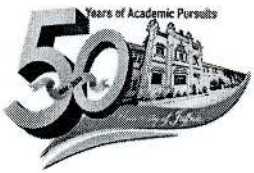
பேராசிரியர் சுவாமிநாதன் சுசீந்திரராஜா
நினைவுப் பேருரை

கணினிமொழியியலின் இன்றைய வளர்ச்சியும் தமிழும்

பேருரை வழங்குபவர்
பேராசிரியர் ந. தெய்வ சுந்தரம் (தமிழ்நாடு)

மொழியியல் மற்றும் ஆங்கிலத்துறை
யாழ்ப்பாணப் பல்கலைக்கழகம்
09.10.2024



மொழியியல் மற்றும் ஆங்கிலத்துறை

யாழ்ப்பாணப் பல்கலைக்கழகம்

பேராசிரியர் சுவாமிநாதன் சுசீந்திரராஜா

நிணைவுப் பேருரை - 2024

கணினிமொழியியலின் இன்றைய வளர்ச்சியும் தமிழும்

பேருரை வழங்குபவர்

பேராசிரியர் ந. தெய்வ சுந்தரம் [தமிழ்நாடு]

09.10.2024

பேராசிரியர் சு.சுசீந்திரராஜா

(1933 - 2019)

மேற்கு நாடுகளிலும் இந்தியாவிலும் இமாலய வளர்ச்சிகண்ட மொழியியல் துறையானது இலங்கையில் கற்பிக்கப்படுவதற்குப் பெரும்பங்காற்றியவர்களுள் ஓய்வுபெற்ற தகைமைசார் பேராசிரியர் சுவாமிநாதன் சுசீந்திரராஜா அவர்கள் என்றும் மதித்துப் போற்றப்படவேண்டிய பெருந்தகையாய் விளங்குகிறார். பேராசிரியர் அவர்கள் அறிவியல் ரீதியிலான அணுகுமுறையில் மொழியை ஆராயும் புதிய ஒரு நோக்கை அறிமுகப்படுத்தியதுடன் அத்துறைசார்ந்த ஆராய்ச்சியிலும் முன்னோடியாக விளங்கியவர் என்பது அவசியம் குறிப்பிடப்படவேண்டிய ஒன்றாகும்.

தமிழ் மொழியியல் துறையில் ஒலியன்கள், உருபன்கள் ஆகியவற்றின் தொழிற்பாட்டினைப் பற்றி மிகவிரிவாக ஆராய்ந்துள்ளார். பேச்சொலிகள், ஒலியன்கள் போன்றவற்றின் விபரிப்பு மற்றும் பயன்பாடு பற்றி அறிவியல் ரீதியில் ஆராய்ந்து துல்லியமான கணிப்பினை வெளிப்படுத்தியமை சிறப்பான விடயமாகும். யாழ்ப்பாணத்தமிழ்க் கிளைமொழியில் உச்சரிக்கப்படும் பேச்சொலிகளின் ஒப்பற்ற தனித்துவப் பண்புகளை அவர் ஆராய்ந்த விதம் அறிவியற் செழுமையின் மேன்மையாய்த் திகழ்கிறது. சமூகத்தில் நிலவும் அடுக்கமைவை அடிப்படையாய்க் கொண்டு சொற்கள் கையாளப்படும் நுணுக்கங்களை ஆராய்ந்து எடுத்தியம்பியுள்ளார். ஆங்கில மொழியிலேயே அதிகளவு வெளியிடப்பட்ட மொழியியல் துறைசார்ந்த விடயப்பரப்புகளைத் தமிழ் மொழியில் எழுதியதன் மூலம் தனது நூல்களினூடாக மாணவர்களும் ஆராய்ச்சியாளரும் சமூக ஆர்வலரும் மொழியியற் புலத்தை நன்கு அறியும்படி செய்தமை அன்னாரது மேன்மைமிகு பணிகளில் ஒன்றாகும். சமூகத்துக்கும் மொழிக்குமிடையில் நிலவும் நெருங்கிய தொடர்பினைப் புதிய நோக்கில் ஆராய்வதற்கான ஊக்கத்தை வலுப்படுத்தினார். அவரது புலமைவாய்ந்த ஆய்வுப் பணிகளினால் உலகம் போற்றும் தகைமையெய்தினார்.

அகராதி உருவாக்கத்திலும் பேராசிரியர் அவர்கள் அளப்பெரும் பங்காற்றியுள்ளமை விதந்து போற்றுதற்குரியது. மொழியியற் துறையோடு தொடர்புபட்ட கலைச்சொற்களை உள்ளடக்கிய கலைச்சொல் அகராதி ஒன்றினையும் தொகுக்கும் அரியமுயற்சியில் ஈடுபட்டார். பேராசிரியர் *R. E. Asher* அவர்களோடு இணைந்தும் தமிழ் மொழியியல் சார்ந்த ஆய்வினை மேற்கொண்டுள்ளார். இந்தியத் தமிழ், இலங்கைத் தமிழ்

ஆகியவற்றை ஒப்பிட்டு ஒலியன், உருபன் நிலைகளில் தென்பட்ட அமைப்புசார் வேறுபாடுகளைத் துல்லியமாக ஆராய்ந்துள்ளார். 'Jaffna Tamil (Phonology and Morphology)', 'Studies in Srilankan Tamil Linguistics and Culture', 'An Introduction to Spoken Tamil' CL போன்ற நூல்களில் அறிவியல் ஆழம் செறிந்தகருத்துச் செழுமையைப் புரிந்துகொள்ளலாம்.

சிங்கள மாணவர் இலங்கையின் இரண்டாம் மொழியாகிய தமிழைக் கற்பதற்கேற்றவகையில் தமிழ் மொழியில் பயன்படுத்தப்படும் ஒலிகள்,ஒலியன்கள், உருபன்கள் இவாக்கிய ஆக்கம் இகட்டுரை ஆக்கம் போன்ற விடயங்களைப் பற்றித் துல்லியமாக ஆராய்ந்து தனது'An Introduction to Spoken Tamil' நூலில் விபரமாகக் குறிப்பிட்டுள்ளார். இரண்டாம் மொழியாகத் தமிழ் கற்கவிரும்பும் மாணவர்களுக்கு இது ஒரு வரப்பிரசாதமாய்த் திகழ்கிறதெனலாம். சொற்களுக்கும் அவை தெரியப்படுத்தும் பொருளுக்குமிடையிலான உள்ளார்ந்த தொடர்பினை விளக்கிய பாங்கு உன்னதமானது. ஆழ்ந்த அறிவியற் புலமையின் சிகரமாய் மிளிரும் தகைமைசார் பேராசிரியர் சுவாமிநாதன் சுசீந்திரராஜா அவர்களது மேன்மை பொருந்திய கல்விசார் பணிகளைப் பேரன்போடும் உளம் நிறைந்த நன்றியுணர்வோடும் நினைவு கூர்வதில் மொழியியற் துறை உவப்படைவதுடன் மேன்மேலும் அன்னாரது கல்விப் பணிகளைக் காலத்துக்கேற்றவகையில் முன்னெடுக்கும் பொறுப்புணர்வையும் உளங்கொள்கிறது!

ந.கவிதா

துறைத்தலைவர் - மொழியியல் மற்றும் ஆங்கிலத்துறை

கலைப் பீடம் - யாழ்ப்பாணப் பல்கலைக்கழகம்.

09.10.2024

யாழ்ப்பாணப் பல்கலைக் கழகத்தின் மொழியியல் மற்றும் ஆங்கிலத்துறையின் முன்னால் தலைவரும் தகைசால் பேராசிரியருமான சுவாமிநாதன் சுசீந்திரராசா அவர்களின் நினைவுப் பேருரை அவரின் தொன்னூற்றோராவது (91) அகவை நாளான இன்று நடைபெறுவதையொட்டி எமது துறையினர் மிகுந்த மகிழ்ச்சி அடைகின்றனர். யாழ்ப்பாணப் பல்கலைக் கழகம் தனது பொன்விழாவினைக் கொண்டாடிக் கொண்டிருக்கும் வேளையில் இந்நினைவுப் பேருரை நடைபெறுவதைபிட்டு பேருவகைகொள்ளும் தருணமாகக் கருதப்படுகிறது.

இலங்கைப் புலமையாளர்களைப் பொறுத்தவரை தமிழ், ஆங்கிலம், சிங்களம், சமஸ்கிருதம் முதலான மொழிகளில் நன்கு தேர்ந்தவரும், தமிழ் மொழியியலில் (Tamil Linguistics) பல்வேறு வகையான பன்முக ஆய்வுகளை மேற்கொண்டவருமான பேராசிரியர் சு. சுசீந்திரராஜா அவர்களின் நினைவுப் பகிர்வை இந்நாளில் நடாத்துவதுபெருஞ் சிறப்பாகும்.

இலங்கையில் உள்ள தமிழ், சிங்கள மொழியியலாளர்கள் மத்தியில் நவீன மொழியியலையும் தமிழ் இலக்கண இலக்கியங்களையும் துறைபோகக் கற்று உலகளாவிய ரீதியில் அங்கீகாரம் பெற்ற ஒரு தமிழ் மொழியியலாளராகப் பேராசிரியர் சு. சுசீந்திரராஜா அவர்கள் திகழ்கின்றார்கள். கடந்த ஐம்பது ஆண்டுகளுக்கு மேலாக இலங்கைத் தமிழின் அமைப்பைப் பற்றியும், அதன் சமய, சமூக பண்பாடுகள் பற்றியும் ஆழமாகச் சிந்தித்தும், ஆராய்ந்தும் அவை பற்றி எழுதியும் தனது ஆற்றலை உலகுக்கு வெளிப்படுத்தினார்கள். அன்றிலிருந்து இன்றுவரை இலங்கையில் தமிழ் மொழி - மொழியியல் பற்றிய ஆய்வில் அவருடைய பங்களிப்பே மிகப் பெறுமதி வாய்ந்ததும் தலையாயதும் என்று நாம் சிறப்பாகக் குறிப்பிடலாம்.

ஈழத்துத் தமிழ்ச் சொற்கள் அகராதியின் தேவையை உணர்ந்த பேராசிரியர் தமிழியல் சார் சிந்தனைத் துளிகள் முதலான நூல்கள் மூலமாக அதுதொடர்பான பல முயற்சிகளை மேற்கொண்டார்கள். துறைசார்ந்தும் வெளியிலும் பல முன்னெடுப்புக்களை மேற்கொண்டு அதில் வெற்றியும் கண்டார். குறிப்பாக சென்னைப் பல்கலைக்கழக அகராதிகள், கிரியாவின் அகராதி போன்றவை இலங்கைத்தமிழ் சொற்களை ஏற்றுக் கொண்டு தம்முள் உள்வாங்ககாரணமாக இருந்தவர்கள் பேராசிரியரும் அவர்தம் ஆசிரியமாணவர்களுமாவர். ஆய்வாளர் என்ற வகையிலும் சிறந்த மொழியியலாளர் என்றவகையிலும் தேசியமட்டத்திலும் சர்வதேசமட்டத்திலும் பெருமதிப்பும் பாராட்டும் கௌரவமும் இவருக்குக் கிடைக்கப் பெற்றன. பல துறைகள் சார்ந்த அறிவும் அனுபவமும் பெற்றிருந்தபோதிலும் அறிவியல் ரீதியாகவும், கிளைமொழி ரீதியாகவும் இலங்கைத்தமிழ், யாழ்ப்பாணத்தமிழ், இந்தியத்தமிழ் மூன்றையும் ஆராய்ந்து பல ஆய்வுக் கட்டுரைகளை வெளியிட்டிருக்கிறார்கள்.

இவற்றின் மூலம் இலங்கைத் தமிழ்மொழி வழக்கு தூய்மையானதும், பிறமொழிக் கலப்பற்றது என்பதையும், சங்ககால வழக்காறுகள் பல இன்றும் பேணப்பட்டு வருவதையும் பேராசிரியர மிகத் தெளிவாகவும் ஆணித்தரமாகவும் உறுதிபட நிரூபித்துள்ளார்கள்.

இவ்வாறு தமிழ் மொழியியல் ஆயிவின் முன்னோடியாக விளங்கியவரும் தமிழ்ப் பண்பாட்டைப் பேணி, தொடர்ந்து முன்னெடுத்துச் சென்ற பேராசிரியர் சு.சுசீந்திரராஜா பற்றிய நினைவுப் பேருரையை மொழியியல் ஆங்கிலத்துறை விரிவுரையாளர்களும், மாணவர்களும் இந்தியாவின் சென்னைப் பல்கலைக்கழக கணினிமொழியியல் **பேராசிரியரான. ந. தெய்வசுந்தரம்** அவர்களின் மூலம் இன்னாளில் குறிப்பாகப் பொன்விழா ஆண்டில் நடாத்துவதில் பெருமகிழ்ச்சி அடைகின்றார்கள்.

பேராசிரியர். சுபதினி ரமேஸ்

மொழியியல் மற்றும் ஆங்கிலத்துறை,

யாழ்ப்பாணப் பல்கலைக்கழகம்,

09.10.2024.

ஐம்பது ஆண்டுகால வரலாற்றுச் சிறப்புடைய யாழ்ப்பாணப் பல்கலைக்கழகத்தின் துணைவேந்தர் அவர்களே! பேராசிரியப் பெருமக்களே! ஆய்வாளர்களே! மாணவ நண்பர்களே! தமிழ் ஆர்வலர்களே! அனைவருக்கும் எனது அன்பு கலந்த வணக்கத்தைத் தெரிவித்துக்கொள்கிறேன்.

மறைந்த பேராசிரியர் சு. சுசீந்திரராஜா அவர்களைப் பெருமைப்படுத்தும் வகையில் ஒரு நினைவுப் பேருரையை உங்கள் முன்னே நிகழ்த்த எனக்கு ஒரு அரிய வாய்ப்பை அளித்த யாழ்ப்பாணப் பல்கலைக்கழகத்திற்கும் , மொழியியல் - ஆங்கிலத் துறைக்கும் எனக்கு மனமார்ந்த நன்றியைத் தெரிவித்துக்கொள்கிறேன்.

பேராசிரியர் அவர்களுடன் நேரிடையான தொடர்புகொள்ளும் வாய்ப்பு எனக்குக் கிடைத்தது இல்லை என்பதை வருத்தத்துடன் தெரிவித்துக்கொள்கிறேன். ஆனால் அவரின் சிறப்புப் பற்றி எனது மொழியியல் பேராசிரியர்கள் என்னிடம் பாராட்டிக் கூறியுள்ளார்கள். பேராசிரியர்கள் அவர்கள் மொழியியல் கற்ற தமிழ் நாட்டின் அண்ணாமலைப் பல்கலைக்கழக மொழியியல் துறையில்தான் நானும் மொழியியல் கற்றேன் என்பதில் எனக்கு மகிழ்ச்சி. பேராசிரியர் அவர்கள் தமிழியல், மொழியியல் துறைகளில் வியக்கத்தக்க ஆய்வுகளை மேற்கொண்ட பெரும்பேராசிரியர்கள் மு. வரதராசனார், தெ.பொ. மீனாட்சி சுந்தரனார், சு. அகத்தியலிங்கம் போன்றவர்களின் தமிழ், மொழியியல் மாணவர். மொழியியல் படிப்பை முடித்தபின்னர் பேரா. தெ. பொ. மீனாட்சி சுந்தரனார் அவர்களின் வேண்டுகோளின் அடிப்படையில் அண்ணாமலைப் பல்கலைக்கழகத்தின் மொழியியல் துறையில் மூன்று ஆண்டுகள் பேராசிரியராகப் பணியாற்றியுள்ளார்.

இலங்கைத்தமிழ் பற்றிய அவருடைய முனைவர் பட்ட ஆய்வேட்டிற்குப் புறத்தேர்வாளர்களாகச் செயல்பட்ட அமெரிக்கப் பேராசிரியர் ஜேம்ஸ் கெயிர் அவர்களும் இங்கிலாந்தைச் சேர்ந்த பேராசிரியர் ஏ. இ. ஆஷர் அவர்களும் ஆய்வேட்டை மிகவும் பாராட்டியுள்ளனர். இரசியாவைச் சேர்ந்த பேராசிரியர் எம். ஆண்டிரனோவ் அவர்கள் ஒரு தனிமடலின் மூலமாகவும் பாராட்டியுள்ளார் என்பது மகிழ்ச்சிக்குரிய ஒரு செய்தி.

இலங்கையில் கொழும்புப் பல்கலைக்கழகத்திலும், பின்னர் களனிப் பல்கலைக்கழகத்திலும் மொழியியல் துறையில் பேராசிரியராகப் பணியாற்றியுள்ளார். பின்னர் யாழ்ப்பாணப் பல்கலைக்கழகத்தின் தமிழ்த்துறைத் தலைவராகப் பொறுப்பேற்றுத் தமது ஆசிரியப் பணியைத் தொடர்ந்தார். அங்கே மொழியியல் துறையையும் நிறுவினார். 1980-இல் ஓராண்டு இங்கிலாந்து சென்று, பேராசிரியர் ஆஷருடன் இணைந்து மொழியியல் ஆய்வை மேற்கொண்டார். 1998- ஆம் ஆண்டில் பணி ஓய்வு பெற்றார்.

"தமிழியல்சார் சிந்தனைத்துளிகள்" என்ற பெயரில் பேராசிரியர் தமது ஆய்வுகளை இரண்டு தொகுதிகளாக வெளியிட்டுள்ளார். பேராசிரியர் சுசீந்திரராஜா அவருடைய பிறந்தநாளைக் கொண்டாடும் வகையில் மாணவர்கள் அவருடைய கட்டுரைகளைத் தொகுத்து "Studies in Srilankan Tamil Linguistics and Culture" என்ற ஒரு தொகுதியை வெளியிட்டுள்ளார்கள். "அடிப்படைத் தமிழ் இலக்கணம்" என்ற பெயரில் இன்றைய எழுத்துத்தமிழுக்கு அனைவரும் பாராட்டுகிற வகையில்ஒரு நூலை எழுதியுள்ள மொழியியல் பேராசிரியர் எம். ஏ. நு:மான் அவர்கள் பேராசிரியருடைய மாணவரே. மறைந்த பேராசிரியர் சு. சுசீந்திரராஜா அவர்கள் தன்னைப் பிரபலப்படுத்துவதில் ஆர்வம் காட்டாது, தமிழ்மொழி ஆராய்ச்சியையே தமது முதன்மைப் பணியாகக் கருதி அமைதியாகச் செயற்பட்டவர் என்பது அவரது தனிச்சிறப்பு.

எனது உரையின் தொடக்கத்தில் மொழி பற்றிய மூன்று வகையான ஆய்வுப் பிரிவுகளைத் தெளிவுபடுத்திக்கொள்ள விரும்புகிறேன். முதலாவது,ஒரு இயற்கைமொழியின் இலக்கணம்; இரண்டாவது, இயற்கைமொழிகள் பற்றியபொது அறிவியலான மொழியியல்; மூன்றாவது, இயற்கைமொழிகளைக் கணினியானது புரிந்துகொண்டு, மனிதர்கள் மேற்கொள்கிற மொழிச் செயல்பாடுகளை கணினி மேற்கொள்வதற்கான வழிமுறைகளைப் பற்றிய அறிவியலான கணினிமொழியியல்.

இலக்கணம்

எந்தவொரு இயற்கைமொழிக்கும் ஒரு கட்டமைப்பு உண்டு. குறிப்பிட்ட மொழியின் அகராதிச் சொற்களின் பொருண்மை, அச்சொற்களைக்கொண்டு மொழிவழிச் செயல்பாடுகளை மேற்கொள்வதற்கான - கருத்தாடலை மேற்கொள்வதற்கான - வரைமுறைகள் ஆகியவைபற்றிய ஒரு பொது உடன்பாடு உண்டு. அவ்வாறு இல்லையென்றால் ஒருவர் கூறுவதை மற்றொருவர் புரிந்துகொள்ளமுடியாது. ஒரு பொருண்மையை வெளிப்படுத்த இந்தக் குறிப்பிட்ட சொல்தான் பயன்படுத்தவேண்டும், குறிப்பிட்ட தொடரமைப்புத்தான் பயன்படுத்தப்படவேண்டும் என்பதில் தெளிவான விதிமுறைகளை அந்த மொழியின் சமுதாயம் உருவாக்கி வைத்திருக்கும். இதையே நாம் அந்த மொழியின் இலக்கணம் என்று கூறுகிறோம். இலக்கணம் இல்லாமல் எந்தவொரு மொழியும் - பழங்குடி மக்களின் மொழி, பேச்சுமொழி ஆகியவை உட்பட - நிலவமுடியாது.

இலக்கணநூல்

ஒரு மொழியின் அமைப்பையும் செயல்பாட்டையும் ஆராய்ந்து விளக்கும் அறிஞர்கள் தங்கள் ஆய்வுகளை இலக்கணநூல்களாக முன்வைக்கின்றனர். இலக்கணநூல்கள் அவற்றை உருவாக்கிய அறிஞர்களின் தனிப்பட்ட விருப்பு வெறுப்புகளைச்சார்ந்த ஒரு அகவயப் படைப்பு இல்லை. அறிஞர்கள் ஒரு மொழியின் இலக்கணத்தை உருவாக்கவில்லை; மாறாக, அந்த மொழியில் புறவயமாக நிலவுகிற விதிமுறைகளை விருப்புவெறுப்பு இல்லாமல் அவர்கள் கண்டறிந்து முன்வைக்கிறார்கள். ஒரு மொழிக்கு இலக்கணநூல் உருவாக்கப்பட்டிருந்தால்தான் அம்மொழியில் இலக்கணம் நிலவுகிறது என்று கருதுவது தவறு. பல மொழிகளுக்கு அவற்றின் இலக்கணங்கள் முறையாக அறிஞர்களால் ஆராயப்பட்டு முன்வைக்கப்படாமல் இருக்கலாம்.

ஆனால் அதற்காக அந்தமொழிகளுக்கு இலக்கணங்கள் இல்லை என்று கருதிவிடக்கூடாது. தமிழுக்குத் தொல்காப்பியம், நன்னூல், வடமொழிக்குப் பாணினீயம் போன்றவை இந்த வகை இலக்கண நூல்களே ஆகும்.

மொழி ஒப்பீட்டு ஆய்வுகள் (Comparative Grammars)

மனிதசமுதாய வரலாற்றில் இன்றுள்ள வளர்ச்சி நிலவாத ஒரு காலகட்டத்தில், மொழி ஆய்வு என்பது குறிப்பிட்ட மொழிகளின் இலக்கணங்களாகவேநிலவின.அவ்வாறுகுறிப்பிட்ட மொழிக்கான இலக்கணத்தை உருவாக்கிய அறிஞர்கள் சிலர்தங்களது தாய்மொழி தவிர தங்கள் சமுதாயத்திற்கு அருகில் உள்ள பிற சமுதாயங்களின் மொழிகளையும் அறிந்துவைத்திருக்கலாம். அதனால் குறிப்பிட்ட மொழியின் தங்கள் இலக்கணத்தில் ஆங்காங்கேபிற மொழிகளின்பண்புகள்பற்றியும் கூறியிருக்கலாம். தொல்காப்பியர் தமிழுக்கான தமது இலக்கணநூலில் தமிழ்மொழியின் மீதான வடமொழியின் தாக்கம் பற்றி ஆங்காங்கே சில கருத்துக்களை முன்வைத்துள்ளார். இருப்பினும் அவரது படைப்பு தமிழ்மொழிக்கான இலக்கணமே ஆகும்.

உலகவரலாற்றில்தொழிற்புரட்சியின்தாக்கத்தால்வேறுபட்டமொழிகளைச் சார்ந்த வேறுபட்ட சமுதாயங்களிடையே வணிக உறவுகள் வளரத்தொடங்கியதால், ஒன்றுக்குமேற்பட்ட மொழிகளைத் தெரிந்துகொள்ளவேண்டிய தேவை வளர்ந்தது. அதையொட்டி மொழி ஆய்வில் ஆர்வம் உள்ள அறிஞர்களுக்குத் தங்கள் தாய்மொழியைத் தாண்டி, பிற மொழிகளைப்பற்றிய அறிவையும் பெறுவதற்கான வாய்ப்புக்கள் உருவாகின. இதன் பயனாக, அறிஞர்களுக்குத் தங்கள் தாய்மொழிகளைப் பிறமொழிகளோடு ஒப்பிட்டு ஆராய்வதற்கான வாய்ப்புக்கள் தோன்றி நிலவின. அவ்வாறு ஆராயும்போது அவர்களுக்குத் தங்கள் தாய்மொழிகளின் தனிச்சிறப்பான இலக்கணங்களைப்பற்றிய அறிவோடு, மொழிகளுக்கிடையேயான பொதுக்கூறுகளையும் அறிந்துகொள்ளும் வாய்ப்புக்கள் நிலவின. குறிப்பாக, 16-ஆம் நூற்றாண்டைத் தொடர்ந்து மேற்குறிப்பிட்ட பன்மொழி ஆய்வுக்கான வாய்ப்புக்கள் பெருகின. அதன் பயனாக, ஒன்றுக்குமேற்பட்ட மொழிகளை ஒப்பிட்டு ஆய்வுசெய்யும் வாய்ப்பும் (Comparative Grammars), ஒன்றுக்கு மேற்பட்ட மொழிகளின் வரலாற்றில் காணப்படுகிற வளர்ச்சிக்கூறுகளைத் தெரிந்துகொள்ளும் வாய்ப்பும் (Historical Grammars) தோன்றி நிலவின. இவற்றின் பயனாகவே ஒப்பிட்டுமொழியாய்வு, வரலாற்றுமொழியாய்வு நடைபெறத் தொடங்கின.

இராபர்ட் கால்டுவெல் (Robert Caldwell) அவர்களின்திராவிடமொழிகள் ஒப்பியல் இலக்கணம், சர் வில்லியம்ஸ் ஜோன்ஸ் (Sir William Jones) அவர்களின்

இந்தோ-ஐரோப்பிய ஒப்பியல் இலக்கணம், பிரான்சிஸ் பாப் (Francis Bopp) அவர்களின் இந்தோ-ஐரோப்பிய ஒப்பியல் ஆய்வு போன்ற ஒப்பியல் மொழி ஆய்வுகள் எல்லாம் 17-ஆம், 18-ஆம் நூற்றாண்டுகளில் முன்வைக்கப்பட்டன. இதே காலக்கட்டங்களில் ஜேகப் கிரிம் (Jacob Grimm) உட்பட அறிஞர்கள்சிலர் குறிப்பிட்ட மொழிகளின் வரலாற்றில் பேச்சொலிகள் என்னென்ன மாற்றங்களுக்கு உட்பட்டன என்ற வரலாற்றுமொழி ஆய்வுகளை மேற்கொண்டனர்.

அமைப்பு மொழியியல் (Structural Linguistics)

மேற்கூறிய வகைகளிலான ஒன்றுக்குமேற்பட்ட மொழிகளுக்கான ஆய்வுகள் தோன்றி வளர்ச்சியடைந்ததன் பயனாக மொழி ஆய்வில் ஒரு குறிப்பிடத்தக்க வளர்ச்சி ஏற்பட்டது. மனித இயற்கைமொழிகளின் அமைப்புக்களும் வளர்ச்சிப்போக்குக்களும் மொழிகளை ஆய்வுசெய்வதற்கான பொது ஆய்வுமுறைகள் முன்வைக்கப்படுவதற்கான வாய்ப்புக்களை உருவாக்கின. சுவிஸ் மொழியியல் அறிஞரான பெர்டினாண்ட் சசூர் (Ferdinand de Sassure), அமெரிக்க மொழியியல் அறிஞரான ப்ளூம்ஃபீல்ட் (Leonard Bloomfield), ஜெல்லிக் ஹேரீஸ் (Zellig S. Harris) ஆகியோர் எந்தவொரு மனித இயற்கைமொழியையும் ஆய்வுசெய்யும் வழிமுறைகளை முன்வைத்தனர். 19, 20 - ஆம் நூற்றாண்டிகளில் ஏற்பட்ட மொழி ஆராய்ச்சி வளர்ச்சி இது. இவர்களின் மொழியியல் ஆய்வுமுறைகளே அமைப்புமொழியியல் (Structural Linguistics) என்று பொதுவாக அழைக்கப்படுகிறது. மேலும் மொழிகள்பற்றிய ஆய்வு பல நிலைகளில் வளரத்தொடங்கின. பேச்சொலியியல்பற்றிய மொழியியல் ஆய்வுகளை ஹென்றி ஸ்வீட் (Henry Sweet), பால் பஸி (Paul Passy), டேனியல் ஜோன்ஸ் (Daniel Jones), பீட்டர் லேட்போஜ்ட் (Peter Ladefoged), டேவிட் ஆபர்கிராம்பி (David Abercrombi) போன்றோர் முன்வைத்தனர். உலக பேச்சொலியியல் மையம் (International Phonetic Association - IPA) தோன்றி நிலவத் தொடங்கியது. அதுபோன்று ஜேகோப்சன் (Roman Jakobson) , டிருபெட்ஸ்காய் (Nikolai Trubetzkoy),கென்னத் பைக் (Kenneth Pike) போன்றோர்ஒலியனியல் ஆய்வுக் கோட்பாடுகளையும், நைடா (Eugene Nida) போன்றோர் உருபனியல் ஆய்வுக் கோட்பாடுகளையும்முன்வைத்தனர். ஹாக்கெட்(Charles Hockett), கிளீசன் (Henry Allan Gleason) போன்றோர் அமைப்புமொழியியல் கோட்பாடுகளை (Structural Linguistics) வரையறுத்து விளக்கினர். ஒரு குறிப்பிட்ட காலகட்டத்தில் ஒரு குறிப்பிட்ட மொழியின் அமைப்புக் கூறுகளை (Synchronic Linguistics) எவ்வாறு ஆய்வுசெய்வது என்பது பற்றித் தெளிவுபடுத்தினர். அதுபோன்று ஒரு மொழியின்வரலாற்றில் ஒவ்வொரு கட்டத்திலும் நிகழ்ந்த மாற்றங்கள், வளர்ச்சிகளை ஆராயும் வரலாறுசார் மொழியியல் பிரிவும் (Diachronic Linguistics) வளர்ந்தது.

செயல்பாட்டு மொழியியல் (Functional / Systemic Grammar)

இதே காலகட்டத்தில் மொழி ஆய்வு என்பது ஒரு மொழியை அதைப் பயன்படுத்தும் சமூகச் சூழலில் வைத்து ஆய்வுசெய்தால்தான் மொழி ஆய்வு நிறைவுபெறும் என்ற அடிப்படையில் போலந்தைச் சேர்ந்த மானிடவியல் அறிஞர் மாலினோவ்ஸ்கி (Bronislaw Malinowski), எட்வேர்டு சபீர் (Edward Sapir) போன்றோர் தங்கள் மொழியியல் கோட்பாடுகளை முன்வைத்தனர். இவர்களைத் தொடர்ந்து இங்கிலாந்தைச் சேர்ந்த ஃபிர்த் (J R Firth) , ஹாலிடே (M A K Halliday), சின்கிளையர் (John Sinclair) ஆகியோர் மொழி ஆய்வு என்பது மொழியையும் அதன் பயன்பாடு அல்லது செயல்பாட்டையும் இணைத்து ஆய்வுசெய்வதுதான் என்று வலியுறுத்தி, செயல்பாட்டு மொழியியல் அல்லது முறைசார் மொழியியல் (Functional Linguistics / Systemic Linguistics) என்ற மொழியியல் கோட்பாட்டை முன்வைத்தனர்.

ஸ்கின்னரின் "மொழிப்புறநடத்தை (Verbal Behaviour)" கோட்பாடு

இயற்கைமொழிகளின் ஆய்வுபற்றிய பொதுமொழியியல் என்ற அறிவுத்துறை இவ்வாறு வளர்ந்துவந்த சூழலில்எவ்வாறு பிறந்த குழந்தைக்கு மொழி வளர்ச்சியடைகிறது,குழந்தைஎவ்வாறுதன் மூளையில் குறிப்பிட்ட மொழிதொடர்பான அறிவைச் சேமித்துவைக்கிறது என்பதுபற்றிய ஆய்வுகள் வளரத்தொடங்கின. அதாவது மனித மூளைக்கும் மொழிக்கும் இடையில் உள்ள உறவுபற்றிய ஆய்வு முக்கியத்துவம் பெறத் தொடங்கியது. 50களில் உளவியல்துறையில் மிகப் பெரிய அறிஞரான ஸ்கின்னர் (B. F. Skinner) என்பவர் முன்வைத்த "மொழிப் புற நடத்தை (Verbal Behaviour)" என்ற கோட்பாடு மிகுந்த முக்கியத்துவம் பெற்றிருந்தது. குழந்தை பிறக்கும்போது அதன் மூளையில் இயற்கைமொழி பற்றிய எந்தவொரு அறிவும் கிடையாது; பிறந்த பிறகுதான் அது தனது சுற்றுப்புறச் சூழலில் உள்ளவர்களின் மொழிச் செயல்பாட்டின்மூலம்தான் - ஏனைய நடத்தைகளைக் கற்றுக்கொள்வது போன்று - மொழியையும் கற்றுக்கொள்கிறது என்பதுதான் மொழிகற்றல்பற்றிய இவருடைய கோட்பாடாகும். மொழி கற்பித்தல் துறையிலும் இந்தக் கோட்பாடே அக்காலகட்டத்தில் செல்வாக்குச் செலுத்தியது.

நோம் சாம்ஸ்கியின்"பொதுமை அல்லது ஞால இலக்கணம் (Universal Grammar)"

மொழிக்கும் மனித மூளைக்கும் உள்ள இந்தக் கோட்பாட்டை அமெரிக்க மொழியியல் அறிஞரான நோம் சாம்ஸ்கி (Noam Chomsky) 50இன் பிற்பகுதியில் மறுத்து, மாறுபட்ட ஆய்வுகளை முன்வைத்தார். பிறந்த குழந்தை

தான் பிறந்து மிகக் குறுகிய கால இடைவெளியில் - மூன்று அல்லது நான்கு வயதுக்குள் - தனது சூழலில் நிலவுகிற மொழியை - தாய்மொழியை - "பெற்றுக்கொள்கிறது"; இயற்கையாகக் குழந்தைக்குத் தனது தாய்மொழிப்பேறு (Language Acquisition) கிடைக்கிறது. மேலும் குழந்தைக்கு இந்தக் குறுகிய கால இடைவெளியில் குறிப்பிட்ட மொழிபற்றிய அறிவுக்கான மொழிச் சான்றுகள் அனைத்தும்கிடைப்பதற்கு வாய்ப்பு இல்லை; அதாவது "மொழித் தூண்டலின் வறுமை (Poverty of Stimulus) "என்ற நிலையே குழந்தைக்கு நீடிக்கிறது. இதன் அடிப்படையில் சில முக்கியமான ஆய்வு முடிவுகளைச் சாம்ஸ்கி முன்வைத்தார். மனிதக் குழந்தை பிறக்கும்போதே அதன் மூளைக்குள் - மனித மூளைக்குள் - மனித இயற்கைமொழிகளுக்கான சில பொதுவான - அடிப்படையான - ஞால அல்லது பொது இலக்கணம் (Universal Grammar - UG)இடம்பெற்றுள்ளது. குழந்தையானது இந்தப் பொதுமை இலக்கணத்தைப் பயன்படுத்தி, தனது சூழலில் கிடைக்கிற மொழிச் சான்றுகளையும் பயன்படுத்தித் தனது மொழிக்கான - தாய்மொழிக்கான - இலக்கணத்தைப் "பெற்றுக்கொள்கிறது அல்லது ஈட்டுகிறது (acquires)" ; மாறாக, "கற்றுக்கொள்ளவில்லை (not learned)" என்று சாம்ஸ்கி கூறினார். மொழியைக் கற்கும் ஆற்றலோடு அல்லது அதற்கான மொழி ஈட்டப் பொறிநுட்பத்தோடுதான்(LanguageAcquisition Device - LAD)ஒவ்வொரு குழந்தையும் பிறக்கிறது; இவ்வாற்றல் மொழிப் பொதுமைகள் நிறைந்தது. குழந்தைக்கு இருக்கிற இந்தப் பொதுமை இலக்கணம் மனித இனத்திற்கே உள்ள ஒரு உயிரியல் திறனாகும்; உள்ளுறை இலக்கணமாகும் (Genetically given - a biological endowment - innate one) என்று அவர் கூறினார். மனிதர் அல்லாத ஏனைய விலங்கினங்களுக்கு இந்தத் திறன் கிடையாது என்று கூறினார். எனவேதான் சிம்பன்சி போன்ற மனிதக் குரங்குகளுக்குக்கூட இயற்கைமொழிகளைப் பெறமுடிவது இல்லை.

மனிதமூளைக்கும் இயற்கைமொழிக்கும் இடையில் உள்ள உறவுபற்றிய சாம்ஸ்கியின் கோட்பாடானது ஸ்கின்னரின் "மொழி நடத்தை" கோட்பாட்டைமொழி ஆய்வுலகில் பின்னுக்குத் தள்ளியது. மேலும் மொழி ஆய்வு என்பது ஒருவரின் வெறும் புறநடத்தையாகக்கொண்டு ஆய்வுசெய்த வண்ணனை அல்லது விளக்க மொழியியலின் கருத்தையும் சாம்ஸ்கி மறுத்தார். இந்தப் மொழிப் புறநடத்தைக்குப் பின்புலமாக மூளையில்நீடிக்கிற மொழியறிவு (Linguistic knowledge) மொழியியல் ஆய்வில் முக்கியம் என்பதை வலியுறுத்தினார். ஒருவரின் மொழிச்செயல்திறத்திற்குப் (linguistic performance) பின்னால் நீடிக்கிற மொழி அறிதிறம் பற்றிய (linguistic competence)ஆய்வுதான் மொழியியலின் முக்கியப் பணியாக அமைதல்வேண்டுமென அவர் வலியுறுத்தினார். குழந்தை பிறக்கும்போது மூளையில் நீடிக்கிற 'பொதுமை இலக்கணத்தைக்"

கண்டறிவதற்கான சாம்ஸ்கியின் முயற்சி கடந்த 75 ஆண்டுகளாகப் பல்வேறு மாற்றங்களையும் வளர்ச்சியையும் பெற்றுவருகிறது. இந்த ஆய்வுமுறையிலான இலக்கணமானது "உருவாக்க மாற்றிலக்கணம் (Transformation Generative Grammar)" என்று அழைக்கப்படுகிறது.

அதுவரை நீடித்துவந்த இலக்கணம் பற்றிய கொள்கைகளோடுஅவர் வேறுபட்டார். ஒரு மொழியின் இலக்கணம் என்பது அம்மொழியின் தரவுகள் அனைத்தையும் சேகரிப்பது (Observational Adequacy) அல்லது சேகரித்தவற்றை விவரிப்பது (Descriptive Adequacy) மட்டும் இல்லை; மாறாக, அவற்றையெல்லாம் விளக்கவேண்டும் (Explanatory Adequacy) என்று அவர் வலியுறுத்தினார். இயற்கைமொழிகளின் அமைப்பைக் கண்டறியும் மொழியியல் துறையானது இயற்கைமொழிகளை மனிதமூளை எவ்வாறு "பெற்றுக்கொள்கிறது" ; பின்னர் எவ்வாறு ஒருவர் தமது சூழலில் பேசப்படுகிற மொழித்தொடர்களின் பொருண்மையைத் தெரிந்துகொள்கிறார்?; பொருண்மை மயக்கங்களுக்கு (Meaning Ambiguity) அடிப்படையான கூறுகள் என்ன? ; அவற்றை எவ்வாறு மனிதமூளை தீர்த்துக்கொள்கிறது போன்றவற்றைப்பற்றிய இயலாகச் சாம்ஸ்கியின் மாற்றிலக்கணம் இன்று வளர்ந்து நிற்கிறது. இந்த அடிப்படையில் பலவகைப்பட்ட "மாற்றிலக்கணங்கள்" இன்று தோன்றி நிலவுகின்றன.

ஒரு மொழியின் தொடரமைப்புக்களை அவற்றில் இடம்பெற்றுள்ள சொற்கள், சொற்றொடர்கள், தொடர் ஆகியவற்றின் அமைப்புக்களை அடிப்படையாகக்கொண்டு மொழி ஆய்வுசெய்கிற மாற்றிலக்கண வடிவங்கள் (Category-based analysis) ஒருபுறமும் தொடரமைப்புக்களில் இடம்பெறுகிற சொற்களுக்கு இடையில் உள்ள உறவுகளின் அடிப்படையில் மொழி ஆய்வுசெய்கிற இலக்கண வடிவங்கள் (Relational Grammatical analysis) மறுபுறமும் இன்று நீடிக்கின்றன.

மொழிவகை ஆய்வு(Language Typology)

இதே காலகட்டத்தில் உலகமொழிகளின்புறவய அமைப்புக்களில் காணப்படுகிற பொதுமைக் கூறுகளைப்பற்றிய ஆய்வும் இன்று மிகவும் வளர்ந்துள்ளது. கிரீன்பெர்க் (Joseph Greenberg), கோம்ரி (Bernard Comrie), கிராப்ட்(William Croft) போன்றோர் "மொழி வகை ஆய்வு (Language Typology)" என்ற மொழிக்கோட்பாடுகளின் அடிப்படையில் மனித இயற்கைமொழிகளின் பொதுமைபற்றிய ஆய்வுகளை மேற்கொண்டனர்.

மொழியியலின் பிற பிரிவுகள்

மேற்கூறிய காலகட்டங்களில் உளவியல்மொழியும் வளர்ச்சிபெற்றுவந்தது. குழந்தையின் மூளை வளர்ச்சி, மன வளர்ச்சி ஆகியவற்றிற்கும் மொழி வளர்ச்சிக்கும்இடையில் உள்ள உறவுகள் பற்றிய உளவியல்மொழியாய்வு (Psycholinguistics) வளர்ந்துள்ளது. மேலும் மனிதமூளையின் அமைப்பு, செயல்பாடு ஆகியவற்றிற்கும்மொழி அமைப்புக்கும் இடையில் உள்ள உறவுகள்பற்றிய நரம்புமொழியியல் துறையும் (Neurolinguistics)வளர்ச்சி அடைந்துள்ளது. மனித மூளையில் உள்ள சில குறிப்பிட்ட பகுதிகளுக்கும் மொழிகளின் குறிப்பிட்ட அமைப்பு நிலைகளுக்கும் இடையில் உள்ள உறவுகள்பற்றிப் புரோகா (Paul Broca), வெர்னிக் (Carl Wernicke) ஆகியோர் பல ஆய்வுகளை முன்வைத்துள்ளனர்.

மேலும் மொழிக்கும் சமுதாயத்திற்கும் இடையில் உள்ள உறவுகள்பற்றிய ஆய்வும் (Sociolinguistics) இன்று வளர்ந்துநிற்கின்றது. இவ்வாறு இன்றைய நிலையில் மொழியியல் துறை பல நிலைகளில் மனித இயற்கைமொழிகள்பற்றிய ஒரு மிகச் சிறந்த அறிவியல்துறையாக வளர்ந்து நிற்கிறது. ஒரு குறிப்பிட்ட மொழியின் அமைப்பைப்பற்றிய ஆய்வாக - குறிப்பிட்ட மொழியின் இலக்கண நூல் (Grammar of a particular language) உருவாக்கமாக - நிலவிய மொழி ஆய்வு இன்று மனித இயற்கைமொழிகள் அனைத்துக்குமான பொதுவான ஒரு அறிவியல்துறையாக (Linguistics) வளர்ச்சியடைந்துள்ளது.

இங்குக் குறிப்பாக முன்வைக்கப்படவேண்டிய ஒரு முக்கியக் கருத்து மொழியியலானது - இயற்கைமொழிகளின் அமைப்பைப் புறவயமாக ஆய்வுசெய்யக்கூடிய, கணித அடிப்படையிலான ஒரு முறைசார் வடிவமாக (Formal grammar - Formalism) முன்வைக்கக்கூடிய ஒரு அறிவியல்துறையாக வளர்ந்துநிற்கிறது.

மேற்கூறிய மொழியியலின் இன்றைய வளர்ச்சிக்கும் இன்று நான் உரையாற்ற உள்ள கணினி மொழியியலின் வளர்ச்சிக்கும் நெருங்கிய உறவு உள்ளது.

மொழியியலுக்கும் கணினிமொழியியலுக்கும் இடையில் உள்ள உறவுகள்

இயற்கைமொழிகளை மனித மூளை எவ்வாறு "பெற்றுக் கொள்கிறது";பெற்றுக்கொண்ட மொழி அறிவை எவ்வாறு, எங்கே சேமித்துவைக்கிறது; அந்த அறிவை நாம் எவ்வாறு பயன்படுத்தி மொழிச் செயல்பாடுகளை மேற்கொள்கிறோம்; கருத்தாடலில் தோன்றுகிற பொருண்மை

மயக்கங்களை எவ்வாறு தீர்த்துக்கொள்கிறோம் போன்ற வினாக்களை முன்வைத்து அவற்றிற்கான விடைகளைத் தர முயல்கிற துறையாக மொழியியல் துறை நீடிக்கிறது என்று முன்பு பார்த்தோம்.

20-ஆம் நூற்றாண்டில் தோன்றி, 21-ஆம் நூற்றாண்டில் உலகந் தழுவிய ஒரு முக்கியத்துவத்தைப் பெற்றுள்ள கணினிமொழியியல் துறையானது மின்னணுக் கருவியான கணினியின் மூளைக்கு - மின்னணுச் சில்லுக்கு - இயற்கைமொழிகளைக் கற்றுக்கொடுக்கமுடியுமா; கற்றுக்கொடுக்கமுடியுமென்றால் எவ்வாறு கற்றுக்கொடுப்பது; கணினி பெற்றுக்கொண்ட "மொழி அறிவை" எவ்வாறு நாம் பயன்படுத்துவது போன்ற வினாக்களை முன்வைத்து அவற்றிற்கான விடைகளைத் தர முயல்கிற துறையாகக் கணினிமொழியியல் துறை நீடிக்கிறது.

மனித மூளையும் இயற்கைமொழியும் (Human Brain and Natural Language) - மின்னணுக் கணினியும் இயற்கைமொழியும் (electronic Computer and Natural Language) என்ற அடிப்படையில் மொழியியல் - கணினிமொழியியல் இரண்டுக்கும் இடையில் உள்ள உறவுகளையும் வேறுபாடுகளையும்பற்றி ஆராய்வதே இந்த உரையின் மைய நோக்கமாகும்.

கணினிமொழியியலின் தோற்றம்

20-ஆம் நூற்றாண்டின் தொடக்கத்தில் வல்லரசு நாடுகளுக்கிடையே ஏற்பட்ட முதலாவது உலகப் போரில் ஒரு வல்லரசு தனதுஇராணுவத்திற்குள் பரிமாறிக்கொள்ளும் சங்கேதமொழியில் அமைந்த இரகசியங்களை மற்றொரு வல்லரசானது பெற்று, அதனைப் புரிந்துகொள்ளும் முயற்சிகளை மேற்கொள்வது வழக்கமாக இருந்தது. அதாவது ஒரு மொழியில் - சங்கேத மொழியில் - அமைந்த செய்திகளை மற்றொரு மொழிக்கு மொழிபெயர்க்கும் முயற்சி மேற்கொள்ளப்பட்டது. இதனுடைய தொடர்ச்சியாக ஒரு இயற்கைமொழியில் அமைந்துள்ள செய்திகளை மற்றொரு மொழிக்கு மொழிபெயர்க்கும் முயற்சிகள் வளரத் தொடங்கின.

முதன்முதலாக வாரன் வீவர் (Warren Weaver) என்ற அறிஞர் 1949இல் இயந்திர மொழிபெயர்ப்பு சாத்தியமான ஒன்றுதான் என்பதைக் கூறும் ஒரு அறிக்கையை வெளியிட்டார். 1952-இல் அமெரிக்காவில் உள்ள எம் ஐ டி ஆய்வு நிறுவனத்தில் இயந்திர மொழிபெயர்ப்பு மாநாடு நடைபெற்றது. அதன் தொடர்ச்சியாக "Mechanical Translation" என்ற பெயரில் ஒரு ஆய்விதழ் தொடங்கப்பட்டது. பின்னர் அந்த இதழின் பெயர்" Mechanical Translation and Computational Linguistics" என்று மாற்றப்பட்டது. இந்த ஆய்விதழைத்

தொடர்ந்துகொண்டுவரும் முயற்சியாக 1962-ஆம் ஆண்டில் உருவாக்கப்பட்ட "Association for Machine Translation and Computational Linguistics"என்ற ஒரு அமைப்பு இந்த இதழை நடத்தும் பொறுப்பை ஏற்றுக்கொண்டது.

இயந்திர மொழிபெயர்ப்புக்கான மேற்கூறிய பணிகள் தொடர்ந்த வேளையில் எந்தஅளவுக்கு இந்த முயற்சிகள் வெற்றிபெற்றுள்ளன என்பதை ஆராய்ந்து கூறுவதற்கு அமெரிக்க அரசாங்கம் தேசிய அறிவியல் கழகத்தின்கீழ் (National Academy of Sciences) தானியங்கு மொழிபெயர்ப்புக்கானஅறிவுரைக்குழு (Automatic Language Processing Advisory Committee - ALPAC) ஒன்றை அமைத்தது.வாரன் வீவர் தலைமையிலான இந்தக் குழு அதுவரை இயந்திர மொழிபெயர்ப்பு முயற்சிகள் எந்த அளவு வெற்றிபெற்றுள்ளது என்பதைப்பற்றி ஒரு விரிவான ஆய்வை மேற்கொண்டது.

இயந்திர மொழிபெயர்ப்புக்காகச் செலவழிக்கப்பட்ட நிதிக்கு ஏற்ப இந்தப் பணி முன்னேற்றம் பெற்றுள்ளதா, அவ்வாறு பெறவில்லையென்றால் அதற்கான காரணங்கள் என்ன, இப்பணி முன்னேறுவதற்குச் செய்யவேண்டிய பணிகள் என்னென்ன என்பதுபற்றியெல்லாம் இந்தக் குழு ஆய்வு மேற்கொண்டது. இறுதியில் அந்தக் குழுவானது தனது அறிக்கையில் இயந்திர மொழிபெயர்ப்புப் பணிகள்எதிர்பார்த்த அளவுக்குத் தக்க வளர்ச்சியடையவில்லை என்ற முடிவுக்கு வந்தது. மொழிகள்பற்றிய ஆழமான ஆய்வுகளை அடிப்படையாகக்கொண்டு இயந்திர மொழிபெயர்ப்புப் பணிகள் மேற்கொள்ளப்பட்டால்தான் அது வெற்றிபெற இயலும் என்ற கருத்தை வலியுறுத்தித் தன் அறிக்கையை (ALPAC Report) வெளியிட்டது. இதன் விளைவாக இயந்திர மொழிபெயர்ப்புக்கான நிதி உதவிகள் குறைந்தன. அதற்கான பணிகளும் தடைபட்டன.

மேற்கூறிய அறிவுரைக்குழுவில் பங்கு வகித்த டேவிட் கேய்ஸ் (David Hays) என்ற அறிஞர் முதன்முதலாகக் "கணினிமொழியியல்" என்ற சொற்றொடரை உருவாக்கியதாகக் கூறப்படுகிறது.

தற்காலிகமாகப் பாதிக்கப்பட்ட இயந்திர மொழிபெயர்ப்புப் பணிகள் 1974-வாக்கில் மீண்டும் முன்னேறத் தொடங்கின. 1974-ஆம்ஆண்டுஅதுவரை வெளிவந்த "Machine Translation and Computational Linguistics" என்ற இதழின் பெயரானது "American Journal of Computational Linguistics"என்று பெயர் மாற்றம் செய்யப்பட்டது. பின்னர் 1980-இல் "Computational Linguistics" என்று பெயர் மாற்றம் பெற்றது. தற்காலிகமாகப் பாதிக்கப்பட்டிருந்த இயந்திர மொழிபெயர்ப்புக்கான நிதி உதவி பெருகத் தொடங்கியது. இயற்கைமொழிகளைப் புரிந்துகொள்ளும்

திறனைக் கணினிக்கு அளிப்பதற்கான ஆய்வுப் பணிகள் பெருகத் தொடங்கின. கணினிமொழியியல் துறை வளரத் தொடங்கி, கடந்த 45 ஆண்டுகளில் குறிப்பிடத்தக்க வெற்றியைப் பெற்றுள்ளது.

கணினிமொழியியலின் அடிப்படை நோக்கம்

நாம் முன்பே பார்த்ததுபோல, மொழியியல் துறையின் பன்முக வளர்ச்சியானது கணினிமொழியியல் வளர்ச்சிக்குச் சிறந்த பங்கை அளித்துள்ளது. மனித மூளையில் வளர்ந்துநிற்கிற இயற்கைமொழி அறிவின் அமைப்பு எவ்வாறானது என்பதை விளக்கும் மொழியியல் துறையின் ஆய்வுகள் கணினி மூளைக்கு - கணினிச் சில்லுக்கு - அந்த இயற்கைமொழி அறிவை எவ்வாறு கொடுக்கலாம் என்பதுபற்றிய ஆய்வாகவும் வளரத்தொடங்கியது.

இங்கு கணினியிடம் நாம் இரண்டு மொழிச் செய்யறிவுத்திறன்களை எதிர்பார்க்கிறோம். ஒன்று, நமது இயற்கைமொழிகளின் தொடர்களை அது புரிந்துகொள்ளும் திறன் (Natural Language Understanding - NLU)மற்றொன்று இயற்கைமொழிகளில் தொடர்களை உருவாக்கித் தரும் திறன் (Natural Language Generation - NLG).

அதனடிப்படையில் நமக்கும் கணினிக்கும் இடையே நேரடிக் கருத்தாடல் (Man - Machine Interface) நடைபெறவேண்டும்.

கோடியேகோடி நரம்பணுக்களைக்கொண்ட மனித மூளையில் நிலவுகிற ஒரு குறிப்பிட்ட மொழியின் அமைப்பு - சொற்களஞ்சியம், இலக்கணம் இரண்டும் - எவ்வாறானது என்பதை விளக்குவதில் மொழியியலில் வேறுபட்ட கோட்பாடுகள் தோன்றி நீடிக்கின்றன. இயற்கைமொழியின் இந்த அமைப்பைக் கணினிக்கு - இலட்சக்கணக்கான மின்னணுச் சில்லுகளைக் கொண்ட கணினி மூளைக்கு - அளிப்பதற்கு எந்தவொரு மொழியியல் கோட்பாடு பயன்படும் என்ற ஆய்வு வளரத் தொடங்கியது.

கணினியானது எண்மங்களை அடிப்படையாகக்கொண்ட ஒரு மின்னணுக் கருவி. எனவே கணித அடிப்படையில் அமைகிற எது ஒன்றையும் கணினிக்குக் கொடுக்கமுடியும். அப்படியென்றால் மனித மூளையில் நீடிக்கிற ஒரு குறிப்பிட்ட மொழியின் அமைப்புக் கூறுகளை அல்லது பண்புகளைக் கணித அடிப்படையில் வெளிப்படுத்தமுடியுமா என்பதே இங்கு தோன்றிய முதல் வினா.

எனவே, கணினிமொழியியல் துறையைச் சேர்ந்தவர்கள்- மொழியியலார் மட்டுமல்லாமல் கணினித்துறையைச் சேர்ந்த அறிஞர்களையும் உள்ளடக்கிய ஆய்வாளர்கள் -மேற்கூறிய திசையில் தங்கள் ஆய்வைத் தொடர்ந்துவருகின்றனர். சாம்ஸ்கி உட்பட மொழியியல் அறிஞர்கள் பலர் இயற்கைமொழிகளின் சொற்களஞ்சியங்கள், இலக்கணங்கள் ஆகியவற்றை முறைசார் வடிவத்தில் - முறைசார் அகராதி, முறைசார் இலக்கண வடிவத்தில் வெளிப்படுத்தப் பலவகை முறைசார் இலக்கணங்களை முன்வைத்து வருகின்றனர். Tagmemics Grammar, Stratificational Grammar, Glossmatics, Functional Grammar, Systemic Grammar, Generative Grammar, Relational Grammar, Cognitive Linguistics, The Prague School Linguistics, Case Grammar, Optimality Theory என்றுபலவகைப்பட்ட முறைசார் இலக்கணங்கள் இன்று மொழியியல் துறையில் முன்வைக்கப்பட்டுள்ளன. மாற்றிலக்கண இலக்கணத்தில்கூடசாம்ஸ்கியின் மாற்றிலக்கணத்தோடு Lexical Functional Grammar (LFG), Generative Semantics, Generalizrd Phrase Structure Grammar (GPSC), Head-driven Phrase Structure Grammar (HDFSG), Tree Adjoing Grammar (TAG) என்று பலவகைப்பட்ட மொழிசார் வடிவங்கள் மொழியியல் துறையில் நீடிக்கின்றன. இவற்றையெல்லாம் இங்குக் குறிப்பிடுவதற்குக் காரணம், மொழியியலில் மொழி அமைப்பை விளக்குவதற்கான கோட்பாடுகள், முறைசார் வடிவங்கள் பல நீடிக்கின்றன என்பதை வலியுறுத்துவதற்கே ஆகும். ஆனால் இவை அனைத்தின் நோக்கமும், மனித இயற்கைமொழிகளின் சொற்களஞ்சியங்களையும் இலக்கணங்களையும் விளக்குவதற்கான முறைசார் வடிவங்களை முன்வைப்பதே ஆகும். இவற்றிற்கிடையே நீடிக்கும் வேறுபாடுகளுக்குக் காரணம், மொழிபற்றிய கோட்பாடுகள், மொழி ஆய்வின் நோக்கம், ஆய்வுமுறைகள் ஆகியவற்றில் நீடிக்கிற வேறுபாடுகளே ஆகும். ஆனால் இவை அனைத்துமே மனித மூளையானது ஒரு குறிப்பிட்ட மொழியின் அமைப்பை எவ்வாறு தனக்குள் கட்டமைத்து வைத்துள்ளது என்பதை விளக்குவதே ஆகும்.

கணினிமொழியியலின் முன் நிற்கும் வினாக்கள்

மேற்கூறப்பட்ட மொழியியலின் முறைசார் இலக்கண வடிவங்களின் அடிப்படையில் கணினியின் மூளைக்கும்இயற்கைமொழிகளின்அமைப்பை க் கொடுக்கமுடியுமா என்பதே இப்போது கணினிமொழியில் முன்பு நீடிக்கிற வினா. எடுத்துக்காட்டாக, ஒரு மொழியின் உருபனியல் (Morphology) அறிவைக் கணினிக்குக் கொடுப்பதற்கு அம்மொழியின் அகராதி (Lexicon) , இலக்கணச் சொற்கள் அல்லது விகுதிகள் (Grammatical Words or Affixes), ஒரு சொற்றொடரில் அகராதிச்சொல்லும் அதனுடன் இணையும் இலக்கண விகுதிகளும் வரும் முறை (Morphotactics), அவ்வாறு அகராதிசொல்லும் விகுதிகளும்

இணைந்துவரும்போது ஏற்படும் உருபொலியன் மாற்றங்கள் (Morphophonemics) ஆகியவற்றைப் பற்றிய அறிவை அளித்தால் போதுமா என்ற வினா. அவ்வாறு அளிக்கமுடியுமென்றால் எந்த வடிவங்களில் அவற்றை அளிப்பது என்ற வினா. இதுபோன்ற வினாக்களே ஒலியனியல் (Phonology), தொடரியல் (Syntax), பொருண்மையியல் (Semantics), சூழல்சார் பொருண்மையியல் (Pragmatics) ஆகியவற்றிற்கும் பொருந்தும்.

அதாவது,மனிதமூளைக்குஒருபுதியமொழியைக்கற்றுக்கொடுக்கும்போது அளிக்கப்படுகிற இலக்கண அறிவும் அதன் வடிவங்களும் அப்படியே கணினிக்கு அளித்தால் கணினியால் அந்த மொழியைக் கற்றுக்கொள்ளமுடியுமா அல்லது இந்த அறிவுகளை வேறு வடிவங்களில் - வேறுபட்ட முறைசார் வடிவங்களில் பொதிந்து - அளிக்கவேண்டுமா என்பதே அடிப்படை வினா.

மேற்கூறிய வகையில்அவ்வாறே - வெறும் மொழிசார் வடிவங்களாகவே (Linguistic Formalism) - அளிக்கமுடியாது என்பதைக் கணினிமொழியியலின் பதிலாகும். எந்தவொரு ஒரு பணியையும் கணினி மேற்கொள்வதற்கான நிரல்கள் உருவாக்கத்திற்கென்று முறைசார் வடிவங்கள் (Computational Formalism) உண்டு; வழிமுறைகள் (Algorithm)தர்க்க வடிவம் (Logic) ஆகியவை உண்டு. முறைசார் மொழியியல் வடிவத்தையும் கணினிக்கான முறைசார் வடிவத்தையும் தர்க்க வடிவத்தையும் உள்ளடக்கிய கணினிமொழியியல் முறைசார் வடிவங்கள் (Computational Linguistic Formalism) தேவை.

எனவே, முறைசார் மொழியியல் வடிவத்தையும் முறைசார் கணினியியல் வடிவத்தையும் முறையாக இணைத்த முறைசார் வடிவத்தின் வழியேதான் கணினியின் மூளைக்கு இயற்கைமொழி அறிவைக் கற்றுக்கொடுக்கமுடியும். அதாவது, மனிதமூளையைச் சார்ந்த மொழியியல் முறைசார் வடிவங்களை அப்படியே கணினி மூளைக்கு அளிக்கமுடியாது.

மேற்கூறியவற்றின் அடிப்படையில் கணினிக்கு மனிதனின் இயற்கைமொழிகளைக் கற்றுக்கொடுப்பதற்கான மொழிசார் வடிவங்களைப்பற்றிய ஆய்வில்இன்று கணினிமொழியியல் வளர்ந்து நிற்கிறது.

இயற்கைமொழி ஆய்வு (Natural Language Processing - NLP)

கணினிமொழியியல் துறையின் அறிவைக் கணினிக்கு அளித்து,மனித சமுதாயத்தின் மொழிவழிச் செயல்பாடுகளைக் கணினிவழியே மேற்கொள்வதற்கான பொறியியல் துறை (Engineering Branch) இன்று

இயற்கைமொழி ஆய்வு என்றும் வழங்கப்படுகிறது. இரண்டும் ஒன்றுதான் என்ற ஒரு கருத்தும் உண்டு. இதை மொழித்தொழில்நுட்பம் (Language Technology) என்றும் அழைக்கலாம்.

இயற்கைமொழி ஆய்வு அல்லது மொழித்தொழில்நுட்பத்தின் பணிகளை மூன்று வகைகளில் பிரித்துப் பார்க்கலாம். ஒன்று, நாம் பேசுவதைக் கணினியே எழுத்துவடிவத்தில் எழுதித்தருவது (Speech to Text) அல்லது நாம் எழுதுவதை அப்படியே பேச்சாக மாற்றித்தருவது (Text to Speech) ஆகும். இதுபற்றிப் பின்னர் விளக்கமாகப் பார்க்கலாம்.

இரண்டாவது, நமது உரைகளை - எழுத்துவழியோ அல்லது பேச்சுவழியோ நாம் முன்வைக்கும்பனுவலை - பல நிலைகளில் ஆய்வுசெய்து தருகிற பனுவல் ஆய்வு (Text Processing) . அதனடிப்படையில் நமது உரைகளின் காணப்படும் சொற்பிழைகள், ஒற்றுப்பிழைகள், தொடர்ப்பிழைகள், பொருண்மைப் பிழைகள் ஆகியவற்றைக் கண்டறிந்து திருத்தித் தரும் மென்பொருள்கள் (Proofing tools) உருவாக்கப்படுகின்றன. மேலும் நமது உரையைச் சுருக்கித் தரும் மென்பொருள் (Text Summarization), ஒரு மொழியில் உள்ள பனுவலை மற்றொரு மொழியில் மொழிபெயர்த்துத் தரும் மென்பொருள் (Automatic Machine Translation)போன்றவற்றை உருவாக்கும் பணிகளை இப்பிரிவு மேற்கொள்கிறது. இதுபற்றிப் பின்னர் விளக்கமாகப் பார்க்கலாம்.

மூன்றாவது, ஒளிவருடிமூலம் (Scanner)நகல் எடுக்கப்பட்ட எழுத்துப் பனுவலை மீள் பதிப்புக்காக எழுத்துருக்களாக மாற்றி அமைக்கிற ஒளிவழி எழுத்தறிவான் மென்பொருள் (Optical Character Recognization - OCR).

மேற்கூறிய மூன்று வகைகளுக்கும் பயன்படும் மென்பொருள்களை உருவாக்கித் தரும்துறையே இயற்கைமொழி ஆய்வு அல்லது மொழித்தொழில் நுட்பம் ஆகும்.

பேச்சுத் தொழில்நுட்பம் (Speech Technology)

பேச்சொலியியல் ஆய்வு (Phonetics) மொழியியலில் ஒரு முக்கியமான துறையாகும். இத்துறையில் நாம் பேசுகிற பேச்சில் இடம்பெறும்ஒலிகள் (phones), ஒலியன்கள் (Phonemes) , மாற்றொலிகள் (Allophones), இசைக்குணம் (Tone), அசை அல்லது சொல் அழுத்தம் அல்லது உரத்தல் (Syllabic or Word Stress), ஒலியேற்ற இறக்கம் (Intonation)மொழியியல் நோக்கில் ஆகியவை ஆய்வுசெய்யப்படுகின்றன.

மொழியியலின் இந்தப் பிரிவில் மூன்று வகைகளில் பேச்சொலிகளை ஆய்வுசெய்யலாம். ஒன்று, ஒரு மொழியில் ஒரு குறிப்பிட்ட பேச்சொலியை ஒருவர் ஒலிக்கும்போது அவருடைய உடல் உறுப்புக்களான வயிற்றுத்தசைகள், மூச்சு உறுப்புக்கள், தொண்டை, வாய், வாயில் உள்ள நாக்கு, பல், அண்ணம் போன்றவை எவ்வாறு செயல்படுகின்றன என்பதை ஆய்வுசெய்யலாம். இப்பிரிவு ஒலிப்பியல் (Articulatory Phonetics) என்று அழைக்கப்படுகிறது. ஒவ்வொரு பேச்சொலியும் தனக்கென்று ஒலிக்கும் கூறுகளைப் பெற்றுள்ளது. மற்றொரு வகையானது ஒருவர் ஒலித்த பேச்சொலியானது கேட்பவரின் காதுகளில் எந்தவகையான பாதிப்பை அல்லது விளைவை உருவாக்குகிறது என்பதுபற்றிய ஆய்வாகும். இதை ஒலியுணர்வியல் (Auditory Phonetics)என்று அழைக்கிறார்கள். மூன்றாவது வகையானது ஒலிக்கப்படுகிற ஒரு பேச்சொலியின் இயற்பியல் பண்புகளை - அலைநீளம் (Wave length), அடர்த்தி (Intensity), நேரம் (Time)ஆகியவற்றை - ஆய்கின்ற ஆய்வாகும். இது ஒலியியக்கவியல் அல்லது இயற்பியல்சார் பேச்சொலியியல் (Acoustic Phonetics) என்று அழைக்கப்படுகிறது.

மேற்கூறிய மூன்று பேச்சொலியியல் பிரிவுகளில் இயற்கைமொழி ஆய்வு அல்லது மொழித்தொழில்நுட்பத்திற்குப் பயன்படும் ஆய்வானது ஒலியியக்கவியல் அல்லது இயற்பியல்சார் பேச்சொலியியல் ஆகும்.

ஆய்வுநோக்கில் ஒருவருடைய பேச்சில் உள்ள பேச்சொலிகளைத் தனித்தனியே பிரித்து ஆய்வுசெய்யலாம். ஆனால் அவர் ஒரு சொல்லில் உள்ள பேச்சொலிகளைத் தனித்தனியே (Discrete) ஒலிப்பதில்லை. மாறாக, ஒரு தொடர் அலையாகத்தான் (Wave - Anlog)சொல்லை ஒலிக்கிறார். எனவே ஒரு சொல்லில் இடம்பெறும் ஒரு பேச்சொலியை உச்சரித்துமுடிக்கும்முன்னரே அடுத்த பேச்சொலிக்கான ஒலிப்புக்கு வாயுறுப்பு தயாராகிவிடும். இடைவெளி இருக்காது. குறிப்பிட்ட பேச்சொலியின்இறுதிப்பகுதியின்மீது அதற்கு அடுத்த பேச்சொலியின் முதற்பகுதி தனது செல்வாக்கைச் செலுத்தும். ஒரு குறிப்பிட்ட தொடரில் ஒரு சொல்லுக்கும் அடுத்த சொல்லுக்கும் இடையில் உள்ள நேர இடைவெளியைவிட ஒரு சொல்லுக்குள் உள்ள அடுத்தடுத்த பேச்சொலிகளுக்கிடையேயான நேர இடைவெளி குறைவாக இருக்கும்.

மேலும் நாம் பேச்சை ஒலிக்கும்போது மொழியசைகளாகத்தான் (Linguistic Syllables) ஒலிக்கிறோம். மெய்யொலிகள் தடையொலிகளாக (Obstructed sounds) இருப்பதால் அவற்றைத் தனித்து ஒலிக்கமுடியாது. அதனால் ஒரு உயிர் ஒலிக்கு (Vowel) முன்னாலோஅல்லது பின்னாலோதான் உயிரோடு இணைத்து அந்த மெய்யை ஒலிக்கமுடியும். எனவே, மொழியசைகளாகக்

கிடைக்கிற பேச்சொலிக்கூட்டைத்தான் நாம் ஆய்வில் பெறமுடியும். இந்தப் பேச்சொலிக்கூட்டானது ஒலியலைகளாக நமக்குக் கிடைக்கிறது. அவற்றை அடிப்படையாகக்கொண்டுதான் நாம் பேச்சொலியியல்ஆய்வை மேற்கொள்ளமுடியும். இதற்கான தொழில்நுட்பங்கள் இன்று வளர்ந்துள்ளன.

ஒரு உயிரையும் ஒரு மெய்யையும் கொண்ட ஒரு பேச்சலையை (Speech Wave) இன்று நாம் உயிர், மெய் என்று பிரித்து ஆய்வுசெய்யமுடியும்.

பேச்சு-எழுத்து மாற்றி (Speech to Text - ASR)

எனவே, ஒருவரது பேச்சில் இடம்பெறும் சொற்களின் பேச்சலைகளை ஆய்வுசெய்து அதனடிப்படையில் மெய், உயிர் பேச்சொலிகளைக் கண்டறியலாம். மேலும் மொழியசைகளையும் கண்டறியலாம். இவ்வாறு பிரித்து ஆய்வுசெய்யும் பேச்சொலித்தொழில்நுட்பம் (Speech Processing Technology) இன்று வளர்ந்துநிற்பதால், ஒரு குறிப்பிட்ட சொல்லுக்கான பேச்சலையைப்பெற்று, அதனை உயிர், மெய், மொழியசை என்று ஆராயமுடியும். அதனடிப்படையில் குறிப்பிட்ட பேச்சொலி, மொழியசைகளை அவற்றிற்குரிய எழுத்துக்களில் (Graphemes or Alphabets) எழுதமுடியும் அல்லது கணினியில் எழுத்துருக்களில் (Fonts) மாற்றிக் காணமுடியும். இதுவே பேச்சை எழுத்தாக மாற்றும் தொழில்நுட்பத்திற்கு (Speech to Text - ASR - Automatic Speech Recognizer) அடிப்படையாக அமைகிறது. இன்று கணினியில் நாம் பேசுவதை அல்லது ஒலிப்பதை எழுத்துருக்களில் பெறமுடியும்.

இங்கு நாம் கவனத்தில் கொள்ளவேண்டிய ஒரு முக்கியமான செய்தி உள்ளது. தமிழ்மொழியில் ஒரு தனி உயிர் அல்லது மெய் ஒலியனுக்கும் (Phoneme) மொழியசைக்கும் (Linguistic Syllable) தனித்தனி வரிவடிவம் அல்லது எழுத்து உண்டு. எனவே ஒலியன்களையும் மொழியசைகளையும் எழுத்து அல்லது எழுத்துரு வடிவில் கொடுப்பதில் சிக்கல் இல்லை. ஆனால் தமிழில் வல்லின ஒலியன் (Stop Phonemes) ஒவ்வொன்றுக்கும் ஒன்றுக்குமேற்பட்ட மாற்றொலிகள் (Allophones) உள்ளன. 'க (/k/)'என்ற ஒலியனுக்குமூன்று மாற்றொலிகள் உண்டு - [k] [g] [x] என்ற மாற்றொலிகள் உண்டு. தமிழில் ஒலியன்களுக்கும் மொழியசைகளுக்கும்தான் எழுத்துவடிவம் இருக்கிறதேதவிர, மாற்றொலிகளுக்குத் தனி எழுத்து கிடையாது. எனவே தமிழ்ச்சொல்லை ஒருவர் ஒலிக்கும்போது அந்த ஒலிக்கற்றையை ஆய்வுசெய்து, அதில் பயின்றுவருகிற மாற்றொலிகளை ஒலியன்களாக மாற்றியபிறகுதான் அதற்கு எழுத்து வடிவத்தை அளிக்கமுடியும்.

இதுபோன்று, ஆங்கிலமொழியிலும் வேறு ஒருவகை சிக்கல் உள்ளது. ஆங்கிலத்தில் சில சொற்களில் பயின்றுவருகிற எழுத்துக்களை அப்படியே ஒலிக்க அல்லது பலுக்கமுடியாது. இச்சொற்களின் ஒலிப்பு அல்லது பலுக்கலுக்கும் அவற்றில் பயின்றுவருகிற எழுத்துக்களுக்கும் நேரடித் தொடர்பு கிடையாது. Plumber என்ற சொல்லை ஒலிக்கும்போது "b" என்பது வெளிப்படாது; ஆனால் எழுதும்போது அந்த எழுத்தை விட்டுவிட்டு எழுதக்கூடாது. Creche, Crush என்று இரண்டு வேறுபட்ட பொருண்மையையும் எழுத்துக்களையும் கொண்ட சொற்களை ஒரே ஒலிப்புமுறையில்தான் (Pronunciation) ஒலிக்கிறார்கள்; ஆனால் இவற்றில் இடம்பெறுகிற எழுத்துக்கள் வேறு. Cat என்ற சொல்லை ஒலிக்கும்போது அதில் உள்ள "c" என்ற எழுத்தை "k"என்று ஒலிக்கிறார்கள். ஆனால் எழுதும்போது அதை "c" என்றுதான் எழுதவேண்டும். ஆனால் cityஎன்ற சொல்லில் உள்ள "c" என்ற எழுத்தை "s"என்றுதான் ஒலிக்கவேண்டும். ஆனால் எழுதும்போது அதை "c" என்று எழுதவேண்டும் (இதற்கான விதிகள் ஆங்கில இலக்கணத்தில் உள்ளது). எனவே ஆங்கிலத்தில் சில சொற்களை அவற்றின் ஒலிப்பை வைத்துக்கொண்டு அவற்றின் எழுத்துக்களை முடிவுசெய்யமுடியாது. அதுபோன்று எழுத்துக்களை வைத்துக்கொண்டு உச்சரிப்பை அல்லது ஒலிப்பை முடிவு செய்ய இயலாது. இவைபோன்ற சிக்கல்களைத் தீர்ப்பதற்கு மொழியியல் ஆய்வு முதலில் தேவைப்படுகிறது என்பதை நாம் கவனத்தில் கொள்ளவேண்டும்.

எழுத்து - பேச்சு மாற்றி (Text to speech - TTS)

அடுத்து,கணினியானது ஒரு மொழிப் பனுவலில் காணப்படுகிற எழுத்துக்களைப்பேச்சொலிகளாகமாற்றுவதிலும் (Text to Speech - TTS) மேலே குறிப்பிட்ட சிக்கல்கள் உண்டு. 'அக்கா' 'காடு' என்ற தமிழ்ச்சொற்களில் சொல் இடையில் 'க்' இரட்டித்து வரும்போதும் சொல்முதலில் வரும்போதும் அவற்றின் ஒலிப்பு [k] தான். 'தங்கம்' போன்ற சொற்களில் சொல் இடையில் மெல்லின ஒலிக்கு அடுத்துவரும்போது 'க' எழுத்தானது [g] ஒலிக்கப்படவேண்டும். 'பகல்' 'காகம்' போன்ற சொற்களில் இரண்டு உயிர்களுக்கிடையில் 'க' வரும்போது [x] என்று ஒலிக்கவேண்டும். இந்த ஒலியன் - மாற்றொலி ஆய்வு இங்கு மிகவும் தேவைப்படுகிறது. அப்போதுதான் ஒரு சொல்லைச் சரியாக ஒலிக்கமுடியும்.

மேலும் தமிழ் போன்ற இரட்டைவழக்கு மொழிகளில் பேச்சுவழக்கில் சொற்களின் இறுதி மெய்களை ஒலிப்பது இல்லை. அல்லது இரட்டித்துவிடுகிறோம். 'மரம்' என்ற சொல்லை ஒலிக்கும்போது சொல்லிறுதி 'ம்' மறைகிறது. ஆனால் அதன் மூக்கொலிப் பண்பானது அதற்கு முந்தைய உயிரொலியை மூக்கொலித்தன்மைகொண்ட உயிராக மாற்றிவிடுகிறது. 'மண்

'கண்' 'நெய்' போன்ற சொற்களைப் பேச்சில் இறுதி எழுத்தை இரட்டித்து, 'இ' அல்லது 'உ' சேர்த்து 'மண்ணு' 'கண்ணு' 'நெய்யி' என்றுஒலிக்கிறோம். எனவே, பேச்சுத்தமிழை எழுத்துத்தமிழில் எழுதிக்காட்டவேண்டுமென்றால் இவைபோன்ற மொழி விதிகளை அறிந்திருக்கவேண்டும்.

இங்கு நான் குறிப்பிட விரும்புவது, கணினிமொழியியல், இயற்கைமொழி ஆய்வுகளுக்கு மொழியியல் அடிப்படைகள் மிகமிகத் தேவை என்பதே ஆகும்.

பேச்சுத்தொழில் நுட்பத்தில் இன்று மிகப் பெரிய மாற்றங்கள் ஏற்பட்டுள்ளன. எனவே, பேச்சு - எழுத்துமாற்றி அல்லது எழுத்து-பேச்சுமாற்றித் தொழில்நுட்பங்கள் நன்றாக வளர்ச்சியடைந்துள்ளன. இருப்பினும் அசையழுத்தம், ஒலி ஏற்ற இறக்கம் போன்ற மேற்கூற்று ஒலிகள் அல்லது ஒலியன்கள் (Suprasegmental Phones / Phonemes or Prosodic features)தொடர்பானவற்றில் சிக்கல்கள் உண்டு. அவையும் இன்றைய மொழித்தொழில்நுட்பத்தில் தீர்க்கப்படும் என்று எதிர்பார்க்கலாம். மேலும் புள்ளியியல் (Probablistic Statistics), செயற்கைச் செய்யறிவுத்திறன் (Artificial Intelligence - AI) போன்ற இன்றைய வளர்ச்சிகள் பேச்சுத் தொழில்நுட்பத்தில் சிறந்த சாதனைகளைப் புரிந்துவருகின்றன. இதனைப் பின்னர் கூறவிருக்கிறேன்.

பனுவல் அமைப்பு ஆய்வு (Text Processing)

எழுத்துப்பனுவல் அமைப்பு ஆய்வானது கணினிமொழியியல், இயற்கைமொழி ஆய்வுஆகியவற்றில்மிக மிகமுக்கியமானஇடங்களைப் பெறுகின்றது. ஒரு பனுவலில் உள்ள சொற்களை எவ்வாறு வேர்ச்சொல் - விகுதிகள் என்று பிரிப்பது, அவற்றின் இலக்கண வகைப்பாடுகளை எவ்வாறு கண்டறிந்து குறிப்பது, பெயர்த்தொடர், வினைத்தொடர் என்றுசொற்றொடர்களாக எவ்வாறு சொற்களை இணைத்து ஆய்வுசெய்வது,வாக்கியங்களை எவ்வாறு சொற்றொடர்களாகப் பிரிப்பது, தனிவாக்கியம்,கூட்டுவாக்கியம்,கலவை வாக்கியம்என்று எவ்வாறு வாக்கியங்களைப் பிரித்து ஆய்வு செய்வது போன்ற பல பணிகளை கணினிமொழியியலின் பனுவல் ஆய்வுப் பிரிவு மேற்கொள்கிறது.

பனுவல் ஆய்வை இரண்டு பிரிவுகளாகப் பிரித்து ஆய்வுசெய்யலாம். ஒன்று, கணினி உருபனியல் (Computational Morphology); மற்றொன்று, கணினித் தொடரியல் (Computational Syntax).

ஒரு மொழியின் அகராதியில் காணப்படும் அகராதிச்சொற்கள் (Lexicon or Word) இரண்டுவகைப்படும். ஒன்று, தனிச்சொல் அல்லது வேர்ச்சொல்; மற்றொன்று, தொகைச்சொல். இந்த இரண்டுவகைச் சொற்களும் அவற்றின் இலக்கண வகைப்பாடுகளுடன் - பெயர், வினை, பெயரடை, வினையடை போன்ற வகைப்பாடுகளுடன் -அகராதியில்இடம்பெறுகின்றன. மேலும் இந்த வகைப்பாடுகளில் உள்வகைப்பாடுகளும்உண்டு. தெரிநிலை வினை - குறிப்புவினை,செயப்படுபொருள் குன்றாவினை - செயப்படுபொருள் குன்றிய வினை, தன்வினை - பிறவினை போன்ற உள்வகைப்பாடுகள் தமிழில் உண்டு. மேலும் வினைச்சொற்கள் கால விகுதிகளை ஏற்பதில் உள்வகைப்பாடு உண்டு. இதுபோன்று பெயர்ச்சொற்களிலும் பன்மை விகுதி சேரும் பெயர், பன்மை விகுதி சேராத பெயர், உயர்திணைப் பெயர், அஃறிணைப் பெயர் போன்று பல உள்வகைப்பாடுகள் உண்டு. இந்த உள்வகைப்பாடுகளைப் பொறுத்துத்தான் சொற்களின் பயன்பாடு சொற்றொடர்களிலும் வாக்கியங்களிலும் வெளிப்படும்.

அகராதிச்சொற்கள் மொழிச்செயல்பாட்டில் தொடர்களிலும் வாக்கியங்களிலும் இடம்பெறும்போது தங்களுக்குரிய இலக்கணக் கூறுகளைப் பெற்று அமையும்.இவ்வாறுஇலக்கணப் பண்புகளைஏற்றசொற்களை (Inflected forms) சொல்வடிவங்கள் (Wordforms) என்று கூறலாம். ஒரு அகராதிச்சொல் மொழித்தொடரில், தான் ஏற்கும் இலக்கணக்கூறுகளின் அடிப்படையில் வேறுபட்ட சொல் வடிவங்களாக அமையும். தமிழில் "படி" என்ற அகராதி வினைச்சொல் 'படித்தான்' 'படித்த' 'படித்து' 'படிக்க' 'படிக்காமல்' என்று பலவகைப்பட்டசொல்வடிவங்களாக அமையலாம். 'புத்தகம்' என்ற அகராதி பெயர்ச்சொல் 'புத்தகத்தை' 'புத்தகத்தால்' புத்தகத்திற்கு' 'புத்தகங்கள்' என்று பலவகைப்பட்ட சொல்வடிவங்களை ஏற்கலாம். உருபனியலின் ஒரு அடிப்படைப் பணியானது ஒரு மொழியின் தொடர்களில் நீடிக்கிறசொல்வடிவங்களை வேர்ச்சொல், இலக்கணவிகுதிகள் என்று பிரித்து ஆராய்வது ஆகும். அதனடிப்படையில் குறிப்பிட்ட சொல்வடிவத்தின் இலக்கணக்குறிப்பை (Word-class / Parts-of-Speech - POS) அளிப்பதாகும். தொடர்களில் அல்லது வாக்கியங்களில் அமைகிற சொற்களின் - அதாவது சொல் வடிவங்களின் - பொருண்மையைப் புரிந்துகொள்ள மேற்குறிப்பிட்ட பகுப்பாய்வு தேவை. இதுவே உருபன் பகுப்பாய்வு(Morphological Parsing) என்று அழைக்கப்படுகிறது.

குறிப்பிட்ட மொழியின் சொல்வடிவங்களைப் பிரித்து ஆராய அந்த மொழியின் இலக்கணநூல் வழிகாட்டுகிறது. மொழியியலின் உருபனியல் பிரிவானது உலகமொழிகள் அனைத்துக்கும் பயன்படக்கூடிய உருபன் பகுப்பாய்வு

முறையை முன்வைக்கிறது. நைடா போன்ற மொழியியல் அறிஞர்கள் மொழிகளின் உருபன் பகுப்பாய்வுக்கான வழிமுறைகளை விதிகளாக விளக்கியுள்ளனர். இந்த விதிமுறைகளைச் சரியாக மேற்கொண்டால், ஒருவருக்குத் தெரியாத மொழியின் சொற்களையும்பகுத்து ஆராயமுடியும்.ஆனால் பிரித்து ஆராய்வதற்குரிய சொல்வடிவங்களின் பொருண்மை கொடுக்கப்பட்டிருக்கவேண்டும். அதன் அடிப்படையில் மொழியியலாளர்களால் எந்தவொரு இயற்கைமொழிக்கும் உருபன் பகுப்பாய்வை மேற்கொண்டு, அகராதிச்சொற்கள், இலக்கண விகுதிகளைப் பிரித்துக் கண்டறியமுடியும். உருபன் பகுப்பாய்வுக்காக மொழியியல் முன்வைக்கிற கோட்பாடுகளையும் வழிமுறைகளையும் கணினிக்கு எவ்வாறு அளிப்பது - கணினி நிரல்களாக எவ்வாறு அமைப்பது - என்பதுபற்றிய பிரிவே கணினி உருபனியல் (Computational Morphology) என்று அழைக்கப்படுகிறது.

ஒரு மொழியின் சொல்வடிவங்களைக் கணினியால் வேர்ச்சொல், விகுதிகள் என்று பிரித்து ஆராய முடிந்தால்தான், அந்தச் சொல்வடிவங்களின் இலக்கண வகைப்பாடுகளைக் - இலக்கணக் குறிப்புக்களைக் - கண்டறியமுடியும். அதனடிப்படையில்தான் அடுத்து சொல்வடிவங்கள் இணைந்து அமைகிற சொற்றொடர் அமைப்பையும் வாக்கியத்தின் அமைப்பையும் வாக்கியத்தின் பொருண்மையையும் கணினியால் அறிந்துகொள்ளமுடியும்.

மேற்கூறப்பட்டஉருபன் பகுப்பாய்வை மேற்கொள்வதற்கு மிக மிக அடிப்படையானது குறிப்பிட்ட மொழியின் அகராதிச்சொற்கள், இலக்கண விகுதிகள் இரண்டையும் கணினியால் எளிதாகப் பயன்படுத்தப்படும் வகையில் தரவுத்தளத்தில் அல்லது கணினி அகராதித்தளத்தில் - வைக்கப்படவேண்டும். இதற்குக் கணினியியலில் பின்பற்றப்படும்Regular Expressionஎன்ற சொல்வடிவம் மிகவும் பயன்படுகிறது. இந்தச் சொல்வடிவமானது சொற்களுக்கிடையே உள்ள பொதுமைக்கூறுகளின் அடிப்படையில் உருவாக்கப்படுகிற ஒரு கணித வடிவமாகும். இந்த வடிவத்தில் ஒரு மொழியின் அகராதிச் சொற்களும் இலக்கண விகுதிகளும் அமைந்திருந்தால், கணினியால் சொற்களை உருபன் பகுப்பாய்வு செய்வது எளிதாக அமையும்.

ஒரு மொழிக்கு மேற்கூறியவாறு அமைக்கப்பட்ட அகராதி உருவாக்கப்பட்ட பிறகு, Automata theory அடிப்படையில் இயங்கும் Finite State Automata என்ற நிரல் இயந்திரத்தின்மூலமாக அந்த மொழியின் சொல்வடிவங்களை வேர்ச்சொல், விகுதிகள் என்று பிரித்து ஆராய்வது எளிது. இந்த அடிப்படையிலான உருபன் பகுப்பாய்வு பொறியானது Finite State Transducer (FST) என்று அழைக்கப்படுகிறது. இந்தக் கணினிப்பொறியால் சொல் வடிவங்கள் வேர்ச்சொற்கள், விகுதிகள் என்று

பிரிக்கப்பட்டபிறகு அவற்றின் இலக்கண வகைப்பாடுகளை முடிவுசெய்வதும் இயலும். இந்த வழிமுறையின் அடிப்படையானது, ஒரு குறிப்பிட்ட மொழியில் வேர்ச்சொற்களும் இலக்கண விதிகளும் ஒரு குறிப்பிட்ட, திட்டவட்டமான பாதையில்தான் பயணித்து இறுதியில் சொல்வடிவங்களாக அமைகின்றன என்பதே ஆகும். எடுத்துக்காட்டாக, தமிழில் 'படித்துக்கொண்டிருக்கிறான்' என்ற சொல்வடிவமானது "படி (வேர்ச்சொல்) + த்து (செய்துவாய்பாட்டு வினையெச்ச விகுதி) + கொண்டிரு (வினைக்கூறு என்ற துணைவினை) + க்கிறு (நிகழ்கால விகுதி) + ஆன் (திணை- எண்-பால் விகுதி)" என்ற பாதையில்தான் திட்டவட்டமாகப் பயணிக்கிறது. அதுபோன்று "-த்து-" என்ற செய்து வாய்பாட்டு வினையெச்ச விகுதி வன்தொடர்க்குற்றியலுகரமாக இருப்பதாலும் அதையடுத்து வருகிற "கொண்டிரு" என்ற துணைவினையின் முதல் எழுத்து வல்லினமாக இருப்பதாலும் இவை இரண்டுக்கும் இடையில் "-க்-" என்ற ஒரு மெய் ஒற்று தமிழ்ப்புணர்ச்சிவிதியின் அடிப்படையில் சேர்கிறது. இறுதியில் "பிடித்துக்கொண்டிருக்கிறான்" என்ற சொல்வடிவம் கிடைக்கிறது.

மேற்கண்ட கணினி உருபனியல் வழிமுறையில் ஒரு வேர்ச்சொல்லோடு குறிப்பிட்ட விகுதிகளை இணைத்து ஒரு சொல்வடிவத்தைப் பெறவும் முடியும் (Generation of a Wordform) ; அதுபோன்று ஒரு சொல்வடிவத்தை வேர்ச்சொல், இலக்கண விகுதிகள் என்று பிரிக்கவும் முடியும் (Parsing of a Wordform).

உருபன் பகுப்பாய்வுக்கு வேறு பல வழிமுறைகளும் கணினிமொழியியலில் பின்பற்றப்படுகிறது. கணினியியலில் பின்பற்றப்படும் தரவுத் தள மென்பொருள்களின் உதவிகொண்டு, ஒரு மொழியின் அகராதியும், இலக்கண விகுதிகளும் கணினியில் சேமித்துவைக்கப்படலாம். அத்துடன் சொல் வடிவங்களின் அமைப்புவிதிகள் (Morphotactics) - அகராதிச்சொல்லும் விகுதிகளும் குறிப்பிட்ட வகையில் இணைவதற்கான விதிகளும் அவ்வாறு இணையும்போது நடைபெறும் சந்தி அல்லது புணர்ச்சி விதிகளும் (Morphophonemic rules) கணினி நிரல்வழியே கணினிக்கு அளிக்கலாம். இவற்றையெல்லாம் உள்ளடக்கிய கணினி நிரல்கள்மூலம் ஒரு மொழியின் சொல்வடிவங்களைப் பிரிக்கவும் அவற்றின் இலக்கண வகைப்பாடுகளைக் குறிக்கவும் இயலும்.

மேற்கூறிய வழிமுறையின் அடிப்படையில் ஒரு குறிப்பிட்ட மொழியின் சொல்வடிவங்களைப் பிரித்து ஆராய, வேர்ச்சொல், பின்னர் விகுதி என்று சொல்வடிவத்தின் முதல் பகுதியிலிருந்தும் தொடங்கலாம். இந்த முறையில் ஆய்வை மேற்கொண்டால் முதலில் வேர்ச்சொல் கிடைக்கும். அதைத் தொடர்ந்து

இலக்கண விகுதிகள் கண்டறியப்படும். இதற்கு மாறாக, ஒரு சொல்வடிவத்தின் பின்னாலிருந்து - அதாவது இலக்கண விகுதிகள், பின்னர் வேர்ச்சொல் என்றும் ஆய்வை மேற்கொள்ளலாம். இந்த இரண்டு வழிமுறைகளில் சாதகங்களும் உண்டு; பாதகங்களும் உண்டு.

கணினித் தொடரியல் (Computational Syntax) :

கணினிவழியே ஒருகுறிப்பிட்ட மொழியின் ஒரு வாக்கியத்தில் அமைந்துள்ள சொல் வடிவங்கள் ஆராயப்பட்டு, அவற்றிற்குரிய இலக்கண வகைப்பாடுகளும் அளிக்கப்பட்டுவிட்டால், அதன்பின்னர் மேற்கொள்ளவேண்டிய பணி, அந்தச் சொல்வடிவங்கள் எவ்வாறு சொற்றொடர்களாகவும் பின்னர் ஒரு வாக்கியமாகவும் அமைகின்றன என்பதை ஆராயும் தொடரியல் பணியே ஆகும். இப்பணிக்காக பலவகைப்பட்ட தொடரியல் கோட்பாடுகள் நிலவுகின்றன. அவற்றை இரண்டு பெரும் பிரிவுகளில் அடக்கலாம். ஒன்று, ஒரு வாக்கியத்தில் அமைந்துள்ள சொல்வடிவங்களைச் சொற்றொடர்களாகவும் பின்னர் வாக்கியமாகவும் ஆராய்கிற சொல், சொற்றொடர் இலக்கணவகைப் பிரிவு அடிப்படையிலான ஆய்வு Category- dependnent Syntax) ;மற்றொன்று, ஒரு வாக்கியத்தில் அமைந்துள்ளசொற்களுக்கு இடையிலான இலக்கண உறவுகள் அடிப்படையிலான ஆய்வு(Relation-dependnent Syntax).

முதல் வகைப்பாட்டில் பலவகைப்பட்ட மாற்றிலக்கணக் கோட்பாடுகள் அடங்கும். Generalized Phrase Structure Grammar (GPSG), Lexical Functional Grammar (LFG), Tree-Adjoining Grammar (TAG), Head-driven Phrase Structure Grammar (HPSC)போன்றவை ஆகும். இரண்டாவது வகைப்பாட்டில் Relational Grammar போன்றவை அடங்கும்.

கணினிப்பொருண்மையியல் - கணினி சூழல்சார் பொருண்மையியல் (Computational syntax - computational Pragmatics)

மேற்குறிப்பிட்ட தொடரியல் கோட்பாடுகளில் ஏதாவது ஒன்றை அடிப்படையாகக்கொண்டுஒரு குறிப்பிட்ட மொழியின் ஒரு தொடரைஆராய்ந்தபிறகு, அந்த ஆய்வின் அடிப்படையில் குறிப்பிட்ட வாக்கியத்தின் பொருண்மையைப் புரிந்துகொள்ளும் முயற்சி இது ஆகும். எந்தவொரு தொடரியல் கோட்பாட்டைப் பின்பற்றினாலும் அதனால் கிடைக்கிறஇறுதித் தொடர் அமைப்பானது அத்தொடரின் பொருண்மையைப் புரிந்துகொள்ள உதவவேண்டும். இந்த அடிப்படையில் குறிப்பிட்ட வாக்கியத்தின் பொருண்மையைப் புரிந்துகொள்வதற்கும் பல வழிமுறைகள்

உருவாக்கப்பட்டுள்ளன. இவையெல்லாம் கணினிப் பொருண்மையியல் (Computational Semantics) என்ற கணினிமொழியியல் பிரிவில் அடங்கும். இதன்வழி கிடைக்கும் ஒரு வாக்கியத்தின் பொருண்மையே இறுதிப் பொருண்மை கிடையாது. அந்த வாக்கியம் பயன்படுத்தப்பட்ட சூழலைப்பொறுத்துதான் இறுதிப்பொருண்மை கிடைக்கும். இதற்காக முன்வைக்கப்படுகிற கணினிமொழியியல் பிரிவானது கணினி சூழல்சார் பொருண்மையியல்(Computational Pragmatics) என்று அழைக்கப்படுகிறது.

மேற்கூறப்பட்ட கணினிமொழியியலின் அனைத்துப் பிரிவுகளும் இணைந்த ஒன்றே கணினிமொழியியல் ஆகும். இவை அனைத்தும் இயற்கைமொழிகளுக்கான மொழியியல் துறையைப் பெரும்பான்மையாகச் சார்ந்தவை ஆகும். மனித மூளையில் இடம் பெற்றிருக்கும் ஒரு குறிப்பிட்ட மொழிக்கான அகராதி, இலக்கணம் ஆகியவற்றைக் கணினிக்கேற்ற தரவுத்தளமாகவும் நிரல்களாகவும் எவ்வாறு மாற்றுவது என்பதுபற்றிய அறிவியலாக இன்று கணினிமொழி வளர்ந்துநிற்கிறது. அதனடிப்படையில் ஒரு குறிப்பிட்ட மொழியில் அமைந்துள்ள ஒரு கணினிக்கோப்பில்சொற்பிழை, ஒற்றுப்பிழை, தொடர்ப்பிழை ஆகியவற்றைத் தானாகக் கண்டறியும் சொல்லாளர் முதல் இயந்திரமொழிபெயர்ப்புவரை மொழித்தொழில்நுட்பம் வளர்ந்து நிற்கிறது. ஐபிஎம், கூகுள் போன்ற பன்னாட்டு கணினிநிறுவனங்கள் பலவகை மென்பொருள்களை உருவாக்கி அளித்துக்கொண்டிருந்தன.

கணினிக்கு உள்ள சிக்கல்

இருப்பினும் மனிதமூளைபோன்று கணினியால் நூறு விழுக்காடு இயற்கைமொழிகளைப் புரிந்துகொள்ளமுடியவில்லை. குறிப்பாக, பொருண்மை மயக்கம் இல்லாத வாக்கியங்களைப் புரிந்துகொள்கிற கணினியால் பொருண்மை மயக்கம் உள்ள வாக்கியங்களைப் புரிந்துகொள்வதில் சிக்கல் நீடிக்கிறது. குறிப்பாக, ஒன்றுக்குமேற்பட்ட பொருண்மையைக்கொண்ட சொற்கள் தொடர்களில் பயின்றுவரும்போது அத்தொடரில் அந்தக் குறிப்பிட்ட சொல்லின் பொருண்மையைப் புரிந்துகொள்வதில் சிக்கல் ஏற்படுகிறது. எடுத்துக்காட்டாக, "மலர்" என்ற சொல்லுக்குப் "பூ" என்ற பெயர்ச்சொல் பொருண்மையும் உண்டு; வினைச்சொல் பொருண்மையும் உண்டு. "மல்லிகை மலர் நல்ல மணத்தைத் தரும்" என்பதில் "மலர்" பெயர்ச்சொல்லாக அமைந்துள்ளது. "மல்லிகையே மலர் என்றால் நீ உடனே மலர்ந்துவிடுவாயா?" என்ற தொடரில் "மலர்" என்பது வினைச்சொல். இந்தக் குறிப்பிட்ட சொல் பயின்றுவருகிற வாக்கியத்தில் நீடிக்கிற பிற சொற்களின் துணைகொண்டுதான் குறிப்பிட்ட வாக்கியத்தில் "மலர்" என்ற சொல் பெயர்ச்சொல் பொருண்மையைத் தருகிறதா அல்லது வினைச்சொல்

பொருண்மையைத் தருகிறதா என்ற முடிவுக்கு வருகிறோம். இதுபோன்ற பொருண்மை மயக்கங்களை "அகராதிச்சொல் பொருண்மை மயக்கம் (Lexical ambiguity)" என்று அழைக்கிறார்கள். தொடரமைப்பினாலும் பொருண்மை மயக்கம் ஏற்படலாம். "முருகன் மருதமலையில் தனது சகோதரியைத் தொலைநோக்கியுடன் பார்த்தான்" என்ற தொடருக்கு "முருகன் தன் கைகளில் தொலைநோக்கியை வைத்துக்கொண்டு மலையில் நின்றுகொண்டிருந்த தனது சகோதரியைப் பார்த்தான்" என்றும் பொருள் கொள்ளலாம்; "தொலைநோக்கியைத் தன் கைகளில் வைத்துக்கொண்டு மலையில் நின்றுகொண்டிருந்ததனது சகோதரியை முருகன் பார்த்தான்" என்றும் பொருள் கொள்ளலாம். இங்குப்பொருண்மை மயக்கத்திற்குக் காரணம், "தொலைநோக்கியுடன்" என்ற சொல்லானது "முருகன்" என்ற சொல்லோடு தொடர்புகொள்கிறதா அல்லது "சகோதரி" என்ற சொல்லோடு தொடர்புகொள்கிறதா என்ற அமைப்பு வேறுபாட்டைப் பொறுத்துத்தான் இந்தத் தொடர் அமைப்பு பொருண்மை மயக்கத்தின் (Structural ambiguity) தீர்வு அமையும். தொடரியல் ஆய்வானது இந்த இரண்டு அமைப்புக்களையும் அளிக்கும். இங்குப் பொருண்மை மயக்கத்தைத் தீர்க்க, இந்தக் குறிப்பிட்ட சொற்களுக்கு முன்னர் அல்லது பின்னர் அமைகிற வாக்கியங்கள் உதவலாம்.

நிகழ்தகவு மொழியியல் (Probabilistic Lingustics)

கணினிமொழியியலில் மேலே விளக்கிய அமைப்புசார்ந்த வழிமுறைகள் தவிர, புள்ளியியல் சார்ந்த அணுகுமுறையும் பயன்படத் தொடங்கியது. ஏராளமான வாக்கியங்களைக்கொண்ட தரவுகளை அடிப்படையாகக்கொண்டு, ஒரு குறிப்பிட்ட சொல் அல்லது வாக்கியம் இரண்டு பொருண்மைகளைத் தரும்போது, நிகழ்தகவுப் புள்ளியியலை (Probablistic Statistics) அடிப்படையாகக்கொண்டு குறிப்பிட்ட சூழலில் எது சரியான பொருண்மை என்பதை முடிவுசெய்யும் வழிமுறைகள் தற்போது பயன்படுத்தப்படுகின்றன.

நிகழ்தகவுப் புள்ளியியலின் வளர்ச்சியும் கோடிக்கணக்கான சொற்களைக்கொண்ட தரவுகளைச் சேமித்துவைத்து, ஆய்வுகளை மேற்கொளவதற்கான கணினித் தொழில்நுட்பத்தின் வளர்ச்சியும் இன்று கணினிமொழியாய்வில் மிகப்பெரும் மாற்றங்ளை ஏற்படுத்தியுள்ளது. மொழியியலிலும் "நிகழ்தகவு மொழியியல் (Probablistic Linguistics)" என்ற ஒரு பிரிவு தோன்றி வளர்ந்து நிற்கிறது. எடுத்துக்காட்டாக, ஒரு மொழிக்கான சொற்பிழைதிருத்தி மென்பொருளானது ஒருவரின் உரை அல்லது பனுவலை ஆய்வுசெய்து, அதில் காணப்படும் பிழைகளைக் கண்டறிந்தபிறகு, அந்தப் பிழையான சொல்லுக்குரிய சரியான சொல்லைத் தரவேண்டும். ஒரு பிழையான சொல்லுக்கு ஏராளமான சரியான சொற்களைக் கணினி அளிக்கும். அவற்றில்

எந்தச் சொல் மிகவும் சரியான சொல்லாக இருக்கிறது என்பதை அறிந்துகொள்ள நிகழ்தகவுப் புள்ளியியல் உதவுகிறது. இதுபோன்று கணினி ஒலியனியல், கணினித் தொடரியல், கணினிப் பொருண்மையியல் ஆகியவற்றிலும் நிகழ்தகவுப் புள்ளியியல் இன்று அதிக அளவில் பயன்படுத்தப்படுகிறது. இயற்கைமொழி ஆய்வில் நிகழ்தகவுப் புள்ளியியல் எவ்வாறு பயன்படுகிறது என்பதுபற்றிமேனிங் (Christopher D. Manning), ஹின்ரிச் (Hinrich Schutze) இருவரும் இணைந்து "Foundationsof Statistical Natural Language Processing"என்ற ஒரு நூலை வெளியிட்டுள்ளார்கள்.

இன்றைக்குக் கணினிமொழியியல், இயற்கைமொழி ஆய்வு இரண்டிலும் நிகழ்தகவுப் புள்ளியியல் மிகவும் பயன்படுகிறது. குறிப்பாக, இன்று பேச்சுத் தொழில்நுட்பம்(Speech Processing), கணினிவழி தானியங்கு மொழிபெயர்ப்பு (Automatic Machine Translation) போன்றவற்றில் இதன் பயன்பாடு அதிகம்.

மொழிகளின் இலக்கணங்கள், மொழியியல் ஆகியவற்றில் முன்வைக்கப்படுகிற விதிகளை அடிப்படையாகக்கொண்டு வளர்ந்துவந்த கணினிமொழியியலில் இன்று நிகழ்தகவுப் புள்ளியியலைப் பயன்படுத்தும் வளர்ச்சிநிலை நன்கு வெளிப்படுகிறது. இதற்கு அடிப்படைக் காரணம், இயற்கைமொழிகளில் அகராதிச்சொற்களும் விகுதிகளும் தொடரமைப்புகளும் பல நேரங்களில் பொருண்மை மயக்கத்தையும், மொழியமைப்பு மயக்கத்தையும் தருகின்றன. ஒரு சொல்லுக்குப் பல பொருண்மைகள், ஒரு தொடரமைப்புக்கு ஒன்றுக்கு மேற்பட்ட பொருண்மைகள் என்று இயற்கைமொழிகள் அமைந்திருப்பதே இதற்குக் காரணம். அதனால்தான் கணினி நிரல்களுக்கான மொழியாக இயற்கைமொழிகளைப் பயன்படுத்தமுடியவில்லை: ஒரு சொல்லுக்குஒரு பொருண்மை, ஒரு தொடரமைப்புக்கு ஒரு பொருண்மை என்ற விதிகளை உள்ளடக்கிய கணினிக்கென்றே உருவாக்கப்பட்ட பல நிரலாக்கமொழிகள் இன்று நீடிக்கின்றன. நடைமுறையில் மனிதர்கள் இயற்கைமொழிகளில் தோன்றுகிற இதுபோன்ற மயக்கங்களை வாக்கியச் சூழல், பேசும் பொருளின் அறிவு, மொழிசாராச் சூழல் ஆகியவற்றின் உதவியுடன் தீர்த்துக்கொள்கிறார்கள். ஆனால் இந்தத் திறன் கணினிக்கு இல்லாத காரணத்தினால், அதற்கு மாற்றாகத் தற்போது நிகழ்தகவுப் புள்ளியியல் நீடிக்கிறது.

இவ்வாறு, இலக்கணங்கள், மொழியியல் ஆகியவை முன்வைக்கும்மொழி அமைப்பு விதிகளைமட்டுமே அடிப்படையாகக்கொண்டு செயல்பட்டுவந்த கணினிமொழியியல், இன்று நிகழ்தகவுப் புள்ளியியலையும் உள்ளடக்கிய ஒரு துறையாக வளர்ந்துள்ளது.

செயற்கை நரம்புப் பின்னல் (Artificial Nerual Network - ANN)

அடுத்த கட்ட வளர்ச்சியாக, மனித மூளையின் அமைப்பு, செயல்பாடுகள் ஆகியவற்றை அடிப்படையாகக்கொண்ட "செயற்கை நரம்புப் பின்னல் (Artificial Neural Network - ANN)' இன்று கணினித் தொழில்நுட்பத்தில் மிகுந்த செல்வாக்கைச் செலுத்திவருகிறது. மனித மூளையின் இயற்கைமொழித் திறனுக்கு மிக மிக அடிப்படை மூளையில்நிலவுகிற கோடிக்கணக்கான நரம்பணுக்களும் அவற்றின் வலைப்பின்னல் அமைப்பும்தான் அடிப்படை. எனவே இந்த நரம்பு வலைப்பின்னல் போன்ற கணினி நிரல்கள்மூலம் நமக்கு உள்ள இயற்கைமொழி அறிவைக் கணினிக்கும் கொடுக்கலாம் என்ற ஒரு கருத்து இன்று வளர்ந்துநிற்கிறது. செயற்கை நரம்புவலைப் பின்னலிலும் ஒன்றுக்குமேற்பட்ட ஆய்வுவழிமுறைகள் நிலவுகின்றன.

20-ஆம் நூற்றாண்டில் தோன்றிய கணினிமொழியியல் மேற்கண்டவாறு மொழி விதிகள் அல்லது இலக்கணம், நிகழ்தகவுப் புள்ளியியல், செயற்கை நரம்பு வலைப்பின்னல் ஆகியவற்றின் அடிப்படைகளைப் பின்பற்றி சிறந்தமுறையில் இன்று வளர்ந்துநிற்கிறது.

செயற்கைச் செய்யறிவுத்திறனும்இயற்கைமொழி ஆய்வும் (Aritifical Intelligence and Natural Language Processing)

கணினிமொழியியலானது மேற்குறிப்பிட்ட வழிமுறைகளின் வாயிலாகத் தான் பெற்ற வளர்ச்சியின் ஒரு உச்சகட்ட நிலையை இன்று - 21-ஆம் நூற்றாண்டில் - பெற்றுள்ளது. மேற்கண்ட கணினிமொழியியலின் மூன்றுவகையான ஆய்வுகளும் இணைந்த ஒன்றாகச் "செயற்கைச் செய்யறிவுத்திறன்சார் இயற்கைமொழி ஆய்வு"இன்று வளர்ந்துநிற்கிறது.

சென்றநூற்றாண்டுவரை இயற்கைமொழிகளின் சொற்களஞ்சியங்களையும் இலக்கண அமைப்புக்களையும் கணினிக்கேற்றவாறு பல வழிமுறைகளில் - மொழி விதிகளைக் கற்றுக்கொடுத்தல், நிகழ்தகவு புள்ளியியல்வழியே மொழி அறிவை அளித்தல், செயற்கை நரம்புவலைப்பின்னல்வழியே மொழி அறிவைக் கொடுத்துச் செயல்படவைத்தல் ஆகிய மூன்று வழிமுறைகளில் - கணினிக்குக் கொடுத்து இயற்கைமொழி ஆய்வுகளையும் மொழித்தொழில்நுட்பச் செயல்பாடுகளையும் கணினிமொழியியல் மேற்கொண்டுவந்தது. இங்கு நாம் கவனத்தில் கொள்ளவேண்டிய ஒரு முக்கியச் செய்தியானது கணினிக்கு இயற்கைமொழிகளைப் பல வழிமுறைகளில் கற்றுக்கொடுத்து, நமக்கு வேண்டிய மொழித்தொழில்நுட்பப் பயன்பாடுகளைப் பெற்றுவந்தோம். இங்குக் கணினிக்கு

நாம்தான் இயற்கைமொழி அறிவைக் கற்றுக்கொடுத்து வந்தோம். கணினியானது நாம் கற்றுக்கொடுத்த மொழி அறிவைப் பயன்படுத்தித்தான் நமக்குப் பலவகை மொழித்தொழில்நுட்பப் பயன்பாடுகளை - பேச்சு - எழுத்து மாற்றி, எழுத்து-பேச்சு மாற்றி, சொல்லாளர், பனுவல் சுருக்கம், மொழிபெயர்ப்பு போன்ற பயன்பாடுகளை-பெற்றுவந்தோம். இங்கு நாம்தான் கணினிக்கு மொழி ஆசிரியர். கணினியானது நம்மிடமிருந்து கற்றுக்கொண்ட மொழி அறிவைத்தான் பயன்படுத்தியது.

21-ஆம் நூற்றாண்டில் ஒரு மகத்தான வளர்ச்சி. கணினி தானே ஒரு இயற்கைமொழியின் செய்யறிவுத்திறனைக் கற்றுக்கொள்ளும் வளர்ச்சி (Self-learning). ஒரு மொழி இல்லை, நூற்றுக்கணக்கான மொழிகளுக்குமான செய்யறிவுத்திறனை ஒரே நேரத்தில் தானே கற்றுக்கொண்டு செயல்படும் தொழில்நுட்பத்தைக் கணினி பெற்றுள்ளது. இது எப்படி சாத்தியம் ஆனது ?

கணினி மூளையின் - மின்னணுச் சில்லுவின் - மகத்தான வளர்ச்சி, கோடியேகோடி மொழித்தரவுகளைச் சேகரிக்கும் வசதிகொண்ட இணையம், அத்தரவுகளைச் சேமித்துவைக்கும் கணினிச் சேமிப்பகம் போன்றவற்றின் வளர்ச்சியால் இது சாத்தியமாகியுள்ளது. கடந்த 20-ஆம் நூற்றாண்டில் நிகழ்த்தப்பட்ட கணினிமொழியியல் ஆய்வுமுறைகள் அனைத்தும் இணைந்த ஒரு புதிய மொழித்தொழில்நுட்பாக இது மலர்ந்துள்ளது.

இவ்விடத்தில் பிறந்த குழந்தை எவ்வாறு மொழியைப் பெற்றுக்கொள்கிறது என்ற மொழியியல் கோட்பாட்டை மீண்டும் நினைவில் கொள்ளவேண்டும். மொழியியல் அறிஞர் நோம் சாம்ஸ்கியின் கோட்பாடா அல்லது நடத்தைசார் உளவியலார் எஸ். எப். ஸ்கின்னர் கோட்பாடா என்பதுபற்றி முன்னரே பார்த்தோம். சாம்ஸ்கியின் கோட்பாட்டின்படி, குழந்தை தன் மூளையில் பிறக்கும்போதே நீடிக்கிற பொதுமை இலக்கணத்தைக்கொண்டு தன்னுடைய தாய்மொழியின் இலக்கணத்தை உருவாக்கிக்கொள்கிறது. சுற்றுப்புற மொழித்தரவுகள்தேவை இல்லைஎன்று சாம்ஸ்கிகூறவில்லை.மாறாக, மூளையில் நீடிக்கிற பொதுமைஇலக்கணத்தைத் தனது குறிப்பிட்ட மொழிக்கான இலக்கணமாக மாற்றி அமைத்துக்கொள்ள குழந்தைக்கு மொழித்தரவுகள் தேவை. குழந்தைக்கு அதன் செவித்திறன் வழியே மூளையில் உள்ள பொதுமை இலக்கண மொழிப்புலத்திற்குத் தரவுகள் போய்ச்சேரவேண்டும். அப்போதுதான் குழந்தையின் மூளையில் குறிப்பிட்ட மொழிக்கான இலக்கணம் தோன்றி வளரமுடியும். இருப்பினும் குழந்தையின் மூளையில் அமைந்துள்ள பொதுமை இலக்கண மொழிப்புலன்தான் தனக்குக் கிடைத்த மொழித்தரவுகளைக் கொண்டு, குறிப்பிட்ட மொழியறிவைக் குழந்தைக்கு அளிக்கிறது. எனவே நாம் குழந்தைக்கு

மொழியைக் கற்றுக்கொடுக்கவில்லை; குழந்தையும் கற்கவில்லை; மாறாக, அதன் மூளையில் உள்ள பொதுமை இலக்கண மொழிப்புலமே குறிப்பிட்ட மொழி அறிவைப் பெறுகிறது.

ஆனால் ஸ்கின்னர் கருத்துப்படி, குழந்தைக்கு நாமும் நமது சுற்றுப்புறமும்தான் மொழியைக் கற்றுக்கொள்கிறது. அதற்குத் தேவையான மொழித்தரவுகளைப் பெற்று, குறிப்பிட்ட மொழிக்கான மொழியறிவை அது கற்றுக்கொள்கிறது. குழந்தை பிறக்கும்போதே அதற்கு பொதுமை இலக்கணம் என்ற மொழிப்புலம் இருப்பதை ஸ்கின்னர் ஏற்றுக்கொள்ளவில்லை.

தற்போதைய செயற்கைச் செய்யறிவுத்திறன்சார் கணினிமொழியியலுக்கு ஸ்கின்னரின் மொழிக்கற்றல் கோட்பாடு அடிப்படை என்று கூறலாம். கணினிக்கு ஒரு மொழியின் அனைத்துப் பண்புகளையும் வெளிப்படுத்தும் மொழித்தரவு அளிக்கப்படவேண்டும். அவ்வாறு அளிக்கப்பட்டால் கணினியின் செய்யறிவுத்திறன் செயல்பட்டு, அந்தக் குறிப்பிட்ட மொழியைத் தானே கற்றுக்கொள்ளும்.இதை ஆங்கிலத்தில் Self-Learning / Self-Attention என்று அழைக்கிறார்கள். அதாவது நாம் எந்தவித மேற்பார்வையும்அல்லதுஉதவி யும்செய்யாமலேயே,கணினி தானே ஒருமொழியின் செய்யறிவுத்திறனைக் கற்றுக்கொள்ளும் (Unsupervised Learning). நாம் எந்தவித உதவியும் செய்யாமலேயே கணினி தானே இத்திறமையை கற்றுக்கொள்வதற்கான கணினி வடிவமைப்பு (Architecture), அதற்குள்ளே செயல்படும் செயல் வழிமுறைகள் (Algorithms) இரண்டையும் கொண்டு, தனக்கு அளிக்கப்பட்ட எண்மத் தரவுகளின் அடிப்படையில் அந்த மொழிக்கான செய்யறிவுத்திறனைப் பெற்றுக்கொள்கிறது. இந்த முறையில் உருவாக்கப்பட்டபெரிய மொழிமாதிரிகளை (LLM) முன்கூட்டியே பயிற்சியளிக்கப்பட்ட மாதிரிகள் (Pre-trained Model)என்று அழைக்கின்றனர்.

மேற்கூறிய மொழி மாதிரிகளில் ஒரு தொடரில் அமைகிற சொல்லின் பொருள், இலக்கண வகைப்பாடு ஆகியவை அந்தக் குறிப்பிட்ட சொல்லுக்கு முன்னர் அமைகின்ற சொற்கள், முன்னர் அமைகின்ற வாக்கியங்கள் ஆகியவற்றை அடிப்படையாகக்கொண்டுதான் தீர்மானிக்கப்படுகின்றன. இதை "சூழல்சார்சாளரங்கள் (Context - Windows)" என்று அழைக்கிறார்கள். தற்போதைய செயற்கை அறிவுத்திறனின் பெரிய மொழி மாதிரிகளில் ஒரு சொல்லுக்கு முந்தைய 300 பக்க அளவிலான தொடர்களை அடிப்படையாகக் கொண்டு, அச்சொல்லின் பண்புகளைத் தீர்மானிக்கமுடியும். இந்த முறையில் நான் முன்னரே குறிப்பிட்ட "நிகழ்தகவுப் புள்ளியியல்" பெரும்பங்குவகிக்கிறது. எனவே நூறு விழுக்காடு அவற்றின் செயல்பாடு சரியாக இருக்கும் என்று கூறமுடியாது.

ஒரு குறிப்பிட்ட சொல்லானது கொடுக்கப்பட்ட மொழித்தரவுகளில் அமைகிற இடத்தின் நிகழ்தகவைக்கொண்டுதான் அறியப்படுகிறது. சில மொழிமாதிரிகளில் குறிப்பிட்ட சொல்லின் அடுத்த சொற்களும் கணக்கில் கொள்ளப்படுகின்றன.

பல ஆயிரம் கோடிகள் அளவில் மொழிகளின் தரவுகள் சேகரிக்கப்பட்டு, கணினியில் சேமித்து வைக்கப்பட்டால், அவற்றைப் பயன்படுத்தி மேற்கூறிய கணினி வடிவமைப்பும் செயல்வழிமுறைகளும் செயல்பட்டு, குறிப்பிட்ட மொழியின் செயலறித் திறனைப் (Language Performance) பெற்றுக்கொள்ளமுடியும். சாம்ஸ்கி கூறுகிற உயிரியல் அடிப்படையான மொழிப்புலம் போன்று (biological-based langauge faculty). இந்தக் கணினிவடிவமைப்பும் செயல்வழிமுறைகளும் செயல்பட்டு மொழிச் செயலறித்திறனைப் பெற்றுக்கொள்கிறது. ஆனால் ஒரு வேறுபாடு. முதலில் அமைக்கப்படுகிற கணினிவடிவமைப்பிலும் வழிமுறைகளிலும் இயற்கைமொழிகளின் மொழி அறிதிறனோ (linguistic competence) அல்லது மொழிச் செயலறிதிறனோ(linguistic peformance) கிடையாது. ஒன்றும் எழுதப்படாத ஒரு கரும்பலகைதான் அது. மொழித்தரவுகள் கிடைத்தபிறகுதான் கணினிக்குக்குறிப்பிட்ட மொழிகளின் மொழிச் செயலறிதிறன் கிடைக்கிறது. இவ்வாறு உருவாக்கப்பட்ட கணினித்தொழில்நுட்பம்தான் "மிகப்பெரிய மொழி மாதிரி (Large Language Model - LLM)" என்று அழைக்கப்படுகிறது.

ஆனால் இதற்கு முதல் தேவை மிகப் பெரிய அளவிலான, எண்ம வடிவத்தில் அமைந்துள்ள மொழித்தரவுகளாகும் (Digital Corpus). தற்போதைய சூழலில் ஆங்கிலம் போன்ற மொழிகளுக்கு இந்தத்தொழில்நுட்பம் எதிர்பார்க்கிற மிகப்பெரிய அளவிலான மொழித்தரவுகள்கிடைக்கின்றன. இதுபோன்ற மொழிகளை "அதிகத் தரவு கிடைக்கும் மொழிகள் (High-Resource languages)" என்று அழைக்கிறார்கள். எனவே ஆங்கிலம் போன்ற மொழிகள்வழியே பலவகை மொழிச்செயலறிவுத் திறனைத் தற்போதைய "செயற்கைச் செய்யறிவுத்திறன்சார் மொழித்தொழில்நுட்பம்" அளிக்கிறது. ஆனால் தமிழ்போன்ற மொழிகளுக்கு அந்த அளவு எண்ம வடிவ மொழித்தரவுகள் கிடைப்பதில்லை. இதுபோன்ற மொழிகளைக் "தரவு குறைவாகக்கிடைக்கி றமொழிகள் (Low-Resource languages) என்று அழைக்கிறார்கள்.

செயற்கைச்செய்யறிவுத்திறன்சார் பெரிய மொழி மாதிரிகள் (Large Language Model - LLM) நாம் வியக்கத்தக்கவகையில் மொழிச்செயல்பாடுகளை மேற்கொள்கின்றன. பேச்சு - எழுத்து மாற்றி, எழுத்து-பேச்சு மாற்றி, சொல்லாளர், பனுவல் சுருக்கம், தானியங்கு மொழிபெயர்ப்பு உட்பட இன்றைய கணினிமென்பொருள்கள் எல்லாம் இந்த மாதிரிகளை அடிப்படையாகக்கொண்டு

உருவாக்கப்பட்டு, மிகப் பெரிய தாக்கத்தைக் கணினிமொழியியல், இயற்கைமொழி ஆய்வு, மொழித்தொழில்நுட்பம் ஆகியவற்றில் ஏற்படுத்திவருகின்றன. விநாடிகளில் நாம் கேட்கிற வினாக்களுக்கு விடைகள் அளிக்கின்றன; நமக்குத் தேவையான கட்டுரைகள், கடிதங்கள், படைப்பிலக்கியங்களைக்கூட இவை தயாரித்துக்கொடுக்கின்றன; பெரிய கட்டுரைகள் அல்லது நூல்களின் சுருக்கங்களை விரைவாகத் தருகின்றன; மொழிபெயர்ப்புப்பணிகளை மிக விரைவாகவும் நன்றாகவும் செய்துகொடுக்கின்றன; மேலும் தற்போது மொழிவழிப் பனுவல்களை (Language texts) மட்டுமல்லாமல், குரல் (Voice), நிழல் உருவம் அல்லது பிம்பம் (Image) ஆகிய பிற வடிவங்களிலும் தங்களது பணிகளை மேற்கொள்கின்றன. இது தற்போது "உருவாக்கச் செய்யறிவுத்திறன் (Generative AI)" என்று அழைக்கப்படுகிறது.

மேற்கூறிய மிகப் பெரிய மொழிமாதிரிகளை உருவாக்குவதில் "செயற்கை நரம்புவலைப்பின்னல்" உட்பட வேறுபட்ட கணினித் தொழில் நுட்பங்கள் பயன்படுத்தப்படுகின்றன. குறிப்பாக, "மிக ஆழமாகக் கற்றல் (Deep Learning)" வழிமுறை அதிக அளவில் பயன்படுத்தப்படுகின்றது. அதிகமான எண்ணிக்கையில்('ஆழமான') கற்பதற்கான அடுக்குகள் (Layers) இதில் இருக்கும். அதுபோன்று "மாற்றி - வடிவமைப்பு(Transformer Architecture)" போன்று பல வடிவமைப்புக்கள் இத்துறையில் பின்பற்றப்படுகின்றன. கூகுள், ஓபன் ஏ ஐ, மெடா என்ற முகநூல் நிறுவனம், ஆந்த்ரோபிக் போன்றபல பன்னாட்டு நிறுவனங்கள் பல பில்லியன் கோடி டாலர்களை முதலீடு செய்து செய்யறிவுத்திறன்சார் பெரிய மொழிமாதிரிகளைத் தயாரித்துள்ளன.

தமிழ் போன்ற எண்மத் தரவுகள் குறைவாக இருக்கிற மொழிகளையும் மேற்கூறிய பெரிய மொழிமாதிரி கையாளக்கூடிய வகையில் வளர்த்தெடுப்பதற்கு அல்லது ஒரு குறிப்பிட்ட மொழிக்கான சில குறிப்பான கணினிவழிச் செயல்பாடுகளுக்கான மென்பொருள்களை உருவாக்குவதற்கு வழிகள் முன்வைக்கப்பட்டுள்ளன. ஏதாவது ஒரு மொழிமாதிரியைத் தேர்ந்தெடுத்து, அதற்குக் குறிப்பிட்ட மொழியின் எண்மத் தரவுகளை மேலும் அளித்து, பெரிய மொழிமாதிரியை நமது தேவைகளுக்கேற்ப நுட்பமாக ஆக்கிக்கொள்ளலாம். இந்த வழிமுறையை "நுட்பத் திறன் மேம்படுத்தம் (Fine-tuning)" என்று அழைக்கிறார்கள்.

மேற்குறிப்பிட்ட "நுட்பத் திறன் மேம்படுத்தலுக்கு" குறிப்பிட்ட மொழியின் எண்மத் தரவுகளை அவற்றிற்கான இலக்கணக் குறிப்புகள், பிற விவரங்களுடன் கணினிக்கு அளிக்கவேண்டும். இதன்மூலம் கணினியில்

ஏற்கனவே நிலவிய "பெரிய மொழிமாதிரியின்"மொழிச் செயலறிதிறன் மேலும் அடுத்த உயர்கட்டத்தை எட்டும். இதை "மனித மேற்பார்வையிலான மொழிக் கற்றல் (Supervised Learning)" என்று அழைக்கிறார்கள்.

ஆயிரக்கணக்கான ஆண்டுகளுக்குமுன் ஒரு குறிப்பிட்ட மொழியின் இலக்கணமே மொழிஆய்வு என்று நிலவிய ஒரு நிலையானது மொழிகள் அனைத்துக்கும் பொதுவான மொழியியலாக மாறி, பின்னர் கணினிமொழியியலாகவும் வளர்ச்சியடைந்து, இன்று பன்மொழித்திறன்கொண்ட செயற்கைச் செய்யறிவுத்திறன்சார் பெரிய மொழிமாதிரிகளின் (Artificial Intelligence Large Language Model - AI LLM) உருவாக்கமாக வளர்ச்சியடைந்துள்ளது என்பது இங்குக் குறிப்பிடத் தக்கது.

ஆனாலும் இங்கு ஒரு செய்தியைக் கவனத்தில் கொண்டுவர விரும்புகிறேன். செயற்கைச் செய்யறிவுத்திறன்சார் மொழிமாதிரிகளைக் கூகுள், மெடா (முகநூல்), மைக்ரோசாஃப்ட் போன்ற நிறுவனங்களும்தான் தற்போதைய நிலையில் உருவாக்கமுடியும். அதற்குக் காரணம், ஆயிரக்கணக்கான கோடி பணம் இதற்குத் தேவைப்படுகிறது. ஆயிரக்கணக்கானவர்களின் பங்களிப்பு அல்லது உழைப்பு தேவைப்படுகிறது. மிகப் பெரிய அளவிலானகணினிக்கட்டமைப்பு தேவைப்படுகிறது. எனவே இந்த நிறுவனங்கள் தயாரித்துக்கொடுக்கும் மென்பொருள்களை நாம் பயன்படுத்தமுடியுமே தவிர நாமே அவற்றை உருவாக்குவது என்பது இன்றைய நிலையில் கடினம். அந்த நிறுவனங்களின் அடிப்படை நோக்கம் வணிகமே. அதற்கு எந்த அளவு மொழிமாதிரிகள் பயன்படும், பயன்படவேண்டும் என்பதில்தான் அவை கவனம்செலுத்தும். வெறும் தரவுகளையும் நிகழ்தகவுப் புள்ளியியலையும் கணினிசார் வழிமுறைகள், கணினிக்கட்டமைப்பு போன்றவற்றையும் வைத்துக்கொண்டு, இயற்கைமொழிகளை நமது பல்வேறு மொழிச்செயல்பாடுகளுக்கு எவ்வாறு பயன்படுத்தலாம் என்பதற்கான வணிக நோக்கிலான பணிகளே இவற்றின் நோக்கமாகும். இயற்கைமொழிகளின் மொழிக்கூறுகளை - அகராதி, இலக்கணம் போன்றவற்றின் அமைப்புக்கூறுகளை - ஆய்வுசெய்து, அதனடிப்படையில் மொழிகள்பற்றிய மொழியியல் ஆய்வுகளை முன்னெடுத்துச்செல்வது அவற்றின் நோக்கம் இல்லை. கணினிக்கு மனிதமூளையில் அமைந்துள்ள "மொழி அறிதிறன்" போன்ற மொழியறிவை அளிப்பது அதன் நோக்கம் இல்லை. எளிமையாகச் சொல்லப்போனால், இன்றைய செய்யறிதிறன்சார் மென்பொருள்களை 1964-67 ஆம் ஆண்டுகளில் ஜோசப் வெய்சன்பாம் (Joseph Weizenbaum) உருவாக்கிய "எலிசா (Eliza)" மென்பொருளின் ஒரு நீட்சியாகப் பார்க்கலாம். நாம் விரும்பும் மொழிச்செயல்களைச் செய்கிறது; நாம் முன்வைத்த

வினாக்களுக்கு விடை தருகிறது. இதில் ஐயம் இல்லை. ஆனால் குறிப்பிட்ட மொழியில் நாம் அதற்கு முன்வைக்கிற ஐயத்தையோ வினாவையோ மனித மூளையானது குறிப்பிட்ட மொழியின் மொழியறிவை - மொழி அறிதிறனை- அடிப்படையாகக்கொண்டு புரிந்துகொள்வதுபோல புரிந்துகொள்வதில்லை. அவற்றில் உள்ள பெரிய மொழி மாதிரிகளுக்கு மனித மூளையில் உள்ள மொழி அறிவு - மொழி அறிதிறன் - கிடையாது. ஆனாலும் தன்னைப் பயிற்றுவிக்க ஆய்வாளர்கள் அளித்த தரவுகளையும் நிகழ்தகவுப் புள்ளியியல், செயற்கை நரம்பு வலைப்பின்னல்போன்றவற்றை அடிப்படையாகக் கொண்ட மொழிச் செயலறிதிறனைப் பயன்படுத்தி, நமக்குத் தேவையான மொழிபெயர்ப்புபோன்ற மொழிவழிச் செயல்பாடுகளைச் செய்து தருகின்றன. அதாவது ஒரு வகையான "மொழிச் செயலறிதிறன்" அவற்றிற்கு உண்டு. "எலிசா" மென்பொருளுக்கு இருந்ததுபோல. பயனாளர்களுக்குத் தேவை கணினியின் "மொழிச் செயலறிதிறன்" தான் ; அந்த மொழித்திறனுக்கு அடிப்படையானதாக "மொழி அறிதிறன்" - மனித மூளையில் இருப்பதுபோன்ற மொழி அறிவு - இருக்கிறதா, அல்லது முன்பு குறிப்பிட்ட "தரவு, நிகழ்தகவு, ஆழக் கற்றல்" போன்றவற்றை அடிப்படையாகக்கொண்ட அறிவு இருக்கிறதா என்பதுபற்றிக் கவலை இல்லை. அதாவது "மொழிச்செயல்" நடைபெறுகிறது. அதுவே பயனாளர்களுக்குத் தேவை.

எனவேதான் இதுபோன்ற மென்பொருள் உருவாக்கத்திற்கு குறிப்பிட்ட மொழிகளின் இலக்கணம், மொழியியல் போன்றவை தேவை இல்லை என்று கருதப்படுகிறது.

இதையொட்டி எனக்கு ஏற்பட்ட ஒரு ஐயம், இயற்கைமொழிகளின் சிக்கலான மொழி அறிதிறனைக் கணினிக்குக் கொடுக்கமுடியாததால், எனவே அவற்றைக்கொண்டு கணினி நிரலாக்கம் செய்யமுடியாது என்பதால்தான், கணினி நிரலாக்கத்திற்கென்று பலகணினிநிரலாக்கமொழிகள் உருவாக்கப் பட்டுள்ளன. இந்த நிலையில் "செயற்கைச் செய்யறிவுத்திறன் சார் மென்பொருள்களுக்கு" இயற்கைமொழிகளின் மொழியறிவு இருந்தால், ஏன் இந்த இயற்கைமொழிகளைக் கொண்டு கணினி நிரலாக்கம் செய்யக்கூடாது? இந்த ஐயத்தை "சேட் ஜிபிடி (Chat GPT), ஜெமினி (Gemini AI), மெடா (Meta AI), கிளாட்(Claude AI)" போன்ற செய்யறிவுத்திறன்சார் மென்பொருள்களிடம் கேட்டேன். அதற்குக் கிடைத்த விடை . . ."இன்று அவ்வாறு செய்யமுடியாது, ஓரளவுக்குப் பயன்படுத்தலாம், அவ்வளவுதான். எதிர்காலத்தில் அந்த வளர்ச்சி ஏற்படலாம்" இதிலிருந்து நமக்குப் புரிகின்ற ஒரு உண்மை, இந்தச் செயற்கைச் செய்யறிவுத்திறன் சார் மென்பொருள்களுக்கு நாம் எதிர்பார்க்கிற 'மனித மூளையிலுள்ள மொழி

அறிவு - மொழி அறிதிறன்" கிடையாது. ஆனால் இந்த "மொழி அறிவு" இல்லாமலேயே நாம் எதிர்பார்க்கிற மொழிவழிச் செயல்பாடுகளை (language functions) மேற்கொள்ளும் "மொழித்திறன் - மொழிச் செயலறிதிறன்" இருக்கிறது. இது அறிவியலின் வளர்ச்சி. இதை நாம் மறுக்கமுடியாது. இருப்பினும் எச்சரிக்கை தேவை. இனி, மனித மூளையில் - மனிதனுக்கு - இருக்கிற மொழி அறிவுபற்றிய ஆய்வு, அதைக் குழந்தைகள் பெறுகிற முறைபற்றிய ஆய்வு, இலக்கண ஆய்வு, சமூகமொழியியல் ஆய்வு போன்றவை தேவை இல்லை என்ற முடிவுக்கு வந்துவிடக்கூடாது. இந்த எச்சரிக்கையோடுதான் நாம் இன்றைய செயற்கைச் செய்யறிவுத்திறன் சார் மென்பொருள்கள், அவற்றில் உள்ள பெரிய மொழிமாதிரிபற்றிய ஆய்வுகளைத் தொடரவேண்டும். மொழிகளின் இலக்கண ஆய்வுகளும் மொழியியல் ஆய்வுகளும் இன்னும் நீண்ட தொலைவைக் கடக்கவேண்டும்.

கணினிமொழியியலும் தமிழ்மொழியும்

கடந்த 125 ஆண்டுகளில் கணினிமொழியியல் துறை எவ்வாறு வளர்ந்து இன்று மிகப் பிரம்மாண்டமாகக் காணப்படுகிறது என்பதை இதுவரை பார்த்தோம். இந்த வரலாற்றுப் பின்னணியில் தமிழ்மொழியானது இத்துறையில் என்னென்ன வளர்ச்சிகளைக் கண்டுள்ளது என்பதை இனிப் பார்க்கலாம்.

80-களில் கணினியில் தமிழ் எழுத்துக்களையே காண இயலாத ஒரு சூழல்தான் நிலவியது. 90-களில் இந்தச் சூழல் மாறியது. தமிழ் வழியே உரைகளைக் கணினியில் உள்ளீடு செய்யவும் பார்க்கவும் வாய்ப்பு ஏற்பட்டது. இதன் முதல் கட்டமாக, தமிழ்ப் பனுவலை ரோமன் எழுத்துக்களில் உள்ளீடுசெய்து, பின்னர் தமிழ் எழுத்துக்களில் அதை மாற்றும் வசதி உள்ள "ஆதமி" போன்ற எழுத்துரு மென்பொருள்கள் உருவாக்கப்பட்டன. தமிழ் ஆர்வம் உள்ள பலர் உலக அளவில் இதுபோன்ற பணிகளில் தங்கள் பங்களிப்பைச் செய்தனர்.

இதன் அடுத்த கட்ட வளர்ச்சியாக, தமிழ் எழுத்துக்கள் வழியே தமிழ்ப்பனுவலை உள்ளீடு செய்யும் வளர்ச்சி ஏற்பட்டது. தமிழுக்கான எழுத்துருக்கள் உருவாக்கத்தில் தமிழ் ஆர்வலர்கள் பலர் பங்கேற்றனர். அப்போது எழுத்துருக்களுக்கான உள்ளீட்டுக் குறியேற்றம் (Encoding) அஸ்கி (ASCII - American Standard Code for Information Interchange).128 + 128இடங்களைக் கொண்ட இந்த குறியேற்றத்தைப் பின்பற்றித்தான் தமிழுக்கு எழுத்துருக்கள் உருவாக்கப்பட்டன. இந்த முறையில் ஆங்கில எழுத்துருக்களுக்கு அளிக்கப்பட்ட இடங்களைக் கொண்டுதான் தமிழுக்கு இடங்கள் அளிக்கப்பட்டன. இதனால் ஒரே நேரத்தில் ஆங்கிலத்தையும் தமிழையும் தட்டச்சிட முடியாது. ஒரே

நேரத்தில் ஒரு கோப்பில் ஆங்கிலத்தையும் தமிழையும் சேர்த்துக் காணமுடியாது. எடுத்துக்காட்டாக, ஆங்கில எழுத்துரு "k" -வுக்கு ஒதுக்கப்பட்ட அஸ்கி குறியீட்டைத் தமிழ் 'க்' -வுக்குப் பயன்படுத்தும்போது, ஒரே போக்கில் ஆங்கில "k" - வையும் தமிழ் "க்" -வையும் பார்க்க இயலாது. நாற்காலி ஒன்றுதான், நபர்கள் இரண்டு என்றால் ஒருவர் உட்காரும்போது மற்றொருவர் அதில் உட்காரமுடியாது. இது ஒரு சிக்கல்.

மற்றொரு சிக்கல், தமிழ் எழுத்துருக்களை உருவாக்கிய தனியார் வணிக நிறுவனங்கள், தனி நபர்கள் அனைவரும் ஒன்றுபோலத் தமிழுக்கு அஸ்கி குறியீட்டைப் பயன்படுத்தவில்லை. எனவே ஒரு குறிப்பிட்ட நிறுவனத்தின் தமிழ் எழுத்துருக்களைக்கொண்டு ஒரு கோப்பை உருவாக்கி, மற்றவர்களுக்கு அனுப்புவதில் சிக்கல் இருந்தது. நாம் கோப்பு அனுப்புகிற நண்பர்களிடம் நாம் பயன்படுத்திய எழுத்துருக்கள் இருக்கவேண்டும். இல்லையென்றால் அவரால் நாம் அனுப்பிய தமிழ்க் கோப்பை வாசிக்கமுடியாது.

1999-ஆம் ஆண்டு தமிழ் நாடு அரசு நடத்திய கணினித்தமிழ் மாநாட்டில் தமிழுக்கான குறியேற்றம் தரப்படுத்தப்பட்டது. தமிழ்99 என்று அதற்குப் பெயரிடப்பட்டது. அயலகத்தில் இருப்பவர்கள் இந்தப் பரிந்துரையை ஏற்கவேண்டும் என்று கட்டாயம் இல்லாததால் சிக்கல் நீடித்தது. கணினி விசைப்பலகைகளும் தரப்படுத்தப்பட்டன. அதைத் தொடர்ந்து பலவகைப்பட்ட தமிழ் எழுத்துருக்கள் சந்தையில் கிடைத்தன. ஆனால் மேற்கூறிய சிக்கல் தொடர்ந்தது. இருப்பினும் தமிழுக்கான சொற்பிழைதிருத்தி, சந்திப்பிழைதிருத்தி போன்ற மென்பொருள்கள் தமிழ் ஆர்வலர்களால் உருவாக்கப்பட்டன. தமிழ் நாடு அரசாங்கம் "தமிழ் இணையப் பல்கலைக்கழகம் (இன்றைய 'தமிழ் இணையக் கல்விக்கழகம்')" ஒன்றை ஏற்படுத்தி, கணினித்தமிழ் வளர்ச்சிக்கான பணிகளை மேற்கொண்டன.

21-ஆம் நூற்றாண்டின் தொடக்கத்தில் ஒருங்குறிக் குறியேற்றம் செயல்பாட்டுக்கு வந்தது. உலகில் உள்ள அனைத்து மொழிகளுக்கும் தனித்தனி குறியேற்றங்கள் அளிக்கப்பட்டன. தமிழுக்கும் 128 இடங்கள் அளிக்கப்பட்டன. தற்போது 72 இடங்கள் நிரப்பப்பட்டுள்ளன. 56 இடங்கள் காலியாக இருக்கின்றன. அதன் பயனாக, ஆங்கில எழுத்துருக்கான அஸ்கி குறியேற்றங்களைப் பயன்படுத்தாமல் தமிழுக்கென்று ஒதுக்கப்பட்ட குறியேற்றங்களைப் பயன்படுத்தமுடிந்தது. ஒருங்குறிக் குறியேற்றத்தின் பயனாகத் தற்போது ஒரு கோப்பு அல்லது ஒரு பக்கத்தில் ஒரே நேரத்தில் பல மொழிகளை உள்ளீடு செய்யும் வசதி உள்ளது.

ஒருங்குறி குறியேற்றத்தினால் நாம் முன்பு கண்ட எழுத்துருச் சிக்கல் தீர்ந்தது. இருப்பினும் இன்றுவரைகூட பலர் பழைய அஸ்கி குறியேற்றமுறையை அடிப்படையாகக்கொண்ட எழுத்துருக்களைப் பயன்படுத்தி வருகிறார்கள். இதன் காரணமாக, அஸ்கி எழுத்துருக்களை ஒருங்குறி குறியேற்ற எழுத்துருக்களாகமாற்றுவதற்கான மென்பொருள்கள் (Font Encoding Converters) தேவைப்பட்டன.

ஒருவாறு தமிழில் நிலவிய எழுத்துருக் குறியேற்றம், விசைப்பலகை சிக்கல்கள் பெருமளவு தீர்ந்தன. இதன் பயனாக, தமிழ்க் கணினிமொழியியல் ஆய்வுத் திட்டங்கள் தமிழ்நாட்டில் பல்கலைக்கழகங்களில் மேற்கொள்ளப்பட்டன. தமிழ் உருபன் ஆய்வு மென்பொருள்கள் உருவாக்கப்பட்டன; சொற்பிழை திருத்திகள் போன்றவை உருவாக்கப்பட்டன; பேச்சுத் தொழில்நுட்பத்திலும் (Speech Processing) அச்சடிக்கப்பட்ட தமிழ் நூல்களை ஒளிவழி நகல் எடுத்து, அவற்றைத் தமிழ் எழுத்துருக்களாக மாற்றியமைப்பதிலும் மென்பொருள்கள் (Optical Character Recognizer - OCR)உருவாக்கப்பட்டன.

இருப்பினும் மேலைநாடுகளில் வளர்ந்திருந்த கணினிமொழியியல், இயற்கைமொழி ஆய்வு போன்ற பிரிவுகள் பெரிய அளவில் தமிழ்நாட்டில் காலூன்றவில்லை. அண்ணா பல்கலைக்கழகம், சென்னை எம் ஐ டி ஏயூ - கேபிசி நிறுவனம் (MIT - AU-KBC), அமிர்தா பல்கலைக்கழகங்கள் போன்ற சில பல்கலைக்கழகங்களில் கணினியியல் துறைகள் கணினிமொழியியல், இயற்கைமொழி ஆய்வுத் திட்டங்களை மேற்கொண்டன. மைசூரில் உள்ள இந்திய மொழிகள் நிறுவனம் தரவகத் திட்டம் (Corpus projects) , பேச்சுத் தொழில்நுட்பம் (Speech Processing) போன்றவற்றில் சில பணிகளை மேற்கொண்டன.

கணினியியல் துறை, மொழியியல் துறையைச் சேர்ந்த பேராசிரியர்கள் சிலரும் இந்தத் திட்டங்களை மேற்கொண்டனர். தமிழ்நாட்டில் உள்ள அறிஞர்கள் மட்டுமல்லாமல் சிங்கப்பூர், மலேசியா, இலங்கை, கனடா, ஆஸ்திரேலியா, அமெரிக்கா போன்ற பல நாடுகளில் பணிபுரியும் கணினியியல், மொழியியல் அறிஞர்கள் தமிழ்க்கணிமொழியியல் வளர்ச்சிக்குப் பெரும்பங்கு ஆற்றிவருகிறார்கள்.

தமிழகத்திலேயே முதன்முதலாகச் சென்னைப் பல்கலைக்கழகத்தில் தமிழ்மொழித்துறை - மொழியியல் ஆய்வுப் பிரிவில் கணினிமொழியியலுக்கான முதுகலை, முனைவர் பட்டப் படிப்புக்கள் தொடங்கப்பட்டன. பல கருத்தரங்கங்கள் நடத்தப்பட்டன. ஆந்திர மாநிலத்தில் குப்பம் என்ற

இடத்தில் உள்ள "திராவிடப் பல்கலைக்கழகத்தில்"திராவிடமொழியியல் & கணினிமொழியியல் என்ற துறை நிறுவப்பட்டுள்ளது. தமிழ் நாட்டு அரசின் "தமிழ் இணையக் கல்விக்கழகம்" தமிழ்க் கணினிமொழியியலுக்காகச் சில பணிகளை மேற்கொண்டன. தமிழ் மென்பொருள் உருவாக்கத்திற்காக ஆய்வாளர்களுக்கும் ஆய்வு நிறுவனங்களுக்கும் நிதி உதவி அளித்தது; இப்போதும் தொடர்ந்து அளித்துக்கொண்டிருக்கிறது. ஹைதராபாத்தில் அமைந்துள்ள நடுவண் பல்கலைக்கழகத்தின் மொழியியல்துறை, அகில உலகத்தகவல் தொழில்நுட்ப ஆய்வுநிறுவனம் (IIIT - International Institute of Information Technology - Hydrabad) ஆகியவையும் ஏனைய இந்திய மொழிகளுடன் தமிழுக்குமான கணினிமொழியியல் ஆய்வுகளை மேற்கொண்டுவருகின்றன. கணினித்தமிழ் ஆர்வலர்கள் "கணித்தமிழ்ச் சங்கம்" என்ற ஒரு அமைப்பை ஏற்படுத்தி, தமிழ் எழுத்துரு, விசைப்பலகை, எழுத்துருக் குறியேற்றம் ஆகியவற்றைத் தரப்படுத்துவதில் பெரும் பங்கு அளித்துள்ளனர். இதுபோன்று உலக அளவிலான கணினிமொழியியல் ஆர்வலர்கள் "உலகத் தமிழ் தகவல் தொழில்நுட்ப மன்றம் - உத்தமம் (International Forum for Information Technology in Tamil - INFITT)" என்ற ஒரு அமைப்பை உருவாக்கி, உலக அளவிலான மாநாடுகள் நடத்திவருகின்றது.

தற்போது தமிழ்க் கணினிமொழியியல் ஆய்வின் பயனாக எழுத்துருக்கள், விசைப்பலகைகள், சொல்லாய்வு, தொடராய்வு, மின்னகராதி, சொற்பிழை திருத்தி, சந்திப்பிழை திருத்தி, பேச்சு - எழுத்து மாற்றி, எழுத்து-பேச்சு மாற்றி, ஒளிவழி அறிவான் போன்ற ஆய்வுக்கருவிகள், செயல்பாட்டு மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. தமிழ் இலக்கியங்கள், இலக்கணங்களைக் கணினியில் ஏற்றி, அவற்றை ஆய்வுசெய்வதற்கான தொடரடைவுபோன்ற மென்பொருள்களையும் உருவாக்கும் பணிகளில் தமிழ் ஆர்வம் உள்ள அறிஞர்கள் மேற்கொண்டுவருகின்றனர்.செம்மொழித் தமிழ் நடுவண் மையம் போன்ற ஆய்வு நிறுவனங்களும் இவைபோன்ற பணிகளை மேற்கொண்டுள்ளன. ஆங்காங்கே செயற்கைச் செய்யறிவுத்திறன்சார்ந்த பணிகளும் நடைபெற்றுவருகின்றன. அதன் அடிப்படையிலான பன்மொழி மொழிபெயர்ப்புப் பணிகளுக்கான மென்பொருள்களும் உருவாக்கப்பட்டுள்ளன. தமிழ் இலக்கிய ஆய்வையும் செயற்கைச் செய்யறிவுத் திறன்சார்ந்த தொழில்நுட்பத்தைப் பயன்படுத்திச் சிலர் மேற்கொண்டுள்ளனர்.

இணையத்தின் வளர்ச்சியானது கணினியில் பரந்த அளவில் தமிழ்ப் பயன்பாடுகளைப் பெருக்கின. வலைப்பூக்கள் தமிழில் நிலவுகின்றன. விக்கிமீடியாவிலும் தமிழ் இடம் பெற்றுள்ளது. இதற்கு முக்கியக் காரணமாக

அமைந்தவர் இலங்கையைச் சேர்ந்த திரு. மயூரநாதன் என்பவர் ஆவார் என்பதை இலங்கை மண்ணில் குறிப்பிடுவதில் மகிழ்வடைகிறேன். தமிழ் இலக்கியங்கள், தமிழ் இலக்கணங்கள் ஆகியவற்றைக் கணினியில் ஏற்றுவதும் அவற்றிற்கான சொல்லடைவு, தொடரடைவுக்கான கருவிகளை உருவாக்குவதிலும் சில நிறுவனங்களும் தனியே சில அறிஞர்களும் மேற்கொண்டுவருகின்றனர்.

செய்யறிவுத்திறன் சார்ந்த பெரிய மொழி மாதிரிகளைத் தமிழுக்குப் பயன்படும் வகையில் வளர்த்தெடுக்கும் முயற்சியில் தனிநபர்களும் ஈடுபட்டுவருகின்றனர்; சில தனியார் நிறுவனங்களும் ஈடுபட்டுவருகின்றன. குறிப்பாக, செயற்கைச் செய்யறிவுத்திறன்சார் மென்பொருள்களைப் பயன்படுத்தி, தமிழுக்கான பல மொழிவழிச் செயல்பாடுகளுக்கான மென்பொருள்களை உருவாக்கிவருகின்றனர். தமிழ் நாட்டில்மட்டும்அல்லாமல் சிங்கப்பூர், இலங்கை, மலேசியா, அமெரிக்கா போன்ற நாடுகளிலும் தமிழார்வலர்கள் இந்தப் பணிகளில் கவனம் செலுத்திவருகின்றனர்.

பன்னாட்டு நிறுவனங்களின் உருவாக்கத்தில் தற்போது வெளியிடப்பட்டுள்ள சேட் ஜிபிடி (Chat GPT), ஜெமினி (Gemini ai), மெடா (Meta ai), க்ளாட் (Claude ai) போன்ற செயற்கைச் செய்யறிவுத்திறன்சார் மென்பொருள்களில் தமிழும் இடம் பெற்றுள்ளது. தமிழ்வழியே ஐயங்களை நாம் முன்வைக்கும்போது, அவற்றிற்கான விடைகளைத் தமிழில் தருகின்றன. அதோடு, தமிழ் இலக்கணம் தொடர்பான வினாக்களுக்கும் விடைகளைத் தருகின்றன. சேட் ஜிபிடி -யில் தமிழ் சேட் ஜிபிடி என்ற ஒரு மென்பொருளும் இடம்பெற்றுள்ளது. க்ளாட் மென்பொருள் 70, 80 விழுக்காடு தமிழ் இலக்கணம்தொடர்பான வினாக்களுக்கும் ஐயங்களுக்கும் விடைகள் தருகின்றது. மெடா மென்பொருள் சில வேளைகளில் நாம் தமிழில் ஐயங்களை முன்வைத்தால் விடைகளை ஆங்கிலத்தில் தருகிறது. ஆனால் தமிழில் தருவதற்கான பணிகளும் நடைபெற்றுவருகின்றன என்று அது தெரிவிக்கிறது.

இவை அனைத்தின் முயற்சிகளும் வெற்றிபெறுவதற்குத் தமிழில் எண்மத் தரவுகள் மிகப் பெரிய அளவில் தேவைப்படுகின்றன. அனைத்துத் துறைகளிலும் தமிழ்மொழியின் செயல்பாடுகள் விரிவடையும்போதுதான் இந்தத் தேவையை நிறைவேற்றமுடியும். எங்கும் தமிழ் எதிலும் தமிழ் என்ற முழக்கத்தைச் செயலில் செய்துகாட்டும்போதுதான் இம்முயற்சியில் வெற்றியடையமுடியும். இன்றைய கணினியுகத்தில் கணினிவழிப் பயன்பாடுகளுக்கு ஒரு மொழி தன்னைத் தயார் படுத்திக்கொள்ளவில்லையென்றால், அதன் வளர்ச்சி உறுதியாகப் பின்னடைவுக்கு உட்படும் என்பதில் ஐயம் இல்லை.

Selected References

1. Allen, James. **Natural Language Understanding.** New Delhi: Pearson, 2007.
2. Anitha, Pillai S. **Automatic Parts of Speech Tagger for Machine Translation (With Special Reference to Malayalam).** Chennai: Ph.D., Thesis (University of Madras), 2007.
3. Bod, Rens, Jennifer Hay and Stefanie Jannedy. **Probabilistic Linguistics.** Massachusetts: The MIT Press, 2003.
4. Horrocks, Geoffrey. **Generative Grammar.** New York: Longman, 1989.
5. John, Patrick and David Christopher. **Computational Linguistics.** New Delhi: Commonwealth, 2011.
6. Jurafsky, Danierl and James H. Martin. **Speech and Language Processing (An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition).** New Jersey: Pearson / Prentice Hall, 2009.
7. Kenneth, Beesley R. **Finite State Morphology.** California: CSLI Publications, 2003.
8. Koehn, Philipp. **Statistical Machine Translation.** Cambridge: Cambridge University Press, 2010.
9. Kumae, Ela. **Natural Language Processing.** New Delhi: I.K. International Publishing Pvt. Ltd., 2011.
10. Manning, Christopher D & Hinrich Schutze. **Foundations of Statistical Natural Language Processing.** Massachusetts: The MIT Press, 1999.
11. Mitkov, Ruslan (Editor). **The Oxford Handbook of Computatioanl Linguistics.** Oxford, 2009.
12. Prabakaran. P, **Context Dependent Allophones in Written Tamil (With Special Reference to TTS).** Chennai: Ph.D. Thesis , University of Madras, 2010.
13. Padmamala, R.**Issues in Syntactic for Modern Tamil.** Chennai: Ph.D., Thesis (University of Madras), 2013.
14. Pustejovsky, James and Amber Stubbs. **Natural Language Annotation for Machine Learning.** Sebastopol, 2012.

15. Shanmugam R, **Issues in Morphological Parsing for Modern Tamil.** Chennai: Ph.D. Thesis, University of Madras, 2010.

16. Renganathan, Vasu. **Computational Approaches to Tamill Linguistics.** Chennai: Cre-A, 2020.

17. Siddiqui, Tanveer and U.S. Tiwary.**Natural Language Processing and Information.** New Delhi: Oxford Universit Press, 2010.

18. Deiva Sundaram, N.**Diglossic Situation in Tamil (Sociolinguistic Approach).** Chennai: Ph.D., Thesis (University of Madras), 1980.

19. Umadevi, K. **English Equivalence for Tamil Inflection (A Transfer Module in Tamil - English Machine Translation System).** Chennai: Ph.D., Thesis (University of Madras), 2014.

20. Venkatachalam, K. **Word Sense Disambiguation in Machine Translation (With Special Reference to Tamil - English).** Chennai: Ph. D., Dissertation, University of Madras, 2009.

21. அரங்கன் கி.**நோம் சாம்ஸ்கி - பன்முக அறிமுகம்.** கோயம்புத்தூர்: மொழியியல் துறை, பாரதியார் பல்கலைக்கழகம், 2013.

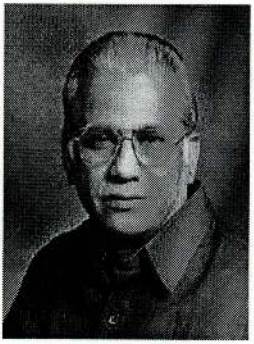
22. குமார்,ப. **இயந்திர மொழிபெயர்ப்பு.** சென்னை: முனைவர் பட்டத்திற்கான ஆய்வேடு (சென்னைப் பல்கலைக்கழகம்) , 2007.

23. இராசாராம். சு. **நோம் சாம்ஸ்கி.** நாகர்கோவில் : காலச்சுவடு பதிப்பகம், 2019.

24. தெய்வ சுந்தரம் ந. **மொழியும் தமிழ் இலக்கணமும்.** சென்னை: அமுத நிலையம், 2021.

25. தெய்வ சுந்தரம் ந. **மொழியியலும் கணினி மொழியியலும்.** சென்னை: அமுத நிலையம், 2021.

26. தெய்வ சுந்தரம் ந. **மொழியியல் பார்வையில் தமிழ் இலக்கணம்.** சென்னை, 2023.



பேராசிரியர் ந. தெய்வ சுந்தரம்

முனைவரும் பேராசியருமான ந. தெய்வ சுந்தரம் அவர்கள் சென்னைப் பல்கலைக் கழகத்தின் தமிழ்மொழித் துறையில் விரிவுரையாளர், இணைப்பேராசிரியர், பேராசிரியராகவும், மொழியியல் பிரிவின் இயக்குனராகவும் 1985 - 2010 காலப்பகுதிவரை கடமையாற்றியுள்ளார்கள். தற்போது மேலாண் இயக்குநராக என்.டி.எஸ். லிங்க் சாப்ட் சொலுசன்ஸ் (NDS LINGSOFT SOLUTIONS PRIVATE LIMITED) சென்னையில் பணியாற்றிக் கொண்டிருக்கின்றார்கள். இவர் ஒரு விஞ்ஞானப்பட்டதாரியாக இருந்து(B. Sc.Physics) பின்னர் தமிழிலும் மொழியியலிலும் பட்டப்பின் கற்கைகளை (MA) மேற்கொண்டதுடன் தனது கலாநிதி பட்டத்தையும் மொழியியலிலேயே பெற்றிருந்தார்கள். மேலும் முதுமுனைவர் ஆய்வுகளை இரு தடவைகள் சென்னைப் பல்கலைக்கழகத்திலும் (1983-1985), உலகத் தமிழாராய்ச்சி நிறுவனத்திலும் (1981-1983) மேற்கொண்டதுடன் இவருடைய சிறப்பு ஆய்வுகளாகத் தமிழ் இலக்கணம், தமிழ்க்கணினிமொழியியல் (Tamil Computational Linguistics), நரம்புமொழியியல் (Neuro Linguistics) சிகிச்சைமொழியியல் (Clinical Linguistics), தமிழ் கற்பித்தல் போன்றவை சிறப்பாகக் குறிப்பிடத்தக்கன.

இவற்றிற்கும் மேலாக பல ஆய்வு மாநாடுகளில் இலங்கை, இந்தியா, சிங்கப்பூர், மலேசியா, டென்மார்க் போன்ற நாடுகளில் பங்குபற்றி 25க்கும் மேற்பட்ட ஆய்வுக் கட்டுரைகளை வெளியிட்டிருக்கின்றார். அத்துடன் அவரின் விருப்பு ஆய்வுகளான கணினிமொழியியல், மொழிகற்பித்தல், அகராதியியல் தொடர்பாகப் பல நூல்களையும் செயற்பாட்டு மென்பொருள்களையும் தமிழ் தொடர்பாக உருவாக்கியுள்ளார்.

1. சுவிதா 99 (தமிழ்ச்சொல்லாளர்),
2. தமிழ்ச் சொல் 2000 (குடியரசுத் தலைவர் கே. ஆர். நாராயணன் அவர்களால் புது டெல்லியில் வெளியிடப்பட்ட பெருமைவாய்ந்தவர்)
3. பேச்சுத் தமிழ் - பல்லூடக மென்பொருள் (2008),
4. மென்தமிழ் (2011),
5. எழுத்துத் தமிழ்- பல்லூடக மென்பொருள்,
6. மின்னகராதிகள்,
7. தமிழ் கணினிவழி மொழிபெயர்ப்பு,
8. கணினிவழி தமிழ் இலக்கணம்,
9. தமிழ் மெய்ப்புதிருத்து கருவி,
10. தமிழ்ச் சொற்பிழை திருத்தி,
11. தமிழ்ச் சந்திப்பிழை திருத்தி

எனப் பலமென் பொருள் நூல்களையும் இத்துடன் தமிழ் இணைச்சொல், எதிர்ச்சொல் அகராதி, ஆட்சிச் சொல் அகராதி, அயற்சொல் அகராதிகளையும் பேராசிரியர். தெய்வசுந்தரம் உருவாக்கியுள்ளார்.

அத்துடன் 100க்கும் மேற்பட்ட எம்.பில் ஆய்வுகள் 25ற்கும் மேற்பட்ட முனைவர் பட்ட ஆய்வுகளுக்கு வழிகாட்டியாக உலகெங்கிலும் பல மாணவர்களை உருவாக்கியுள்ளார்கள். இவரின் பெருமைக்கு சான்றாக **தமிழ்நாடு முதலமைச்சர் கணினித் தமிழ் விருதை** 2014 ஆண்டிலும், வாழ்நாள் சாதனையாளர் விருதை 2019 மலேசியாவிலும் பெற்றுள்ளார் என்பதை இங்குச் சிறப்பாகக் குறிப்பிடலாம். குறிப்பாகக் கணினிமொழியியல் பற்றிய அவர்களின் ஆய்வுகள் தமிழ் மொழிவளர்ச்சிக்கு அவர் ஆற்றிவரும் பெரும் பங்களிப்பாக விளங்குகின்றன.

பேராசிரியர் ந. தெய்வ சுந்தரம்

முனைவரும் பேராசியருமான ந. தெய்வ சுந்தரம் அவர்கள் சென்னைப் பல்கலைக் கழகத்தின் தமிழ்மொழித் துறையில் விரிவுரையாளர், இணைப்பேராசிரியர், பேராசிரியராகவும், மொழியியல் பிரிவின் இயக்குனராகவும் 1985 - 2010 காலப்பகுதிவரை கடமையாற்றியுள்ளார்கள். தற்போது மேலாண் இயக்குநராக என்.டி.எஸ். லிங்க் சாப்ட் சொலுசன்ஸ் (NDS LINGSOFT SOLUTIONS PRIVATE LIMITED) சென்னையில் பணியாற்றிக் கொண்டிருக்கின்றார்கள். இவர் ஒரு விஞ்ஞானப்பட்டதாரியாக இருந்து(B. Sc.Physics) பின்னர் தமிழிலும் மொழியியலிலும் பட்டப்பின் கற்கைகளை (MA) மேற்கொண்டதுடன் தனது கலாநிதி பட்டத்தையும் மொழியியலிலேயே பெற்றிருந்தார்கள். மேலும் முதுமுனைவர் ஆய்வுகளை இரு தடவைகள் சென்னைப் பல்கலைக்கழகத்திலும் (1983-1985), உலகத் தமிழாராய்ச்சி நிறுவனத்திலும் (1981-1983) மேற்கொண்டதுடன் இவருடைய சிறப்பு ஆய்வுகளாகத் தமிழ் இலக்கணம், தமிழ்க்கணினிமொழியியல் (Tamil Computational Linguistics), நரம்புமொழியியல் (Neuro Linguistics) சிகிச்சைமொழியியல் (Clinical Linguistics), தமிழ் கற்பித்தல் போன்றவை சிறப்பாகக் குறிப்பிடத்தக்கன.

இவற்றிற்கும் மேலாக பல ஆய்வு மாநாடுகளில் இலங்கை, இந்தியா, சிங்கப்பூர், மலேசியா, டென்மார்க் போன்ற நாடுகளில் பங்குபற்றி 25க்கும் மேற்பட்ட ஆய்வுக் கட்டுரைகளை வெளியிட்டிருக்கின்றார். அத்துடன் அவரின் விருப்பு ஆய்வுகளான கணினிமொழியியல், மொழிகற்பித்தல், அகராதியியல் தொடர்பாகப் பல நூல்களையும் செயற்பாட்டு மென்பொருள்களையும் தமிழ் தொடர்பாக உருவாக்கியுள்ளார்.