J. J. Nay, Natural Language Processing for Legal Texts, DOI=10.1017/9781316529683.011, Cambridge University Press, 2021, p. 99–113.

[3] E. Mumcuoğlu, C. E. Öztürk, H. M. Ozaktas, A. Koç, Natural language processing in law: Prediction of outcomes in the higher courts of turkey, Information Processing & Management 58 (2021) 102684.

[4] J. Frankenreiter, J. Nyarko, Natural language processing in legal tech, Legal Tech and the Future of Civil Justice (David Engstrom ed.) (2022).

[5] G. Sturge, Court statistics for England and Wales, Technical Report, House of Commons Library, 2021. URL: https://commonslibrary.parliament.uk/research-briefings/cbp-8372/.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, OpenAI (2022).

[7] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, et al., CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings, 2020.

[8] S. Feng, O. Kudina, B. M. Halpern, O. Scharenborg, Quantifying bias in automatic speech recognition, arXiv preprint arXiv:2103.15122 (2021).

[9] Y. Zhang, Mitigating bias against non-native accents, Delft University of Technology (2022).

[10] L. Mai, J. Carson-Berndsen, Unsupervised domain adaptation for speech recognition with unsupervised error correction, Proc. Interspeech 2022 (2022) 5120–5124.

[11] Z. Huo, D. Hwang, K. C. Sim, S. Garg, A. Misra, N. Siddhartha, T. Strohman, F. Beaufays, Incremental layer-wise self-supervised learning for efficient speech domain adaptation on device, arXiv preprint arXiv:2110.00155 (2021).

[12] H. Sato, T. Komori, T. Mishima, Y. Kawai, T. Mochizuki, S. Sato, T. Ogawa, Text-Only Domain Adaptation Based on Intermediate CTC, Proc. Interspeech 2022 (2022) 2208–2212.

[13] S. Dingliwa, A. Shenoy, S. Bodapati, A. Gandhe, R. T. Gadde, K. Kirchhoff, Domain prompts: Towards memory and compute efficient domain adaptation of ASR systems, https://tinyurl.com/2a9jp88t, 2022.

[14] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, F. Metze, Asr error correction and domain adaptation using machine translation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6344–6348.

[15] M. Del Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong,

P. Zelasko, M. Jetté, Earnings-21: a practical benchmark for asr in the wild, arXiv preprint arXiv:2104.11348 (2021).

[16] H. Wang, S. Dong, Y. Liu, J. Logan, A. K. Agrawal, Y. Liu, ASR Error Correction with Augmented Transformer for Entity Retrieval., in: Interspeech, 2020, pp. 1550–1554.

[17] N. Das, D. H. Chau, M. Sunkara, S. Bodapati, D. Bekal, K. Kirchhoff, Listen, Know and Spell: Knowledge-Infused Subword Modeling for Improving ASR Performance of OOV Named Entities, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 7887–7891.

[18] A. Graves, Sequence transduction with recurrent neural networks, arXiv preprint arXiv:1211.3711 (2012).

[19] J. Guo, G. Tiwari, J. Droppo, M. Van Segbroeck, C.-W. Huang, A. Stolcke, R. Maas, Efficient minimum word error rate training of rnn-transducer for end-to-end speech recognition, arXiv preprint arXiv:2007.13802 (2020).

[20] G. Bouma, Normalized (pointwise) mutual information in collocation extraction, Proceedings of GSCL 30 (2009) 31–40.

[21] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50. http://is.muni.cz/publication/884893/en.

A. Appendix: Examples of ASR ouput with and without domain-adaptation

BASE MODEL

Names of people
"Mrs Agbaje"

Legal jargon
"my learned"

with my lonely junior Miss Eleanor Harris. I appear on behalf of Mrs Ebadi, the appellate um, and then your friend Mr Timothy Scott, Queen's counsel, and Mr Peter Mitchell appear on behalf of Mr and Party, the respondent to the appeals. If I may, I will have dropped the Family Law convention of calling the party's husband and wife. The principal issue arising on this appeal is, in what circumstances is it appropriate for the English court to make a further financial order when a foreign court has already divorced the spouses and made a financial order? At first blush, it may seem unusual that courts in two different countries can, at different times make financial orders are rising out of the same cause of action, namely, the party's divorced. The power in English court to make a further financial order derives from Part three of the Matrimonial and Family Proceedings Act 1984 and it was intended by its frame as the law commissioners to remit financial hardship arising from two distinct circumstances. I suppose I should make it absolutely plain that this is the work of my predecessor. Yes, the Law Commission. I had no hand in part three at all. Yes, we saw the authors that listed in the report. The two circumstances that the law commissioners were addressing were these firstly, when a foreign court had made no financial order at all and secondly, where the foreign Court had made a financial order. But that order was inadequate and this appeal is concerned only with the second category. It is a condition precedent to the exercise of powers under Part three that the wife has to show in the circumstances of her case that a serious injustice has arisen. The serious injustice that this wife relies upon is that after a marriage of 32 years which produced five Children of the founded and where the assets were around £700,000 the Nigerian court in June of 2005 awarded her a lump sum of £21,000 and a life interest in the house on Tin Can Island in Lagos. This very modest award produced a very significant disparity or discrepancy in the allocation of the assets between the parties

Grammar mistakes
"parties divorced"

Legal Terms
"its framers"

Vocabulary
"children of the family

CLM MODEL Correcting Mistakes

Names

Legal jargon

with my learned Junior Miss Eleanor Harris. I appear on behalf of Mrs Agbaje. The appellate, um, my learned friend, Mr Timothy Scott, Queen's counsel, and Mr Peter Mitchell appear on behalf of Mr M. Party, the respondent to the appeal. If I may, I will have dropped the Family Law Convention of calling the parties husband and wife. The principal issue arising on this appeal is in what circumstances is it appropriate for the English court to make a further financial order when a foreign court has already divorced the spouses and made a financial order? At first blush, it may seem unusual that courts in two different countries can at different times make financial orders arising out of the same cause of action, namely, the parties divorced. The power in the English court to make a further financial order derives from Part three of the Matrimonial and Family Proceedings Act 1984 and it was intended by its framers, the law commissioners, to remit financial hardship arising from two distinct circumstances. I suppose I should make it absolutely plain that this is the work of my predecessor. Yes, the Law Commission. I had no hand in part three at all. Yes, we saw the authors listed in the report. The two circumstances that the law commissioners were addressing were these firstly, when a foreign court had made no financial order at all and secondly, where the foreign Court had made a financial order. But that order was inadequate and this appeal is concerned only with the second category. It is a condition precedent to the exercise of powers under Part three that the wife has to show in the circumstances of her case that a serious injustice has arisen. The serious injustice that this wife relies upon is that after a marriage of 32 years which produced five Children of the family and where the assets were around £700,000 the Nigerian court in June of 2005 awarded her a lump sum of £21,000 and a life interest in the house on Tin Can Island in Lagos. This very modest award produced a very significant disparity or discrepancy in the allocation of the assets between the parties"

Grammar

Legal terms

Vocabulary